RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 28
(2007)

David B. Pisoni, Ph.D.
Principal Investigator

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405-1301

Research Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. DC-00111

and

National Institutes of Health
Training Grant No. DC-00012

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)

Table of Contents

INTRODUCTION

This is the twenty-eighth annual progress report summarizing research activities on speech perception and spoken language processing carried out in the Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize our accomplishments over the past year and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief summaries of "on-going" research projects in the laboratory. From time to time, we also have included new information on instrumentation and software developments when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating research.

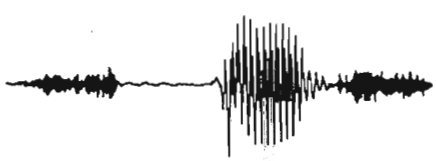
We are distributing progress reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of spoken language processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception and spoken language processing and we would be grateful if you and your colleagues would send us copies of any recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405-1301
USA

Telephone: (812) 855-1155, 855-1768
Facsimile: (812) 855-1300
E-mail: pisoni@indiana.edu
Web: http://www.indiana.edu/~srlweb

Copies of this report are being sent primarily to libraries and research institutions rather than individual scientists. Because of the rising costs of publication and printing, it is not possible to provide multiple copies of this report to people at the same institution or issue copies to individuals. We are eager to enter into exchange agreements with other institutions for their reports and publications. Please write to the above address for further information.

The information contained in this progress report is freely available to the public and is not restricted in any way. The views expressed in these research reports are those of the individual authors and do not reflect the opinions of the granting agencies or sponsors of the specific research.



SPEECH – THE FINAL FRONTIER

SPEECH RESEARCH LABORATORY
FACULTY, STAFF, AND TECHNICAL PERSONNEL

(January 1, 2007–December 31, 2007)

RESEARCH PERSONNEL

David B. Pisoni, Ph.D. ..... Chancellor's Professor of Psychology and Cognitive Science[1,2]

Steven B. Chin, Ph.D. ..... Associate Professor in Otolaryngology–Head and Neck Surgery[3]
Derek Houston, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3]
Tonya Bergeson-Dana, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3]
Marcia Hay-McCutcheon, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3]
Thomas M. Gruenenfelder, Ph.D. ..... Research Scientist/Visiting Professor of Psycholgy/Brain Sciences

Nathaniel Peterson, M.D. ..... NIH Postdoctoral Trainee[3]
Mary Fagan, Ph.D. ..... NIH Postdoctoral Trainee[3]
Christopher M. Conway, Ph.D. ..... NIH Postdoctoral Trainee
Jeremy Loebach, Ph.D ..... NIH Postdoctoral Trainee
Tessa Bent, Ph.D. ..... NIH Postdoctoral Trainee
Robert Felty, Ph.D. ..... NIH Postdoctoral Trainee
Susannah Levi, Ph.D. ..... NIH Postdoctoral Trainee[4]
Jessica Beer, Ph.D. ..... Postdoctoral Trainee
Adam B. Buchwald, Ph.D. ..... NIH Postdoctoral Trainee[5]
Vanessa Taler, Ph.D. ..... Postdoctoral Trainee

Joshua Radicke, B.S. ..... NIH Predoctoral Trainee
Caitlin M. Dillon, B.A. ..... NIH Predoctoral Trainee[6]
Vsevold Kapatsinski, B.A. ..... NIH Predoctoral Trainee
Rebecca Ronquest, B.A ..... NIH Predoctoral Trainee
Esperanza Anaya, B.A ..... NIH Predoctoral Trainee
Nick Altieri, B.A. ..... Graduate Research Assistant

---

[1] Also Adjunct Professor of Linguistics, Indiana University, Bloomington, IN.

[2] Also Adjunct Professor of Otolaryngology–Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

[3] Department of Otolaryngology–Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

[4] Now at Department of Linguistics, University of Michigan.

[5] Now at Department of Speech-Language Pathology and Audiology, New York University.

[6] Now at Haskins Laboratories, New Haven, CT.

TECHNICAL PERSONNEL

Luis R. Hernández, B.A. ..... Research Associate in Psychology
Darla J. Sallee ..... Administrative Assistant
Wesley Alford. ..... Programmer
Pablo VanWoerkam ..... Programmer
Clint Wylie ..... Programmer

Melissa Troyer ..... Undergraduate Research Assistant
Jennifer Karpicke ..... Undergraduate Research Assistant
Althea Bauernschmidt ..... Undergraduate Research Assistant
Larry Phillips ..... Undergraduate Research Assistant
Jaymi Light ..... Undergraduate Research Assistant
Katie Scherschel ..... Undergraduate Research Assistant

E-MAIL ADDRESSES

Althea Bauernschmidt ..... abauerns@indiana.edu
Jessica Beer ..... jesbeer@indiana.edu
Tessa Bent ..... tbent@indiana.edu
Tonya Bergeson-Dana ..... tbergeso@iupui.edu
Adam B. Buchwald ..... adambuchwal@nyu.edu
Steven B. Chin ..... schin@iupui.edu
Christopher M. Conway ..... cmconway@indiana.edu
Tom Gruenenfelder ..... tgruenen@indiana.edu
Marcia Hay-McCutcheon ..... rmhaymccu@indiana.edu
Luis R. Hernández ..... hernande@indiana.edu
Derek Houston ..... dmhousto@indiana.edu
Vsevold Kapatsinski ..... vkapatsi@indiana.edu
Jennifer Karpicke ..... jkarpick@indiana.edu
Susannah Levi ..... svlevi@umich.edu
Jeremy L. Loebach ..... jlloebac@iupui.edu
David B. Pisoni ..... pisoni@indiana.edu
Josh Radicke ..... jradicke@indiana.edu
Rebecca Ronquest ..... rronques@indiana.edu
Darla J. Sallee ..... dsallee@indiana.edu
Melissa Troyer ..... mltroyer@indiana.edu
Vanessa Taler ..... vtaler@indiana.edu

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Efficacy and Effectiveness of Cochlear Implants in Deaf Children[1]

David B. Pisoni,[2] Christopher M. Conway, William Kronenberger,[2]
David L. Horn,[2] Jennifer Karpicke and Shirley Henning[2]

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] The research described in this chapter was supported by NIH-NIDCD Training Grant T32DC00012 and NIH-NIDCD Research Grants R01DC00111, NIH-NIDCD R01DC00064 to Indiana University. We thank Luis Hernandez and Darla Sallee for their help and assistance on various phases of this work over the years. Chapter to appear in M. Marschark and P. Hauser (Eds.), "Deaf Cognition: Foundations and Outcomes" in 2008.
[2] Indiana University School of Medicine, Indianapolis, IN.

Efficacy and Effectiveness of Cochlear Implants in Deaf Children

Abstract. A large body of clinical research over the last decade demonstrates that cochlear implants work and provide significant speech and language benefits to profoundly deaf adults and prelingually deaf children. The most challenging research problem today is that cochlear implants do not work equally well for everyone who has a profound hearing loss and cochlear implants frequently do not provide much benefit at all under highly degraded listening conditions. Some individuals do extremely well on traditional audiologic outcome measures with their cochlear implants when tested under benign listening conditions in the clinic and research laboratory while others have much more difficulty. However, all patients with cochlear implants uniformly have difficulty in a number of challenging perceptual domains such as: listening in noise, talking on the telephone, localizing sounds, recognizing familiar voices and different dialects, identifying environmental sounds and listening to music. The enormous variability in outcome and benefit following implantation is not surprising because none of the current generation of cochlear implants successfully restores normal hearing or supports robust speech perception and spoken language processing across all of these difficult and highly variable listening conditions. The traditional outcome measures of audiologic benefit were never designed to assess, understand or explain individual differences in speech perception and spoken language processing. In this chapter, we summarize recent findings that suggest several promising new directions for understanding and explaining variability in outcome and benefit after implantation. These results have implications for the design of new cochlear implants as well as the development of radically new approaches to intervention, training and habilitation following implantation.

Introduction

One aspect of our research program at the Indiana University School of Medicine has been concerned with understanding the large individual differences in speech and language outcomes in deaf children who have received cochlear implants (CIs). We are interested in explaining and predicting the enormous variability observed in a wide range of conventional measures of speech and language following cochlear implantation. The degree of variation in clinical outcome measures is enormous and is a robust finding observed universally at all implant centers around the world. The variability observed in outcome and benefit following cochlear implantation remains a significant problem for both clinicians and researchers alike. Why do some profoundly deaf children do so well with their CIs and why do other children do more poorly? The problem of individual differences in outcome and benefit is a major clinical issue in the field that has been addressed repeatedly over the years by the two earlier NIH Consensus Conferences on CIs (1988, 1995).

Despite the importance of understanding and explaining variability and individual differences following CI, very little solid progress has been made in identifying the neurobiological substrates and cognitive factors that are responsible for individual variation in speech and language outcomes. Knowledge and understanding of these factors and the information-processing subsystems that are affected by profound deafness and language delay is critical for diagnosis, prediction and treatment and for explaining why some children do poorly with their CIs. Several reasons can be proposed for the unsatisfactory state of affairs concerning variability and individual differences.

First, most of the people who work in the field of hearing impairment and CIs are clinicians. The CI surgeons, audiologists and speech language pathologists are primarily interested in the medical care of the patient and demonstrating the efficacy of CIs as a medical treatment for profound deafness. For them, individual differences and variability in speech and language outcome are viewed as a source of undesirable noise, a "nuisance variable" so to speak, that needs to be reduced or eliminated in order to reveal the true underlying benefits of cochlear implantation. When a child does well with his or her CI, the family, clinical team, teachers and other professionals are all delighted with the outcome. However, when a child does poorly with an implant, the clinical team is at a loss to explain the anomaly or suggest alternatives about what to do next. At the present time, given the nature of the clinical research carried out on CIs, it is unclear even how to approach the study of individual differences in this clinical population. What factors are responsible for the individual differences in outcome and benefit? What behavioral and neurocognitive domains should be investigated? What kinds of measures should be obtained? What theoretical approach should be adopted to study this problem?

Second, the conventional battery of speech and language tests that is routinely administered to measure clinical outcome and benefit was developed by the CI manufacturers to establish efficacy as part of the clinical trials for FDA approval. These behavioral tests were never designed to measure individual differences or assess variability in outcome. Moreover, and perhaps more importantly, the foundational assumptions and theoretical framework underlying the selection and use of the conventional speech and language outcome measures that speech perception and spoken language processing recruit formal rules and context-free symbolic representations is now being seriously questioned and undermined. The formalist assumption that everyone comes up with the same grammar of language despite vastly different individual developmental histories has been questioned in recent years in light of new knowledge about brain structure and function and the development of adaptive self-organizing systems like speech and language. The old static views of language as an idealized homogeneous context-free system of abstract linguistic knowledge are being replaced by new conceptions linking mind, body and world together in a complex interactive system (Clark, 1997).

Third, because the primary focus of most of the research on CIs has been clinical in nature, that is, demonstrating efficacy and safety and establishing that CIs work well under quiet testing conditions in the clinic or research laboratory, the typical battery of conventional behavioral tests only provide measures of the final "product" or "end-point" of a long series of neural and cognitive processes. All of the current outcome measures routinely used in the clinic and research laboratory rely on accuracy and percent correct as the primary dependent variable to assess performance and document benefit following cochlear implantation. Unfortunately, end-point measures of performance while they have strong face validity and are used successfully to demonstrate efficacy of CIs, are fundamentally unable to measure and assess the underlying elementary information processing variables like speed, capacity, learning and memory, inhibition, attention, cognitive control and the neurocognitive operations that are used in performing the specific individual behavioral tasks used to assess the benefits of CIs.

In addition, because the field of clinical audiology is an applied science drawing knowledge and methods from several different related disciplines, there is no common integrated theoretical framework to motivate the choice of specific outcome measures and tests, interpret the results and findings, provide explanations or make predictions. Without the benefit of a well-defined conceptual framework and additional theoretically-motivated "process-based" measures of performance, it is impossible to gain any new knowledge about the underlying neural and neurocognitive factors that are responsible for the observed variability in the traditional audiological outcome measures of performance. Without knowing what factors are responsible for the individual differences and understanding the basis for variation in performance, it is difficult to motivate and select a specific approach to habilitation and therapy after

cochlear implantation. Moreover, all of the clinical research on CIs has been primarily descriptive in nature and not experimentally motivated by hypothesis testing or specific predictions leading to understanding and explanation of process and mechanism. The bulk of CI research has focused on medical, demographic and educational factors, not the underlying neurobiological or neurocognitive processes that link brain and behavior.

Given what we know about population variability in biology, it is very likely that deaf children who are performing poorly with their cochlear implants are a heterogeneous group that differs in numerous ways from each other reflecting dysfunction of multiple processing systems associated with deafness and language delays. Adopting a common uniform approach to assessment, therapy and habitation after cochlear implantation will be inadequate to accommodate a wide range of individual differences and subtypes in outcome and benefit. Without knowing how and why poorer performers differ from each other and from the exceptionally good performers, as well as typically-developing hearing children, it is difficult to establish realistic goals and generate expectations for treatment and intervention following implantation. Moreover, it is unlikely that an individual child will be able to achieve optimal benefits from his/her implant without knowing why this child is having problems and what specific neurocognitive domains are involved.

Deaf Children as a "Model System" for Development. Two reasons motivate our interest in studying deaf children with CIs. The first is clinical in nature. CIs provide a medical treatment for profound deafness and have been shown to facilitate the development of spoken language. Without some kind of medical or behavioral intervention, a profoundly deaf child will not learn language normally from caretakers in his or her surrounding environment and will be unable to achieve his/her full intellectual potential as productive members of society. No one argues with this reason for studying deaf children. Sensory deprivation is a significant neurodevelopmental problem that has lasting and permanent effects on brain development and intellectual achievement. A profound hearing loss at birth is uniformly viewed by hearing people as a clinically significant sensory disability, an impairment that affects cognitive, social and intellectual development. Almost all of the clinical research on CIs has been concerned with device efficacy, that is, demonstrating that CIs work and provide benefit to profoundly deaf children and adults. In contrast, very little research has been devoted to effectiveness and, specifically, to understanding the reasons for the enormous variability in outcome and benefit following implantation.

When considering the efficacy of a treatment or intervention, we mean the power to produce a desired effect in an individual, that is, does a CI work and provide benefit to a profoundly deaf person? In contrast, when considering the effectiveness of a treatment or intervention, we mean actually producing the expected effect, that is, does a CI work equally well and provide the desired benefit in everyone who is a candidate and receives a CI?

A second major reason for our interest in studying deaf children with CIs is more basic in nature in terms of theoretical implications for gaining fundamental new knowledge about learning, development and neural plasticity. Deaf children with CIs represent a unique and unusual clinical population because they provide an opportunity to study brain plasticity and neural reorganization after a period of auditory deprivation and a delay in language development. In some sense, the current research efforts on deaf children with CIs can be thought of as the modern equivalent of the so-called "forbidden experiment" in the field of language development but with an unusual and somewhat unexpected and positive consequence. The forbidden experiment refers to the proposal of raising a child in isolation without exposure to any language input in order to investigate the effects of early experience on language development. These kinds of isolation experiments are not considered ethical with humans although they

are a common experimental manipulation with animals to learn about brain development and neural reorganization in the absence of sensory input.

Following a period of sensory deprivation from birth, a medical intervention is now available that can be used to provide a form of "electrical" hearing to a congenitally deaf child. A CI provides electrical stimulation to the auditory system, the brain and nervous system, therefore facilitating development of the underlying neurobiological and cognitive systems used in speech and language processing as well as other domains of neuropsychological function.

The current population of deaf children who use cochlear implants also provides an unusual opportunity for developmental scientists to study the effects of early experience and activity-dependent learning and to investigate how environmental stimulation and interactions with caretakers shapes the development of perception, attention, memory, and a broad range of other neurocognitive processes such as sensory-motor coordination, visual-spatial processing and cognitive control, all of which may be "delayed" or "reorganized" as a consequence of a period of early auditory deprivation resulting from congenital or prelingual deafness prior to implantation and the associated delays in language development. When viewed in this context, the clinical and theoretical implications of research on deaf children with CIs are quite extensive. Research on this clinical population will contribute new knowledge and understanding about important contemporary problems in cognitive development and developmental cognitive neuroscience.

Perceptual Robustness of Speech. Research on deaf children who use CIs will also contribute new knowledge about perceptual learning and adaptation in speech perception and spoken language understanding. The most distinctive property of human speech perception is its perceptual robustness in the face of diverse physical stimulation over a wide range of environmental conditions that produce significant changes and perturbations in the acoustic signal. Hearing listeners adapt very quickly and effortlessly to changes in speaker, dialect, speaking rate and speaking style and are able to adjust rapidly to acoustic degradations and transformations such as noise, filtering, and reverberation that introduce significant physical changes to the speech signal without apparent loss of performance (Pisoni, 1997). Investigating the perceptual, neurocognitive and linguistic processes used by deaf listeners with CIs and understanding how hearing listeners recognize spoken words so quickly and efficiently despite enormous variability in the physical signal and listening conditions will provide fundamental new knowledge about the sources of variability in outcome and benefit in patients who use CIs.

What is a Cochlear Implant? A cochlear implant is a surgically implanted electronic device that functions as an auditory prosthesis for a patient with a severe to profound sensorineural hearing loss. The device provides electrical stimulation to the surviving spiral ganglion cells of the auditory nerve bypassing the damaged hair cells of the inner ear to restore hearing in both deaf adults and children. The device provides patients with access to sound and sensory information from the auditory modality.

The current generation of multi-channel cochlear implants consist of an internal multiple electrode array and an external processing unit. The external unit consists of a microphone that picks up sound energy from the environment and a signal processor that codes frequency, amplitude and time and compresses the signal to match the narrow dynamic range of the ear. Cochlear implants provide temporal and amplitude information. Depending on the manufacturer, several different place coding techniques are used to represent and transmit frequency information in the signal.

For postlingually profoundly deaf adults, a CI provides a transformed electrical signal to an already fully developed auditory system and intact mature language processing system. Postlingually

deaf patients have already acquired spoken language under typical listening conditions so we know their central auditory system and brain have developed normally. In the case of a congenitally deaf child, however, a CI provides novel electrical stimulation through the auditory sensory modality and an opportunity to perceive speech and develop spoken language for the first time after a period of auditory deprivation.

Congenitally deaf children have not been exposed to speech and do not develop spoken language normally. Although the brain and nervous system continue to develop and mature in the absence of auditory stimulation, there is now increasing evidence suggesting that some cortical reorganization has already taken place during the period of sensory deprivation before implantation and that several aspects of speech and language as well as other cognitive processes and neural systems may be delayed and/or disturbed and develop in an atypical fashion after implantation. Although both peripheral and central differences in neural and cognitive function are likely to be responsible for the wide range of variability observed in outcome and benefit following implantation, increasing evidence suggests that the enormous variability in outcome and benefit following cochlear implantation cannot be explained as a simple sensory impairment in detection and/or discrimination of auditory signals. Other more complex cognitive and neural processes are involved.

Cochlear Implants Do Not Restore Normal Hearing. Although CIs work reasonably well with a large number of profoundly deaf children and adults under quiet listening conditions, it is important to emphasize that CIs do not restore normal hearing and they do not provide support for the highly-adaptive robust speech perception and spoken language processing routinely observed in hearing listeners under a wide range of challenging listening conditions. The difficulties consistently reported by CI patients under difficult listening conditions are both theoretically and clinically important because they reflect fundamental differences between acoustic hearing and electrical stimulation of the auditory system. These difficulties demonstrate that the rapid adaptation, tuning and continuous adjustment of the perceptual processes that are the hallmarks of robust speech perception by hearing listeners have been significantly compromised by the processing and stimulation strategies used in the current generation of CIs as well as any neural reorganization that may have taken place before implantation.

While everyone working in the field acknowledges the difficulties that CI patients have listening in noise, these problems are not explicitly discussed extensively in the literature nor are they considered to be major research questions. Because of their fundamental design, CIs create highly degraded "underspecified" neural representations of the phonetic content and indexical properties of speech which propagates and cascades to higher processing levels. Although the degraded electrical signal can often be interpreted by most deaf listeners as human speech and can support spoken word recognition and lexical access under quiet listening conditions, the fine episodic acoustic-phonetic details of the original speech waveform are not reliably reproduced or transmitted to the peripheral auditory nerve, central pathways or higher cortical areas that are used for recognition, categorization and lexical discrimination and selection. Moreover, the internal perceptual spaces that are used to code and represent linguistic contrasts are significantly warped and deformed in ideopathetic ways by the unique pathology of each individual patient (Harnsberger et al., 2001). When confronted with different sources of variability which transform and degrade the speech signal in various ways, patients with CIs often have a great deal of difficulty perceiving speech and understanding the linguistic content of the talkers' intended message.

The speech perception and spoken word recognition problems experienced by patients with CIs also reflect impairments and disturbances in the neural circuits and categorization strategies that are routinely used to compensate and maintain perceptual constancy in the face of variability in the speech signal. Hearing listeners routinely have similar problems in noise and under high cognitive load but they

can cope and overcome the variability and degradation. In some cases, such as listening in high levels of noise or against a background of multi-talker babble, patients are unable to derive any benefits at all from their CI and often turn their device off because the speech signal is unpleasant or becomes an aversive stimulus to them.

Key Findings on Outcome and Benefit Following Cochlear Implantation

What do we know about outcome and benefit in deaf children with CIs? Table I lists seven key findings that have been observed universally at all implant centers around the world. These findings indicate that a small number of demographic, medical and educational factors are associated with speech and language outcome and benefit following implantation. In addition to the enormous variability observed in these outcome measures, several other findings have been consistently reported in the clinical literature on cochlear implants in deaf children. An examination of these findings provides some initial insights into the possible underlying cognitive and neural basis for the variability in outcome and benefit among deaf children with cochlear implants. When these contributing factors are considered together, it is possible to begin formulating some more specific hypotheses about the reasons for the variability in outcome and benefit.

Table I

Key Findings on Outcome and Benefit Following Cochlear Implantation

• Large Individual Differences in Outcomes
• Age of Implantation (Sensitive Periods)
• Effects of Early Experience (Auditory-Oral vs. Total Communication)
• No Preimplant Predictors of Outcome
• Abilities "Emerge" after Implantation (Learning)
• "Cross-Modal Plasticity" and "Neural Reorganization"
• Links Between Speech Perception & Production

Much of the past research on CI's has been concerned with questions of assessment and device efficacy using outcome measures that were based on traditional audiological criteria. These clinical outcome measures included a variety of hearing tests, speech discrimination, word recognition and comprehension tests, as well as some standardized vocabulary and language assessments as well as other assessments of speech production, articulation and speech intelligibility. The major focus of most clinical research has been concerned with the study of demographic variables as predictors of these outcome measures. The available evidence suggests that age at onset of deafness, length of deprivation and age at implantation are all strongly associated with the traditional audiological outcome measures (Fryauf-Bertschy et al., 1997; Osberger, Miyamoto, Zimmerman-Phillips et al., 1991; Staller, Pelter, Brimacombe, Mecklenberg, & Arndt, 1991; Waltzman et al., 1994, 1997).

Age at Implantation. Age at implantation has been shown to influence all outcome measures of performance. Children who receive an implant at a young age do much better on a whole range of outcome measures than children who are implanted at older ages. Length of auditory deprivation or duration of deafness is also related to outcome and benefit. Children who have been deaf for shorter periods of time before implantation do much better on a wide variety of clinical measures than children who have been deaf for longer periods of time. Both findings demonstrate the contribution of sensitive periods in sensory, perceptual, and linguistic development and serve to emphasize the close links that exist between neurobiological development and behavior, especially development of hearing, speech and

spoken language and the neural systems that support these processes (Ball & Hulse, 1998; Konishi, 1985; Konishi & Nottebohm, 1969; Marler & Peters, 1988).

Effects of Early Experience. Early sensory and linguistic experience and processing activities after implantation have also been shown to affect performance on a wide range of outcome measures. Deaf children who are immersed in "auditory-oral" communication environments after implantation do much better on a wide range of clinical tests of speech and language development than deaf children who are enrolled in "total communication" programs (Kirk, Pisoni, & Miyamoto, 2000). Auditory-oral communication approaches emphasize the use of speech and hearing skills and actively encourage children to produce spoken language to achieve optimal benefit from their implants. Total communication approaches employ the simultaneous use of some form of manual-coded English (i.e., Signed-Exact English) along with speech to help the child acquire language using both sign and spoken language inputs. The differences in performance between groups of children who are immersed in auditory-oral or total communication education settings are observed in both receptive and expressive language tasks that involve the use of phonological coding and rapid phonological processing skills such as open-set spoken word recognition, language comprehension and measures of speech production, especially measures of speech articulation and intelligibility, expressive and receptive language development and nonword repetition skills (Pisoni et al., 2000).

Preimplant Predictors. Until recently, clinicians and researchers were unable to find reliable preimplant predictors of outcome and success with a CI (see, however, Bergeson & Pisoni, 2004; Horn et al., 2005 a,b; Tait, Lutman & Robinson, 2000). The absence of preimplant predictors is a theoretically significant finding because it suggests that many complex interactions take place between the newly acquired sensory capabilities of a child after a period of auditory deprivation, properties of the language-learning environment and various interactions with parents and caregivers that the child is exposed to after implantation. More importantly, however, the lack of reliable preimplant predictors of outcome and benefit makes it difficult for clinicians to identify those children who may be at risk for poor outcomes with their CI at a time in perceptual and cognitive development when changes can be made to modify and improve their language processing skills.

Learning, Memory and Development. Finally, when all of the outcome and demographic measures are considered together, the available evidence strongly suggests that the underlying sensory, perceptual and cognitive abilities for speech and language "emerge" after implantation. Performance with a CI improves over time for almost all children. Success with a CI therefore appears to be due, in part, to perceptual learning and exposure to a language model in the environment. Because outcome and benefit with a CI cannot be predicted reliably from conventional clinical audiological measures obtained before implantation, any improvements in performance observed after implantation must be due to sensory and cognitive processes that are linked to maturational changes in neural and cognitive development (see Sharma, Dorman & Spahr, 2002).

Although traditional demographic factors are associated with a large portion of the variance in outcomes, there are still substantial gaps in our basic knowledge of how the electrical stimulation provided by a CI works in the brain. Moreover, several other neurocognitive factors related to the "information processing" capacities of the children have also been found to contribute to outcome. These cognitive information processing factors involve the sensory and perceptual encoding of speech, the storage, maintenance and processing of phonological and lexical information in short-term memory and the coordination, integration and connectivity of multiple brain systems as well as response output processes.

Our current working hypothesis about the source of individual differences in outcome following cochlear implantation is that while some proportion of the variance in performance is associated with peripheral factors related to audibility and the initial sensory encoding of the speech signal into "information-bearing" sensory channels in the auditory nerve, several additional sources of variance are associated with more central cognitive and linguistic factors that are related to perception, attention, learning, memory, and cognitive control. How a deaf child uses the initial sensory input from the CI and the way the environment modulates and shapes language development are fundamental research problems in cognitive neuroscience and cognitive psychology. These problems deal with sensory and perceptual encoding, verbal rehearsal, storage and retrieval of phonetic and phonological codes and the transformation and manipulation of phonological and neural representations of the initial sensory input used in a wide range of language and neuropsychological processing tasks. In addition to these issues which are related directly to language and language processing activities, there are also a set of additional questions that deal with the organization and integration of sensory and motor information from multiple brain regions and the processes involved in coordination and interconnectivity of these neural systems.

Moreover, as summarized in the sections below, several converging sources of evidence suggest that other neural systems and circuits secondary to deafness and hearing loss may also be disturbed by the absence of sound and auditory stimulation early in development before implantation takes place. Because of the rich interconnections of sensory and motor systems and auditory and visual signals in the brain, there are additional reasons to suspect that the absence of sound and delays in language during early development produce effects on processes that are not necessarily related to the early sensory processes of hearing and audition. These processes are uniquely associated with the development of neural circuits in the frontal cortex that are involved with executive function and cognitive control processes, such as allocation of conscious attention and control, self-regulation, monitoring of working memory, temporal coding of patterns, particularly memory for sequences and temporal order information, inhibition, planning and problem solving and the ability to act on and make use of prior knowledge and experiences in the service of perception, learning, memory and action.

To investigate individual differences and the sources of variation in outcome, we began by analyzing a set of data from a long-term longitudinal project on CIs in children (see Pisoni et al., 1997; 2000). Our first study was designed to study the "exceptionally" good users of CIs—the so-called "Stars." These are the children who did extremely well with their CIs after only two years of implant use. The "Stars" acquired spoken language quickly and easily and appeared to be on a developmental trajectory that parallels hearing children although delayed a little in time (see Svirsky et al., 2000). The theoretical motivation for initially studying the exceptionally good children was based on an extensive body of research on "expertise" and "expert systems" theory in the field of cognitive psychology (Ericsson & Smith, 1991). Many novel insights have come from studying expert chess players, radiologists and other individuals who have highly developed skills in specific knowledge domains.

Correlations Among Outcome Measures. The results of these analyses revealed that the exceptionally good performers did well on measures of speech feature discrimination, spoken word recognition and language comprehension. They also did well on other tests of receptive and expressive language, vocabulary knowledge and speech intelligibility (see Pisoni et al., 1997; 2000). Until our investigation of the exceptionally good CI users, no one had studied individual differences in outcome in this clinical population or investigated the underlying perceptual, cognitive and linguistic processes.

To assess the relations between these different clinical tests, we carried out a series of correlations on the speech perception scores and several of the other outcome measures. We were interested in whether a child who performs exceptionally well on the Phonetically Balanced Kindergarten

test (PBK) (Haskins, 1949) also performs exceptionally well on other tests of speech feature discrimination, word recognition and comprehension? Is the exceptionally good performance of these children restricted only to open-set word recognition tests or is it possible to identify a common underlying variable or core process that can account for the relations observed among the other outcome measures?

Correlations were carried out separately for the "Stars" and "Low-Performers" using the test scores obtained after one year of implant use (see Pisoni et al., 1997; 2000 for the full report). The results revealed a strong and consistent pattern of intercorrelations among all of the test scores for the "Stars." This pattern was observed for the speech perception tests as well as vocabulary knowledge, receptive and expressive language and speech intelligibility. The outcome measures that correlated the most strongly and most consistently with the other tests were scores on the Lexical Neighborhood Test (LNT), another open-set spoken word recognition test (Kirk, Pisoni & Osberger, 1995).

The finding that performance on open-set spoken word recognition was strongly correlated with all of the other outcome measures is theoretically important because it suggested that the pattern of intercorrelations among all these dependent measures reflects a shared common underlying source of variance. The extremely high correlations with the open-set word recognition scores on the LNT suggested that the common source of variance may be related to the perception and processing of spoken words, specifically, the rapid encoding, storage, retrieval and manipulation of the phonological representations of spoken words in working memory.

Process measures of performance that assess what a child does with the sensory information provided by his/her CI were not part of the standard research protocol used in our longitudinal study so it was impossible at that time to examine differences in information processing capacity, speed, learning, memory, attention or cognitive control (see Pisoni, 2000). It is very likely that fundamental differences in processing capacity and speed are responsible for the individual differences observed between these two groups of children. Differences in learning, memory, attention and cognitive control may also contribute to the variance in outcome and benefit. These types of measures are not routinely collected at most CI centers as part of the routine clinical assessment of CI patients.

For a variety of theoretical reasons, we redirected our research efforts to study "working memory" in deaf children with CIs. One reason for pursuing this particular research direction is that working memory processes have been shown to play a central role in human information processing (Cowan, 2005). Working memory serves as the primary "interface" between sensory input and stored knowledge and procedures in long-term memory. Another reason is that working memory has also been found to be a major source of individual differences in processing capacity across a wide range of information processing domains from perception to memory to language (Ackerman, Kyllonen & Roberts, 1999; Baddeley, Gathercole, & Papagno, 1998; Carpenter, Miyake & Just, 1994; Gupta & MacWhinney, 1997; see Bavelier, Supalla, & Newport, in press).

Process Measures of Performance

Immediate Memory Capacity. Measures of immediate memory capacity were obtained from a group of 176 deaf children following cochlear implantation in a study carried out in collaboration with Ann Geers and her colleagues at Central Institute for the Deaf (CID) in St. Louis (Geers, Brenner & Davidson, 2003; Pisoni & Geers, 2001). Geers et al. had a large-scale clinical research project already underway and they collected a large number of different outcome measures of speech, language and

reading skills from 8 and 9 year old children who had used their CIs for at least three and one-half years. Thus, chronological age and length of implant use were controlled in their study.

Using the test lists and procedures from the WISC III (Wechsler, 1991), forward and backward auditory digit spans were obtained from four groups of 45 deaf children who were tested separately during the summers of 1997, 1998, 1999 and 2000. Forward and backward digit spans were also collected from an additional group of 45 age-matched hearing 8- and 9- year old children who were tested in Bloomington, Indiana, and served as a comparison group.

The WISC-III memory span task requires the child to repeat back a list of digits that is spoken live-voice by an experimenter at a rate of approximately one digit per second (WISC-III Manual, Wechsler 1991). In the "digits-forward" condition, the child was required to repeat the list as heard. In the "digits-backward" condition, the child was told to "say the list backward." In both subtests, the lists begin with two items and increase in length until a child gets two lists incorrect at a given length, at which time testing stops. Points are awarded for each list correctly repeated with no partial credit for incorrect recall.

A summary of the digit span results for all five groups of children is shown in Figure 1. Forward and backward digit spans are shown separately for each group. The children with CIs are shown in the four panels on the left by year of testing; the hearing children are shown on the right. Each child's digit span in points was calculated by summing the number of lists correctly recalled at each list length.

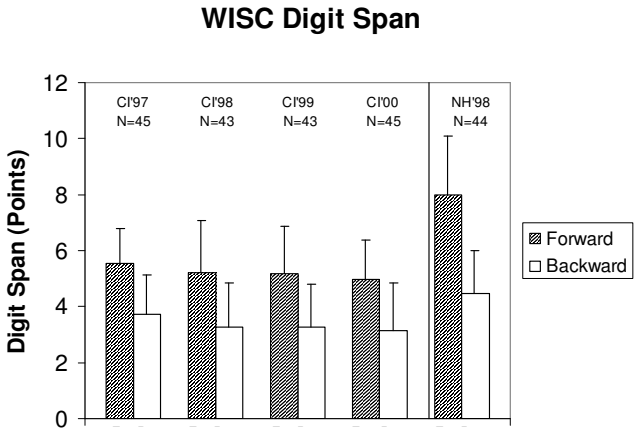


Figure 1. WISC digit spans scored by points for the four groups of 8- and 9-year old children with cochlear implants and for a comparison group of 8- and 9-year-old hearing children. Forward digit spans are shown by the shaded bars, backwards digit spans by the open bars. Error bars indicate one standard deviation from the mean (Adapted from Pisoni & Cleary, 2003).

The forward and backward digit spans obtained from the group of age-matched hearing children are shown in the right-hand panel of Figure 1. These results show that the digit spans for the hearing children differ in several ways from the spans obtained from the children with CIs. First, both forward and backward digit spans are longer for the hearing children than the children with CIs. Second, the forward digit span for the hearing children is much longer than the forward digit spans obtained from the children with CIs. This latter finding is particularly important because it demonstrates for the first time

that the short-term immediate memory capacity of deaf children with CIs is atypical and suggests several possible differences in the underlying processing mechanisms that are used to encode and maintain verbal information in immediate memory (Pisoni & Cleary, 2003; Pisoni & Geers, 2001).

Numerous studies have suggested that forward digit spans reflect coding strategies related to phonological processing and rehearsal mechanisms used to maintain verbal information in short-term memory for brief periods of time before retrieval and output response. Differences in backward digit spans, on the other hand, are thought to reflect the contribution of controlled attention and the operation of higher-level "executive" processes that are used to transform and manipulate verbal information for later processing operations (Rosen & Engle, 1997; Rudel & Denckla, 1974).

The digit spans for the hearing children shown in Figure 1 are age-appropriate and fall within the published norms for the WISC III. However, the forward digit spans obtained from the children with CIs are atypical and delayed and suggest possible differences in encoding and/or verbal rehearsal processes used to maintain phonological information in immediate memory. These differences may cascade and influence other information processing tasks that make use of working memory and verbal rehearsal processes. Because all of the clinical tests that are routinely used to assess speech and language outcomes in this clinical population rely heavily on component processes of working memory, verbal rehearsal and cognitive control, it seems reasonable to assume that these tasks will also reflect variability due to basic differences in immediate memory and processing capacity.

Correlations with Digit Spans. To learn more about the differences in auditory digit span and the limitations in processing capacity, we examined the correlations between forward and backward digit spans and several traditional speech and language outcome measures that were also obtained from these children as part of the larger clinical project at CID (see Pisoni & Cleary, 2003). Of the various demographic measures available, the only one that correlated strongly and significantly with digit span was the child's communication mode. Children who were in educational environments that primarily emphasized auditory-oral skills displayed longer forward digit spans than children who were in total communication environments. However, the correlation between digit span and communication mode was highly selective in nature because it was restricted only to the forward digit span scores; the backward digit spans were not correlated with communication mode or with any of the other demographic variables.

Digit Spans and Spoken Word Recognition. Although these results indicate that early experience and activities in an educational environment that emphasizes auditory-oral language skills is associated with longer forward digit spans and increased capacity of working memory, without additional converging measures of performance, it is difficult to identify precisely what specific information processing mechanisms are actually affected by early experience and which ones are responsible for the increases in forward digit spans observed in these particular children.

Several studies of hearing children have demonstrated close "links" between working memory and learning to recognize and understand new words (Gathercole et al., 1997; Gupta & MacWhinney, 1997). Other research has found that vocabulary development and several other important milestones in speech and language acquisition are also associated with differences in measures of working memory, specifically, measures of digit span, which are commonly used as estimates of processing capacity of immediate memory (Gathercole & Baddeley, 1990).

To determine if immediate memory capacity was related to spoken word recognition, we correlated the WISC forward and backward digit span scores with three different measures of spoken

word recognition that were obtained from the same children. A summary of the correlations between digit span and the spoken word recognition scores based on these 176 children is shown in Table II.

Table II

Correlations between WISC digit span and three measures of spoken word recognition (Adapted from Pisoni & Cleary, 2003).

	Simple Bivariate Correlations	
	WISC Forward Digit Span	WISC Backward Digit Span
Closed Set Word Recognition (WIPI)	.42***	.28***
Open Set Word Recognition (LNT-E)	.41***	.20**
Open Set Word Recognition in Sentences (BKB)	.44***	.24**

*** p <.001, ** p<.01

The Word Intelligibility by Picture Identification Test (WIPI) is a closed-set test of word recognition in which the child selects a word's referent from among six alternative pictures (Ross & Lerman, 1979). The LNT is an open-set test of word recognition and lexical discrimination that requires the child to imitate and reproduce an isolated word (Kirk et al., 1995). Finally, the BKB is an open-set word recognition test in which key words are presented in short meaningful sentences (Bench, Kowal & Bamford, 1979).

Table II displays the simple bivariate correlations of the forward and backward digit spans with the three measures of spoken word recognition. The correlations for both the forward and backward spans reveal that children who had longer WISC digit spans also had higher word recognition scores on all three word recognition tests. This finding was observed for both forward and backward digit spans. The correlations are all positive and reached statistical significance.

These results demonstrate that children who have longer forward WISC digit spans also show higher spoken word recognition scores; this relationship was observed for all three word recognition tests even after other contributing sources of variance were removed. The present results suggest a common source of variance that is shared between forward digit span and measures of spoken word recognition that is independent of other mediating factors that have been found to contribute to the variation in these outcome measures.

Digit Spans and Verbal Rehearsal Speed. While the correlations of the digit span scores with communication mode and spoken word recognition suggest fundamental differences in encoding and rehearsal speed which are influenced by the nature of the early experience a child receives, measures of immediate memory span and estimates of information processing capacity are not sufficient on their own to identify the specific underlying information processing mechanism responsible for the individual differences. Additional converging measures are needed to pinpoint the locus of these differences more precisely. Fortunately, an additional set of measures was obtained from these children for a different purpose and made available for several new analyses.

As part of the research project, speech production samples were obtained from each child to assess their speech intelligibility and measure changes in articulation and phonological development

following implantation (see Tobey et al., 2000). The speech samples consisted of three sets of meaningful English sentences that were elicited using the stimulus materials and experimental procedures originally developed by McGarr (1983) to measure intelligibility of deaf speech. All of the utterances produced by the children were originally recorded and stored digitally for playback to groups of naïve adult listeners who were asked to transcribe what they thought the children had said. In addition to the speech intelligibility scores, the durations of the individual sentences in each set were measured and used to estimate each child's speaking rate.

The sentence durations provided a quantitative measure of a child's articulation speed which we knew from a large body of earlier research in the memory literature was closely related to speed of subvocal verbal rehearsal (Cowan et al., 1998). Numerous studies over the past 30 years have demonstrated strong relations between speaking rate and memory span for digits and words (for example, Baddeley, Thompson & Buchanan, 1975). The results of these studies with hearing children and adults suggest that measures of an individual's speaking rate reflect articulation speed and this measure can be used as an index of rate of covert verbal rehearsal for phonological information in working memory. Individuals who speak more quickly have been found to have longer memory spans than individuals who speak more slowly (see Baddeley et al., 1975).

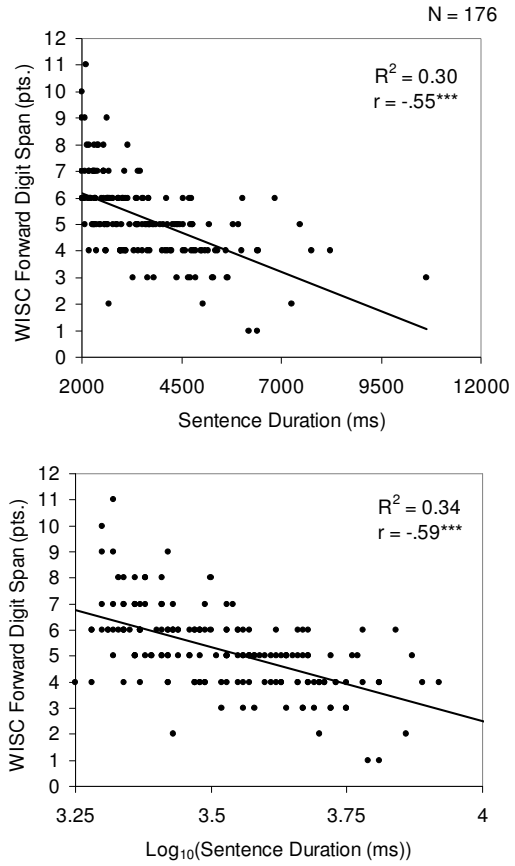


Figure 2. Scatterplots illustrating the relationship between average sentence duration for the seven-syllable McGarr Sentences (abscissa) and WISC forward digit span scored by points (ordinate). Each data-point represents an individual child. Measured duration scores are shown in the top panel, log-transformed duration scores in the bottom panel. R-squared values indicate percent of variance accounted for by the linear relation (Adapted from Pisoni & Cleary, 2003).

A scatterplot of the forward digit span scores for the 168 children are shown in Figure 2 along with estimates of their speaking rates obtained from measurements of their productions of meaningful English sentences. The digit spans are plotted on the ordinate; the average sentence durations are shown on the abscissa. The top panel shows mean sentence durations; the bottom panel shows the log sentence durations. The pattern of results in both figures is very clear; children who produce sentences with longer durations speak more slowly and, in turn, have shorter forward digit spans. The correlations between forward digit span and both measures of sentence duration were strongly negative and highly significant. It is important to emphasize once again, that the relations observed here between digit span and speaking rate were selective in nature and were found only for the forward digit spans. No correlation was observed between backward digit span scores and sentence duration in any of these analyses.

The dissociation between forward and backward digit spans and the correlation of the forward spans with measures of speaking rate suggests that verbal rehearsal speed is the primary underlying factor that is responsible for the variability and individual differences observed in deaf children with CIs on a range of behavioral speech and language tasks. The common feature of each of these clinical outcome measures is that they all make use of the storage and processing mechanisms of verbal working memory (Archibold & Gathercole, 2007).

Verbal Rehearsal Speed and Word Recognition. To determine if verbal rehearsal speed is also related to individual differences in spoken word recognition performance, we examined the correlations between sentence duration and the same three measures of spoken word recognition described earlier. All of these correlations were also positive and suggest once again that a common processing mechanism, verbal rehearsal speed, is the factor that underlies the variability and individual differences observed in these word recognition tasks.

Our analysis of the digit span scores from these deaf children uncovered two important correlations linking forward digit span to both word recognition performance and speaking rate. Both of the correlations with forward digit span suggest a common underlying information processing factor that is shared by each of these dependent measures. This factor reflects the speed of verbal rehearsal processes in working memory. If this hypothesis is correct, then word recognition and speaking rate should also be correlated with each other because they make use of the same processing mechanism. This is exactly what we found. As in the earlier analyses, differences due to demographic factors and the contribution of other variables were statistically controlled for by using partial correlation techniques. In all cases, the correlations between speaking rate and word recognition were negative and highly significant. Thus, slower speaking rates were associated with poorer word recognition scores on all three word recognition tests. These findings linking speaking rate and word recognition suggest that all three measures, digit span, speaking rate and word recognition performance are closely related because they share a common underlying source of variance.

To determine if digit span and sentence duration share a common process and the same underlying source of variance which relates them both to word recognition performance, we re-analyzed the intercorrelations between each pair of variables with the same set of the demographic and mediating variables systematically partialled out. When sentence duration was partialled out of the analysis, the correlations between digit span and each of the three measures of word recognition essentially approached zero. However, the negative correlations between sentence duration and word recognition were still present even after digit span was partialled out of the analysis suggesting that it is processing speed that is the common factor that is shared between these two measures.

The results of these analyses confirm that the underlying factor that is shared in common with speaking rate is related to the rate of information processing, specifically, the speed of the verbal rehearsal process in working memory. This processing component of verbal rehearsal could reflect either the articulatory speed used to maintain phonological patterns in working memory or the time to retrieve and scan verbal information already in working memory or both (see Cowan et al., 1998). In either case, the common factor that links word recognition and speaking rate is the speed of information processing operations used to store and maintain phonological representations in working memory (see Pisoni & Cleary, 2003).

Scanning of Information in Immediate Memory. In addition to our studies on verbal rehearsal speed, we also obtained measures of memory scanning during the digit recall task from a group of deaf children with cochlear implants and a comparison group of typically-developing age-matched hearing children (see Burkholder & Pisoni, 2003; 2006). Our interest in studying scanning of verbal information in short-term memory in these children was motivated by several earlier findings reported by Cowan and his colleagues who have carefully measured the response latencies and interword pause durations during recall tasks in children of different ages (Cowan, 1992; Cowan et al., 1994; 1998).

To investigate scanning of information in short-term memory, we obtained several new measures of speech-timing during immediate recall from a group of deaf children who use CIs (see Burkholder & Pisoni, 2003). Measures of speaking rate and speech timing were also obtained from an age-matched control group of hearing, typically-developing children. Articulation rate and subvocal rehearsal speed were measured using sentence durations elicited with meaningful English sentences. Relations between articulation rate and working memory in each group of children were then compared to determine how verbal rehearsal processes might differ between the two populations. To assess differences in speech timing during recall, response latencies, durations of the test items, and interword pauses were also measured in both groups of children.

For the analysis of the speech-timing measures during recall, we analyzed only the responses from the digit span forward condition. Analysis of the speech-timing measures obtained during recall revealed no differences in the average duration of articulation of the individual digits or response latencies at any of the list lengths. There was no correlation between the average articulations obtained from the forward digit span scores when all children were considered together or when the children were evaluated in groups according to hearing ability or communication mode.

However, we found that interword pause durations in recall differed significantly between the two groups of children. The average of individual pauses that occurred during digit recall in the forward condition was significantly longer in the deaf children with CIs than in the hearing children at list lengths three and four. Although the deaf children with CIs correctly recalled all the items from the three- and four-digit lists, their scanning and retrieval speeds were three times slower than the average retrieval speed of age-matched hearing children (Burkholder & Pisoni, 2003).

The results of this study also replicated our previous findings showing that profoundly deaf children with CIs have shorter digit spans than their hearing peers. As expected, deaf children with CIs also displayed longer sentence durations than hearing children. Total communication users displayed slower speaking rates and shorter forward digit spans than the auditory-oral communication users. In addition to producing longer sentence durations than hearing children, the deaf children with CIs also had much longer interword pause durations during recall. Longer interword pauses reflect slower serial scanning processes which affects the retrieval of phonological information in short-term memory (Cowan, 1992; Cowan et al., 1994). Taken together, the pattern of results indicates that both slower

subvocal verbal rehearsal and slower serial scanning of short-term memory are associated with shorter digit spans in the deaf children with CIs.

The effects of early auditory and linguistic experience found by Burkholder and Pisoni (2003) suggest that the development of subvocal verbal rehearsal and serial scanning processes may not only be related to developmental milestones in cognitive control processes, such as the ability to effectively organize and utilize these two processes in tasks requiring immediate recall. Efficient subvocal verbal rehearsal strategies and scanning abilities also appear to be experience- and activity-dependent reflecting the development of basic sensory-motor circuits used in speech perception and speech production.

Because the group of deaf children examined in the Burkholder and Pisoni (2003) study fell within a normal range of intelligence, the most likely developmental factor responsible for producing slower verbal rehearsal speeds, scanning rates, and shorter digit spans is an early period of auditory deprivation and associated delay in language development prior to receiving a cochlear implant. Sensory deprivation results in widespread developmental brain plasticity and neural reorganization, further differentiating deaf children's perceptual and cognitive development from the development of hearing children (Kaas, Merzenich & Killackey, 1983; Riesen, 1975; Shepard & Hardie, 2001). Brain plasticity affects not only the development of the peripheral and central auditory systems but other higher cortical areas as well both before and after cochlear implantation (Ryugo, Limb, & Redd, 2000; Teoh, Pisoni & Miyamoto, 2004a, b).

Sequence Memory and Learning

All of the traditional methods for measuring memory span and estimating the capacity of immediate memory use recall tasks that require a subject to explicitly repeat back a sequence of test items using an overt articulatory-verbal motor response (Dempster, 1981). Because deaf children may also have disturbances and delays in other neural circuits that are used in speech motor control and phonological development, it is possible that any differences observed in performance between deaf children with CIs and age-matched hearing children using traditional full-report memory span tasks could be due to the nature of the motor response requirements used during retrieval and output. Differences in articulation speed and speech motor control could magnify other differences in encoding, storage, rehearsal or retrieval processes.

To eliminate the use of an overt articulatory-verbal response, we developed a new experimental methodology to measure immediate memory span in deaf children with CIs based on Simon, a popular memory game developed by Milton-Bradley. Figure 3 shows a display of the apparatus which we modified so it could be controlled by a PC. In carrying out the experimental procedure, a child is asked to simply "reproduce" a stimulus pattern by manually pressing a sequence of colored panels on the four-alternative response box.

In addition to eliminating the need for an overt verbal response, the Simon methodology permitted us to manipulate the stimulus presentation conditions in several systematic ways while holding the response format constant. This particular property of the experimental procedure was important because it provided us with a novel way of measuring how auditory and visual stimulus dimensions are analyzed and processed alone and in combination and how these stimulus manipulations affected measures of sequence memory span. The Simon memory game apparatus and methodology also offered us an opportunity to study learning processes, specifically, sequence learning and the relations between working memory and learning using the same identical experimental procedures and response demands (see Karpicke & Pisoni, 2004; Conway et al., 2007a).

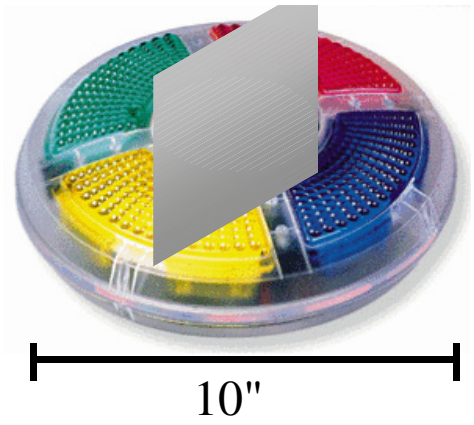


Figure 3. The memory game response box based on the popular Milton Bradley game "Simon."

Simon Sequence Memory Spans. In our initial studies with the Simon apparatus, three different stimulus presentation formats were employed (Cleary, Pisoni & Geers, 2001; Cleary, Pisoni & Kirk, 2002; Pisoni & Cleary, 2004). In the first condition, the sequences consisted only of spoken color names (A). In the second condition, sequences of colored lights (L) were presented in the visual modality. In the third presentation condition, the spoken color names were presented simultaneously with correlated colored lights (A+L).

Forty-five deaf children with CIs were tested using the Simon memory game apparatus. Thirty-one of these children were able to complete all six conditions included in the testing session. They also were able to reliably identify the color-name stimuli used in this task when these items were presented alone in isolation before the experiment began. Thirty-one hearing children who were matched in terms of age and gender with the group of children with CIs were also tested. Finally, 48 hearing adults were recruited to serve as an additional comparison group (see Pisoni & Cleary, 2004).

Of the six conditions tested, three measured immediate memory skills and three measured sequence learning skills. In the immediate memory task, the temporal sequences systematically increased in length as the subject progressed through successive trials in the experiment. Within each condition, the subject started with a list length of one item. If two lists in a row at a given length were correctly reproduced, the next list was increased by one item in length. If a list was incorrectly reproduced, the next trial used a list that was one item shorter in length. Sequences used for the Simon memory game task were generated pseudo-randomly by a computer program, with the stipulation that no single item would be repeated consecutively in a given list. A memory span score was computed for each subject by finding the proportion of lists correctly reproduced at each list length and averaging these proportions across all list lengths.

A summary of the results from the Simon immediate memory task for the three groups of subjects is shown in Figure 4. Examination of the memory span scores for the hearing adults shown in the left-hand panel of Figure 4 reveals several findings that can serve as a benchmark for comparing and evaluating differences in performance of the two groups of children. First, we found a "modality effect" for presentation format. Auditory presentation (A) of sequences of color names produced longer immediate memory spans than visual presentation (L) of sequences of colored lights. Second, we found a "redundancy gain." When information from the auditory and visual modalities was combined together

and presented simultaneously (A+L), the memory spans were longer compared to presentation using only one sensory modality.

The modality effect and the redundancy gains observed with the adults demonstrate subtle differences in the sensory modality used for presentation of the stimulus patterns. As in other studies of verbal short-term memory, longer memory spans were found for auditory stimuli compared to visual stimuli in the hearing adults, suggesting the active use of phonological coding and verbal rehearsal strategies (Penny, 1989; Watkins, Watkins & Crowder, 1974). In addition, the memory spans reflected cross-modal redundancies between stimulus dimensions when the same information about a stimulus pattern was correlated and presented simultaneously to more than one sensory modality (Garner, 1974). This latter finding demonstrates that adults are not only able to combine redundant sources of stimulus information across different sensory modalities, but the consequence of the integration and redundancy gains is an increase in immediate memory capacity when the stimulus dimensions are correlated in the auditory and visual modalities.

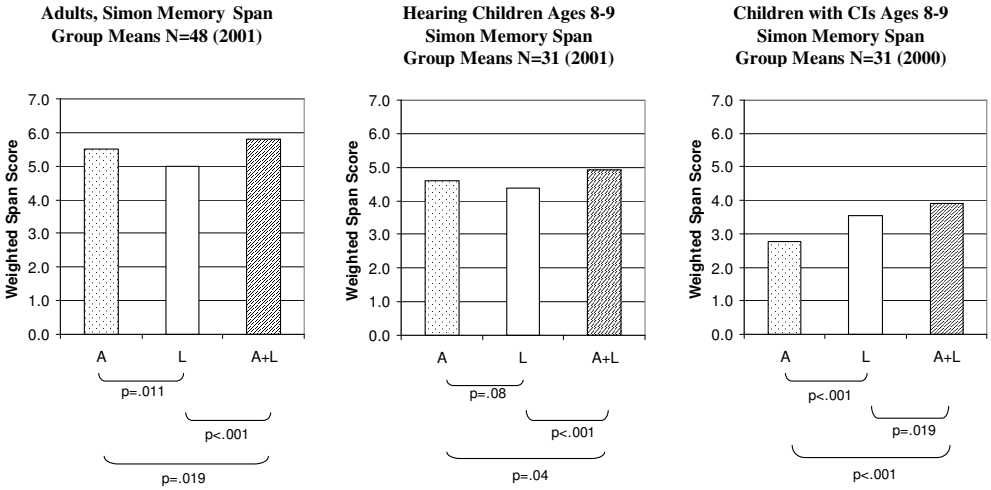


Figure 4. Mean sequence memory spans in each of the three presentation conditions using the "Simon" memory game (Adapted from Pisoni & Cleary, 2004).

The middle panel of Figure 4 shows the results of the same three presentation conditions for the group of hearing 8-and 9-year old children who were age-matched to the group of deaf children with CIs. Overall, the pattern of the Simon memory span scores is similar to the findings obtained with the hearing adults shown in the left-hand panel of Figure 4 although several differences were observed. First, the absolute memory spans for all three presentation conditions were lower for the hearing children than the memory spans obtained from the adults. Second, while the modality effect found with the adults was also present in these data, it was smaller in magnitude suggesting possible developmental differences in the rate and efficiency of verbal rehearsal between adults and children in processing auditory and visual sequential patterns. Third, the cross-modal "redundancy gain" observed with the adults was also found with the hearing children although it was also smaller in magnitude.

The memory spans for the deaf children with CIs are shown in the right-hand panel of Figure 4 for the same three presentation conditions. Examination of the pattern of these memory spans reveals several striking differences from the memory spans obtained for the hearing children and adults. First, the

memory spans for all three presentation conditions were consistently lower overall than the spans from the corresponding conditions obtained for the age-matched hearing children. Second, the modality effect observed in both the hearing adults and hearing children was reversed for the deaf children with CIs. The memory spans for the deaf children were longer for visual-only sequences than auditory-only sequences. Third, although the cross-modal "redundancy gain" found for both the adults and hearing children was also observed for the deaf children and was statistically significant for both conditions, the absolute size of the redundancy gain was smaller in magnitude than the AV gain observed with the hearing children.

The results obtained for the visual-only presentation conditions are of particular theoretical interest because the deaf children with CIs displayed shorter memory spans for visual sequences than the hearing children. This finding adds additional support to the hypothesis that phonological recoding and verbal rehearsal processes in working memory play important roles in perception, learning and memory in these children (Pisoni & Cleary, 2004). Capacity limitations of working memory are closely tied to speed of processing information even for visual patterns which can be rapidly recoded and represented in memory in a phonological or articulatory code for certain kinds of sequential processing tasks. Verbal coding strategies may be mandatory in memory tasks that require immediate serial recall of temporal patterns that preserve item and order information (Gupta & MacWhinney, 1997). Although the visual patterns were presented using only sequences of colored lights, both groups of children appeared to recode these sequential patterns using verbal coding strategies to create stable phonological representations in working memory for maintenance and rehearsal prior to response output.

The deaf children with CIs also showed much smaller redundancy gains under the multi-modal presentation conditions (A+V), which suggests that in addition to differences in working memory and verbal rehearsal, automatic attention processes used to perceive and encode complex multi-modal stimuli are atypical and disturbed relative to age-matched hearing children. The smaller redundancy gains observed in these deaf children may also be due to the reversal of the typical modality effects observed in studies of working memory that reflect the dominance of verbal coding of the stimulus materials. The modality effect in short-term memory studies is generally thought to reflect phonological coding and verbal rehearsal strategies that actively maintain temporal order information of sequences of stimuli in immediate memory for short periods of time (Watkins et al., 1974). Taken together, the present findings demonstrate important differences in both automatic attention and working memory processes in this population. These basic differences in information processing skills may be responsible for the wide variation in the traditional clinical speech and language outcome measures observed in deaf children following cochlear implantation (Cleary, Pisoni & Kirk, 2002).

Simon Sequence Learning Spans. The initial version of our Simon memory game used novel sequences of color names and colored lights (Pisoni & Cleary, 2004). All of the sequences were generated randomly on each trial in order to prevent any learning. Our primary goal was to obtain estimates of working memory capacity for temporal patterns that were not influenced by sequence repetition effects or idiosyncratic coding strategies that might increase memory capacity from trial to trial.

In addition to measuring immediate memory capacity, we have also used the Simon memory game procedure to study sequence learning and investigate the effects of long-term memory on coding and rehearsal strategies in working memory (Cleary & Pisoni, 2001; Conway, Karpicke & Pisoni, 2007; Karpicke & Pisoni, 2004). To accomplish this goal and to directly compare the gains in learning and the increases in working memory capacity to our earlier Simon memory span measures, we examined the effects of sequence repetition on immediate memory span by simply repeating the same pattern over again if the subject correctly reproduced the sequence on a given trial. In the sequence learning

conditions, the same stimulus pattern was repeated on each trial for an individual subject and the sequences gradually increased in length by one item after each correct response until the subject was unable to correctly reproduce the pattern. This change in the methodology provided an opportunity to study nondeclarative learning processes based on simple repetition and to investigate how repetition of the same pattern affects the capacity of immediate memory (see Hebb, 1958; Melton, 1962).

Figure 5 displays a summary of the results obtained in the Simon learning conditions that investigated the effects of sequence repetition on memory span for the same three presentation formats used in the earlier conditions, auditory-only (A), lights-only (L) and auditory+lights (A+L). Examination of the two sets of memory span scores shown within each panel reveals several consistent findings. First, repetition of the same stimulus sequence produced large learning effects for all three groups of subjects. The sequence repetition effects can be seen clearly by comparing the three scores on the right-hand side of each panel of Figure 5 to the three scores on the left-hand side. For each of the three groups of subjects, the learning span scores on the right were higher than the memory span scores on the left.

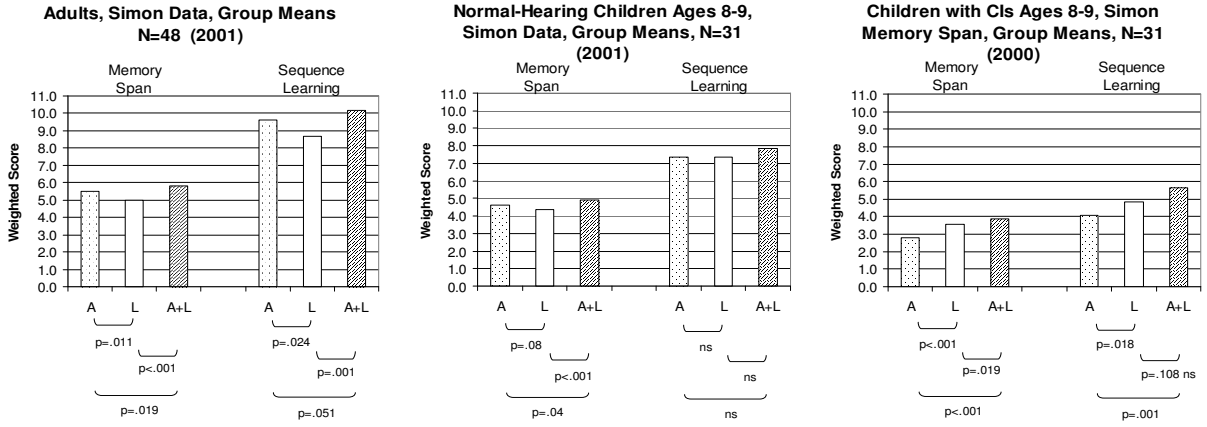


Figure 5. Mean immediate memory spans and sequence learning scores in each of the three conditions tested using the "Simon" memory game (Adapted from Pisoni & Cleary, 2004).

Repetition of a stimulus pattern increased immediate memory span capacity, although the magnitude of the learning effects differed systematically across the three groups of subjects. The memory spans observed for the adults in the learning condition were about twice the size of the memory spans observed when the sequences were generated randomly from trial to trial. Although a repetition effect was also obtained with the deaf children who use CIs in the right panel, the size of their repetition effect was about half the size of the repetition effect found for the hearing children shown in the middle panel of Figure 5.

Second, the rank ordering of the three presentation conditions in the sequence learning conditions was similar to the rank ordering observed in the memory span conditions for all three groups of subjects. The repetition effect was largest for the A+L conditions for all three groups. For both the hearing adults and hearing children, we also observed the same modality effect in learning that was found for immediate memory span. Auditory presentation was better than visual presentation. And, as before, the deaf children also showed a reversal of this modality effect for learning. Visual presentation was better than auditory presentation.

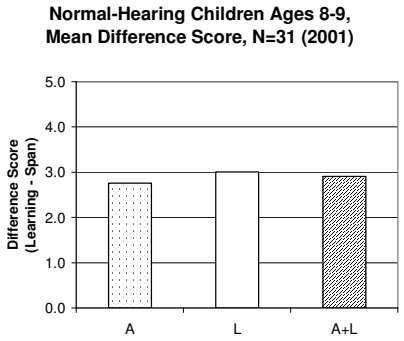


Figure 6. Difference scores between memory and learning for each of the three conditions (A, L, A+L) for the three groups of participants tested using the "Simon" memory game (Adapted from Pisoni & Cleary, 2004).

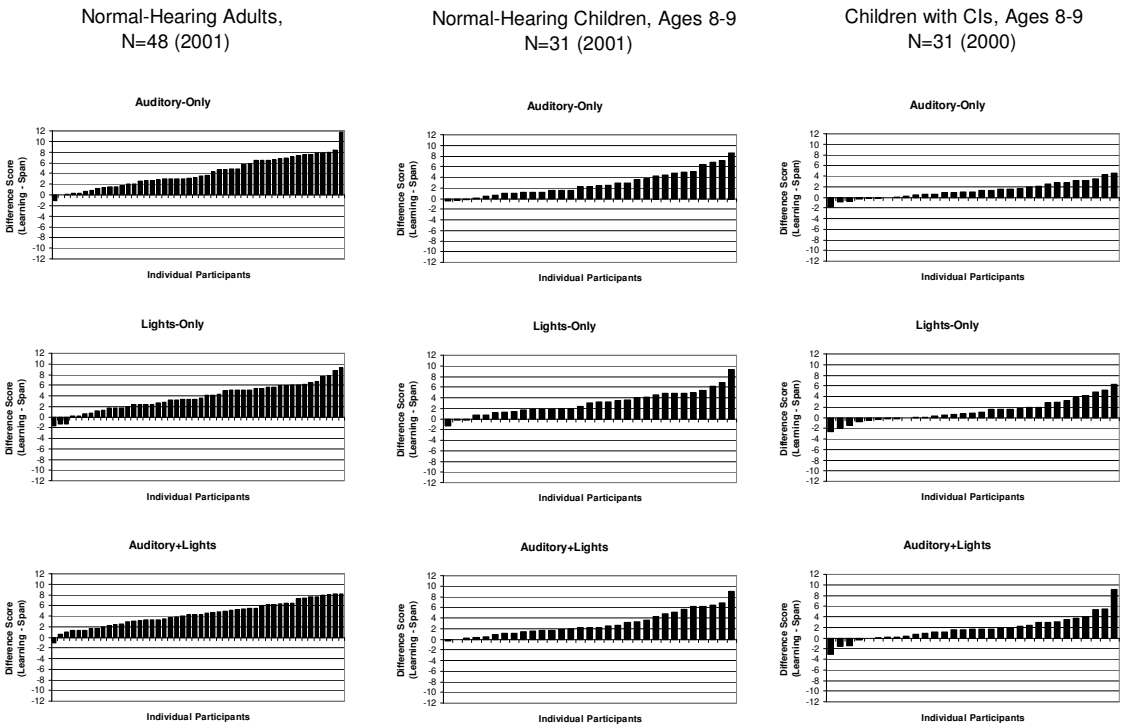


Figure 7. Difference scores for individual subjects showing sequence learning score minus memory span score. Data for the auditory-only (A) condition is shown on the top, lights-only (L) condition in the middle, and auditory-plus-lights (A+L) condition on the bottom. Data from hearing adults are shown on the left, scores for hearing 8- and 9-year-old children in the center, and scores for 8- and 9-year-old cochlear implant users on the right (Adapted from Pisoni & Cleary, 2004).

To assess the magnitude of the repetition learning effects, we computed difference scores between the learning and memory conditions by subtracting the memory span scores from the learning span scores for each subject. The average difference scores for the three groups of subjects are shown in Figure 6, while the data for individual subjects in each group for the three presentation formats are displayed in Figure 7. Inspection of the distributions in Figure 7 reveals a wide range of performance for all three groups of subjects. While most of the subjects in each group displayed some evidence of learning in terms of showing a positive repetition effect, there were a few subjects in the tails of the distributions who either failed to show any learning at all or showed a small reversal of the predicted repetition effect. Although the number of subjects who failed to show a repetition effect was quite small in the adults and hearing children, about one-third of the deaf children with CIs showed no evidence of a repetition learning effect at all and failed to benefit from having the same stimulus sequence repeated on each trial.

Sequence Learning and Outcome Measures. To study the relations between sequence learning and speech and language development in these children, Cleary and Pisoni (2001) computed a series of correlations between the three learning scores obtained from the Simon learning task and several of the traditional audiological outcome measures of benefit that were obtained from these children as part of the larger CID project (see Geers, Nicholas & Sedey, 2003). None of the demographic variables were found to be correlated with any of the Simon sequence learning scores. However, moderate positive correlations were obtained for three measures of spoken word recognition, the WIPI, BKB sentences and the LNT and the auditory-only Simon learning condition. Moreover, the auditory-only Simon learning span was also found to be correlated with the TACL-R measure of receptive language as well as the backwards WISC digit span.

Thus, sequence learning in the auditory-only condition was positively correlated with outcome measures that involve more complex cognitive processing activities that reflect executive functions and controlled attention (Engle, Kane & Tuholski, 1999; Miller & Cohen, 2001). Performance on the TACL-R reflects the ability to comprehend subtle morphological and syntactic distinctions in spoken sentences. Similarly, performance on the backward digit span task assesses the ability to explicitly manipulate the serial order of items actively maintained in working memory. Both of these measures, along with measures of open-set word recognition on the LNT, assess the storage and maintenance of verbal items in short-term memory and the subsequent processing operations of working memory, controlled attention and executive function.

In a follow-up study, Pisoni and Davis (2003) assessed the relations between measures of sequence learning and several speech and language outcome measures with a different group of deaf children who use CIs. They examined two additional measures of sequence learning. The first measure, a redundancy gain learning score, was computed by subtracting the V-weighted span from the AV-weighted span on the Simon learning task in the first interval a child was tested. The difference in performance between the AV and V conditions can be thought of as a measure of how much gain the child received from the addition of redundant auditory information to the visual pattern.

The second measure, a sequence learning gain score, was computed by subtracting the Simon learning span from the first interval a child was tested (for both V and AV conditions) from the span obtained in the last interval a child was tested, and dividing by the total number of years between the scores. This measure of sequence learning was designed to assess changes in the rate of sequence learning over time, while eliminating any baseline differences. Unlike the first learning gain measure, which was used to assess the contribution of redundant auditory information on visual sequence learning,

the second gain measure provided a way to measure changes in sequence memory and learning over time after a period of CI use.

To examine the relationship between these two measures of learning and outcomes, correlations were performed using several traditional speech and language outcome measures. Measures of open-set word recognition (PBK words), sentence comprehension (Common Phrases A, V and AV), vocabulary knowledge (Peabody Picture Vocabulary Test-III (PPVT) (Dunn & Dunn, 1997), language development (Reynell Developmental Language Scales-3rd Edition (RDLS) (Reynell & Huntley, 1985) and Clinical Evaluation of Language Function (CELF) (Semel, Wiig, Secord, 1995), and speech intelligibility (Beginner's Intelligibility Test (BIT) (Osberger, Robbins, Todd & Riley, 1994) were examined. In each of these analyses, the outcome measures were obtained from the first interval a child was tested in using the Simon learning procedure.

A moderate correlation was found between the redundancy gain learning score and the Common Phrases auditory-alone scores, even after controlling for age and length of implant use. Correlational analyses also revealed that the learning gain score was related to the vocabulary knowledge of the child at the first time of testing using the Simon memory game, although the relationship was in different directions for the AV and V conditions. The amount of auditory+visual improvement in learning over time was positively related to the child's initial vocabulary knowledge, while the amount of visual-only gain over time was negatively related. This pattern suggests that greater vocabulary knowledge is associated with better sequence learning skills. Higher PPVT vocabulary scores were associated with increases in AV span and decreases in V span scores.

The results obtained by Pisoni and Davis (2003) showed that measures of sequence learning in deaf children with CIs are associated with changes over time in several clinical outcome measures of speech and language. These findings are of interest both clinically and theoretically because they suggest that the individual differences in outcome of deaf children who receive CIs may also reflect fundamental learning processes that affect the encoding and retention of temporal information in both short-term and long-term memory. Large improvements in immediate reproductive memory span for sequences of visually-presented colored lights were obtained following repetition of a familiar sequence. Differences in the susceptibility to repetition effects were associated with several traditional clinical outcome measures of speech and language.

The findings obtained on learning and memory suggest that differences in the development of basic sequence learning mechanisms in this population may contribute an additional unique source of variance to the overall variation observed in a range of different outcome measures following cochlear implantation. Additional studies of sequence learning and memory in hearing children, adults and deaf children with CIs have been carried out recently and are reported elsewhere (Conway, Karpicke & Pisoni, 2007).

Neuropsychological Measures

Examination of the findings obtained on immediate memory capacity, speed of verbal rehearsal and scanning of items correctly retrieved from short-term memory, suggests that the verbal coding strategies and automatized phonological processing skills of deaf children with CIs are atypical and differ in several significant ways from age-matched typically-developing hearing children. Deaf children with CIs demonstrated shorter forward digit spans, slower verbal rehearsal speeds and significant processing delays in scanning and retrieval of verbal information from short-term memory even for items that were successfully retrieved and correctly recalled. Disturbances were also found in visual sequence memory

and learning. In particular, deaf children with CIs showed significant declines in sensitivity to sequence repetition effects in the Simon learning conditions which suggests fundamental differences in repetition priming, procedural learning and processes involved in encoding and retention of temporal sequences in long-term memory.

The overall pattern of results obtained in these studies is not surprising or unexpected because all of the children were congenitally deaf for some period of time before receiving their CI. What was surprising, however, and what turned out to be both theoretically and clinically significant were the results obtained from the sequence memory and learning experiments using the Simon memory game, especially the findings obtained from the visual-only sequence conditions and the multimodal conditions involving presentation of redundant auditory and visual patterns. The memory and learning results obtained under these two conditions suggest that the effects of deafness and delay in language development, the cognitive and behavioral sequelae following a period of auditory deprivation before implantation, are not modality-specific nor are they restricted to only the perception and processing of auditory signals. The effects of deafness appear to be broader and more global in scope involving the processing of sequences and temporal patterns independently of input modality and the allocation of attentional resources to perceptual dimensions of complex multidimensional stimuli (see Marschark & Wauters, in press; Pelz, in press).

The present findings suggest that multiple information processing systems and the neural circuits underlying their operation are affected by a period of deafness and associated delay in language development prior to implantation. The memory, attention and sequence learning effects observed in these studies are not directly related to the peripheral coding and sensory aspects of hearing or the perception of auditory signals although these factors contribute to establishing and maintaining distinctiveness and discriminability of phonological information at the time of initial encoding and registration in sensory and short-term memory.

It is very likely that many of the deaf children with CIs tested in our studies have other co-morbid disturbances and delays in the development of neural circuits that underlie other information processing systems that are secondary to their profound hearing loss and delay in language development. The absence of sound and auditory experience during early development prior to implantation affects neurocognitive development in a wide variety of ways. Differences resulting from deafness and language delays and subsequent neural reorganization of multiple brain systems may be responsible for the enormous variability observed in speech and language outcome measures following implantation.

One of the new directions our research program has pursued is the investigation of basic elementary neurocognitive abilities of prelingually-deaf children. These are processes that are not specific to hearing, audition or to spoken language processing per se, although they may play important roles in perceiving speech, acquiring spoken language, and developing the underlying sensory-motor abilities and control structures needed for articulation and production of highly intelligible speech and spoken language.

In addition to identifying early predictors of outcome and uncovering additional sources of individual variability, research on elementary neurocognitive factors may provide the theoretical basis for the development of new therapeutic interventions for deaf children who, despite having access to sound with a CI, show significant delays and disturbances in spoken language acquisition and processing. These delays would be especially evident under challenging listening conditions where listeners must rapidly encode and maintain phonological representations of temporal patterns in working memory and monitor and examine the contents of these representations to meet specific task demands.

To explore these findings further, we shifted our research efforts in two new directions. First, we began searching for preimplant predictors of outcome and benefit that did not involve any direct measures of speech or language processing or perception of auditory signals. Second, adopting a broader integrated functional systems approach to brain, behavior and development, we collected several new sets of data using several standardized neuropsychological measures of visual-motor integration, sensory-motor processes as well as executive function and cognitive control so that age-equivalent comparisons can be made based on normative data. Finally, we have recently obtained some preliminary data using the Behavior Rating Inventory of Executive Functions (BRIEF) (Gioia, Isquith, Guy, & Kenworthy, 2000), a behavioral rating inventory filled out by a parent or caretaker to study behavioral regulation, metacognition and executive function in real-world environments outside the clinic and research laboratory. We have also obtained several additional measures of learning, memory and attention using the Learning, Executive, and Attention Functioning (LEAF) (Kronenberger, 2006) and the Conduct-Hyperactive-Attention Problem-Opposition Scale (CHAOS) (Kronenberger, Dunn & Giauque, 1998) rating scales that were developed in our ADHD clinic to assess learning, executive function and attention-hyperactivity. We present a summary of these new findings in the sections below.

Development of Motor Skills. In our research center, as part of the process for determining candidacy prior to implantation, a battery of standardized psychological tests is administered to each child by a clinical psychologist who has extensive experience working with deaf children. These psychological tests include: the Vineland Adaptive Behavior Scales (VABS) (Sparrow, Balla & Cicchetti, 1984), the Bayley Scales of Infant Development (Bayley, 1993), the Beery Visual Motor Integration Scale (VMI) (Beery, 1989), the Weschler Intelligence Scale for Children, Third Edition (WISC-III), the Developmental Assessment of Young Children (DAYC) (Voress & Maddox, 2003) and the Child Behavior Checklist (CBC) (Achenbach & Rescorla, 2005)as well as several additional specialized tests depending on whether the child presents with any developmental disabilities.

Other tests involve parental reports of the children's behavior and adaptive functioning in real-world settings. Historically, these tests were not considered as research data because they were administered prior to implantation and were designed primarily to rule out mental retardation and other developmental disorders that were thought to be possible risks for cochlear implantation. Currently, almost all children who present with a bilateral profound hearing loss at our center are implanted and receive CIs regardless of whether they have any developmental delays or disabilities. Only a small number of children who are medically at risk for surgery are excluded from candidacy.

One of the parental reports used in our psychological assessments is the VABS (Sparrow, Balla & Cicchetti, 1984) which is used to obtain information about the child's adaptive functioning in four functional domains: daily living skills, socialization, motor, and communication. Because the test questions for the communication subscale of the VABS rely heavily on hearing and spoken language skills, they are not considered valid for this clinical population and were excluded from our analyses. However, the other three domains on the VABS provide valuable normative information about the child's adaptive behaviors prior to implantation and offered an opportunity to assess whether a period of profound deafness and language delay prior to cochlear implantation affects adaptive behaviors in these domains.

We examined data for 43 deaf children from the VABS for the motor development, daily living and socialization scales as a function of duration of deafness prior to implantation (Horn et al., 2006). All of the children subsequently received a CI at our center and all of them also provided scores on a range of traditional speech and language outcome measures obtained at several test intervals following

implantation. Because the children in this study received their CIs at different ages, we were able to assess the effects of length of deprivation (i.e., duration of deafness) prior to implantation on these three adaptive behaviors to determine whether these skills developed in an age-appropriate fashion before cochlear implantation.

Children with known or suspected neurological impairment or developmental delay were excluded from the study. Standard scores from hearing tables of the VABS were used to assess preimplant adaptive behavioral functioning. The effects of several demographic variables on VABS standard scores were investigated to determine if preimplant measures of behavioral functioning on the VABS are related to post-implant speech perception and spoken language outcomes following implantation.

For each of the three VABS domains, children were divided into two groups based on a median split. Using this design, spoken language outcomes were compared for each group. If a given VABS domain is predictive of spoken language outcomes after implantation, children in the high group should show higher scores on spoken language measures than children in the low group.

When compared to the results obtained from the daily living skills and socialization domains, the effect of the median split on spoken language outcomes was more robust for the motor domain. Children in the high-motor domain group demonstrated significantly better performance on all spoken language measures than children in the low motor domain group. For the GAEL-P, a closed-set test of spoken word recognition, estimated mean score of children in the high motor domain group was 60.5% words correct compared with 34.1% for children in the low motor domain group. Children in the high motor domain group also demonstrated language and vocabulary skills that were closer to their chronological-age peers than children in the low motor domain group as shown by the differences in mean RDLS-rec, RDLS-exp and PPVT language quotients between the two groups.

We also found that the average motor domain score was age-appropriate and within the typical range of variability compared to the other two domains of the VABS. This finding differs from earlier studies that have reported delays in motor skills of deaf children compared with hearing children. The earlier studies of motor development used children attending residential schools for the deaf who used American Sign Language rather than oral or manual English (Wiegersma & Van der Velde, 1983). Moreover, these studies did not report or control for etiology of deafness or other potential confounding variables such as neurological impairment or age at diagnosis. The present findings suggest that deaf children who present for a CI in infancy or early childhood do not display evidence of general motor impairments, as measured by the VABS.

Multivariate analyses also revealed that nonmotor VABS scores were negatively related to chronological age at testing. Children who were older at the time the VABS data were obtained showed greater delays in socialization and daily living skills than children who were younger. These results suggest that motor development proceeds more typically in these children than other two developmental domains. Because age at testing and duration of auditory deprivation are highly correlated in this population of infants and children, the relations observed between age at testing and VABS domain scores can be recast in terms of duration of auditory deprivation; longer periods of profound deafness before cochlear implantation are associated with greater delays in socialization and daily living skills but not motor development.

One goal of the Horn et al. study was to determine whether preimplant VABS scores could be used to predict post-implant spoken language skills. The results revealed several new preimplant

predictors of spoken language outcomes. Moreover, the pattern of results indicated that not all VABS domains were related to the development of spoken language skills. Motor development was related to performance on spoken word recognition, receptive language, expressive language, and vocabulary knowledge tests obtained over a 3-year period after implantation. Children in the low motor domain group demonstrated poorer spoken word recognition scores and lower age-adjusted language and vocabulary skills than children in the high motor domain group.

Links between motor development and perceptual and linguistic skills have been widely reported in the developmental literature with both hearing and deaf children. In hearing children, motor development assessed in infancy has been shown to be strongly associated with language outcomes in later childhood. The study carried out by Horn, Pisoni et al. (2005) was the first investigation to demonstrate that preimplant measures of motor development predict post-implant language outcomes in profoundly deaf infants and young children who have received a CI.

One explanation of the relations observed between motor development and spoken language acquisition in deaf children with CIs is that motor and language systems are closely coupled in development and share common cortical processing resources that reflect the organization and operations of an integrated functional system used in language processing. This hypothesis is not new. Eric Lenneberg (1967), one of the first theorists to propose a biological explanation for the links between motor and language development, argued strongly that correlations between motor and language milestones in development reflected common underlying rates in brain maturation. Recently, a number of studies have explored the basic neural mechanisms behind these links in greater depth (Iverson & Fagan, 2004). These findings suggest an articulatory or motor-based representation of speech in which brain areas traditionally known to be involved in regulating motor behavior are also recruited during language processing tasks (Wilson, 2002 ; Teuber, 1964).

Divergence of Fine vs. Gross Motor Skills. In a follow-up study, Horn et al. (2006) assessed whether gross or fine motor skills on the VABS showed any evidence of a developmental divergence. Three hypotheses were explored. The first hypothesis was that fine motor skills which are conceptually linked to the "complex motor skills" should be delayed relative to the gross motor skills in these children. The second hypothesis was that fine motor skills should be negatively related to length of auditory deprivation: older deaf participants with longer periods of auditory deprivation should show lower fine motor scores than younger deaf participants. The third hypothesis was that gross motor skills should not be related to length of auditory deprivation.

Horn et al. also assessed whether pre-implant measures of fine or gross motor skills predict of spoken language outcomes in prelingually deaf children with CIs. In the earlier VABS paper, Horn et al. found that pre-implant motor development scores were significantly correlated with post-implant scores on tests of word recognition, receptive and expressive language, and vocabulary knowledge. In the second study, fine and gross motor skills were analyzed separately using correlational analyses with several different post-implant spoken language scores.

As in the earlier study, three spoken language outcome measures were collected longitudinally at various times after implantation. The first test assessed closed-set spoken word recognition, the second assessed both receptive and expressive language skills and the third assessed vocabulary knowledge. Correlations between gross motor scores and the three outcome measures were weakly positive while correlations between fine motor scores and the three language outcome measures were more strongly positive. The only correlations to reach significance were between fine motor scores and expressive

language quotients obtained at the 1 year and 2 year post-implant intervals. In contrast, the correlations between gross motor scores and expressive language scores were all lower and non-significant.

The findings from this study reveal a dissociation in development between gross and fine motor skills in prelingually deaf children. Although the average differences for fine and gross motor skills did not differ, the two motor subdomains showed a developmental divergence as a function of chronological age. For gross motor skills, a positive relationship between age and motor development was observed: older deaf children tended to show more advanced gross motor behaviors compared to younger deaf children. In contrast, the opposite trend was observed for fine motor skills: older deaf children tended to show less advanced fine motor behaviors than younger deaf children. Although these findings are correlational, they are consistent with the hypothesis that a period of auditory deprivation and associated language delay affects the development of fine motor skills differently than gross motor skills. In both of these studies, degree of hearing loss and other demographics were partialled out in the correlation analyses.

Horn et al. also found evidence that pre-implant fine motor skills predict post-implant expressive language acquisition. Infants and children with more advanced fine motor behaviors on the VABS prior to implantation demonstrated higher expressive language scores after 1 or 2 years of CI use than children with less advanced fine motor behaviors. In contrast, gross motor skills measured prior to implantation were not related to post-implant expressive language skills. Although the sample sizes in this study were small, the overall trend suggests that pre-implant fine motor skills are better predictors of post-implant spoken language skills than gross motor skills.

The results reported by Horn et al. provide new evidence that fine motor development and spoken language acquisition are closely coupled processes in deaf infants and children with CIs. These findings suggest that a common set of cortical mechanisms may underlie both the control of fine manual motor behaviors and spoken-language processing, especially the development of expressive language skills.

Links Between Visual-Motor Integration and Language. Numerous researchers have recognized that perceptual-motor development and language acquisition are closely linked and develop together in a predictable fashion with several behavioral milestones correlated across systems (Lenneberg, 1967; Locke, Bekken, McMinn-Larson & Wein, 1995; Siegel et al., 1982). In addition to motor development, visual-motor integration skills have also been found to be closely linked to spoken-language development in numerous studies. Traditionally, visual-motor integration is measured using design-copying tasks in which adults and children are asked to copy a series of increasingly complex geometric figures (Beery, 1989). Performance on design copying tasks has been shown to be correlated with language development, reading ability, and general academic achievement in hearing children (Taylor, 1999) as well as deaf children who use American Sign Language (Bachara & Phelan, 1980; Spencer & Delk; 1985).

Several studies have reported that deaf children display atypical performance on visual-motor integration tasks as well as other perceptual-motor tasks involving balance, running, throwing, and figure drawing (Erden, Otman & Tunay, 2004; Savelsbergh, Netelenbos & Whiting, 1991; Wiegersma & Van der Velde, 1983). In fact, more than 50 years ago, Myklebust and Brutten (1953) carried out one of the earliest studies investigating the visual perception skills of deaf children. They found that performance on the marbleboard test which required children to reproduce visual patterns using marbles on a 10x10 grid was significantly lower for deaf children than hearing age-matched controls. They concluded that deafness disturbs the visual perceptual processes required for constructing continuous figures from

models consisting of discrete elements and causes an alteration in the normal response modes of the organism including disruptions in visual perceptual organization. Myklebust and Brutten (1953) argued that deafness should not be viewed as an isolated autonomous sensory-perceptual impairment but rather as a modification of the total reactivity of the organism.

Many of these early studies included deaf children who had other neurological and cognitive sequelae. And, all of the earlier studies were conducted before deaf children could be identified at birth through universal newborn hearing screening (NIH, 1993). Other studies tested deaf children who were immersed in a manual language environment in which auditory-oral spoken language skills were not emphasized. Thus, the results from these earlier studies cannot be generalized easily to the current population of prelingually deaf children who present for a CI. Two recent studies carried out in our center by Horn et al. (2005, 2006) addressed several questions about the development of visual-motor integration skills.

In the first study, the Beery Test of Visual Motor Integration (VMI-Beery, 1989) was administered prior to implantation to 42 who children were identified from the large cohort of pediatric CI patients followed longitudinally at our center. The Beery VMI test contains a sequence of 24 geometric forms of increasing complexity ranging from a simple vertical line to a complex three-dimensional star. Children are asked to copy each item as accurately as they can.

Several clinical spoken-language measures were also obtained at 6-month intervals in this longitudinal study. Open-set word recognition was measured using the PBK test. Sentence comprehension was assessed with the Common Phrases (CP) test (Osberger et al., 1991), using auditory-only, live voice presentation. Speech intelligibility scores were obtained using the Beginner's Intelligibility Test (BIT). Vocabulary knowledge was assessed with the PPVT. Finally, the Reynell Developmental Language Scales (RDLS) was administered to assess receptive and expressive language skills. The receptive scales (RDLS-r) measured 10 skills, including spoken word recognition, sentence comprehension, and verbal comprehension of ideational content. The expressive language scales (RDLS-e) assessed skills such as spontaneous expression of speech and picture description.

The speech and language measures were obtained during the pre-implant period, within 6 months before implantation, and then at 6-month intervals after implantation. Scores were collapsed into one of five intervals of CI use: pre-implant, 1-year post, 2-years post, 3-years post, and 4-years post. The mean pre-implant VMI score for the 40 deaf children was 0.98 which did not differ significantly from the expected mean of 1.0 for hearing children. For all of the language outcome measures, the scores increased significantly as a function of CI use. Moreover, children with higher pre-implant VMI showed higher percent correct scores on the post-implantation word recognition, comprehension and intelligibility tests.

Several new findings were obtained in this study. First, the pre-implant visual-motor integration scores of the deaf children in this study were age-appropriate when compared with the normative data. This result contrasts with earlier reports showing delays in deaf children compared to hearing children (Erden, Otman & Tunay, 2004; Tiber, 1985). The differences may be due to several factors. First, the sample of deaf children used in our studies was likely to have been diagnosed earlier and received earlier audiological and speech-language intervention than the children used in the earlier studies. Second, children with gross cognitive or motor delays were excluded from the present study.

Second, the longitudinal analyses revealed that VMI scores were robust predictors of post-implant outcomes of speech perception, sentence comprehension, and speech intelligibility. Children

with higher pre-implant VMI scores displayed better performance on all of the outcome measures following CI. Higher VMI scores were also associated with larger increases in speech intelligibility scores over time than lower VMI scores. Thus, pre-implant VMI not only predicts overall performance, but it also predicts rate of improvement with CI experience.

VMI was not an independent predictor of expressive and receptive language scores or vocabulary knowledge. One important difference between the PBK, BIT, and CP tests, compared to the language and vocabulary tests is that the former tests are all administered using auditory-only presentation format whereas the latter are administered using the child's preferred mode of communication. It is very likely that the relations observed between visual-motor integration and these three language processing measures are heavily influenced by the specific information processing demands of the task and the degree to which the behavioral tests require the use of controlled attention, working memory and verbal rehearsal strategies.

One limitation of the first VMI study reported by Horn et al. was that the children were only tested at early ages before implantation as part of their initial preimplant psychological assessment. Variability of visual-motor integration skills in prelingually-deaf children and the associations observed with spoken-language outcomes might not be fully realized until children are a little older and have had more experience using their CI. To pursue these questions further, a second study was carried out with prelingually-deaf children who had used their implants for longer periods of time. The Design Copying and Visual-Motor Precision tests from the NEPSY (Korkman, Kirk & Kemp, 1998), a standardized battery of neuropsychological tests widely used in clinical settings to assess neurocognitive functions of children between 3 and 12 years of age, were administered to determine if the preimplant findings obtained in the first study would generalize to other visual-motor tasks obtained post-implantation.

A total of 30 school-aged children, ages 6 to 14 years, were recruited for this study. Criteria for inclusion in the study were: prelingually deaf prior to age 4, implantation prior to age 6 years, and use of a CI for at least two years. Age of implantation ranged from 1 to 6 years. Duration of CI use varied from 3 to 11 years. All of the children were enrolled in mainstream educational environments. Twenty-five participants were in oral educational environments (auditory-verbal or auditory-oral) and five were in total communication environments. All of the children had hearing parents. The measures reported here were collected as part of a larger study investigating neuropsychological functioning, phonological processing, and reading skills in prelingually-deaf children with CIs (Dillon, 2005; Fagan et al., 2007; Horn, Fagan et al., 2007). Each participant was tested in a single 1.5 hour testing session during which several standardized tests of nonverbal development, vocabulary, and spoken-language processing were administered.

Design Copying is very similar to the Beery VMI test used in the first study. This test is a pencil-and-paper test that measures a child's ability to copy two-dimensional geometrical figures of increasing complexity with no time limits. Visual-Motor Precision is a timed maze-tracing task containing two mazes, a Simple Maze and a Complex Maze. Children were instructed to draw a line down the track as fast as they could without crossing the lines or rotating the paper. Composite raw scores for each maze reflected number of errors (number of times the line crossed the track) and speed (time to complete the task). Fewer errors and faster speed contributed to higher raw scores.

Several conventional speech and language outcome measures were also obtained from each child. Open-set word recognition was assessed with the PBK test. The PPVT was administered to assess receptive vocabulary knowledge. The Forward Digit Span and Backward Digit Span subtests of the WISC-III (Wechsler, 1991) were also administered to measure information processing capacity. Forward

span was included to measure immediate memory capacity and verbal rehearsal; backward span was used to measure working memory capacity. Test sentences developed by McGarr (1983) were used to estimate verbal rehearsal speed (Pisoni, & Cleary, 2003; Baddeley et al., 1975). The children were asked to repeat the sentences aloud and their utterances were recorded and then later measured for length of utterance in seconds.

If average Design Copying performance of prelingually-deaf children with CIs is similar to their age-matched hearing peers, we would expect that mean age equivalent score to be close to the mean age of the sample. Mean Design Copying was 8.14 years while the mean age of the sample was 9.13 years. This difference was statistically significant. While most children fell within normal limits, the mean performance on Design Copying was lower than would be expected from a sample of age-matched hearing peers. The same pattern was observed for the Visual-Motor Precision scores.

Correlations were carried out on both sets of visual-motor scores. The only demographic factor found to correlate significantly with these scores was age at implantation. Children who received a CI at an earlier age tended to show higher Design Copying and Visual-Motor Precision scores than children implanted at later ages. Several correlations were also carried out on the language measures. For the correlations that were significant, partial correlations were conducted to control for the effect of age at implantation. Design Copying showed significant correlations with PPVT, PBK and with backward digit-span scores. Each of these relationships remained significant after partial correlations were carried out to control for age at implantation. Visual-Motor Precision scores were also significantly correlated with PBK scores.

Overall, performance on both Design Copying and Visual-Motor Precision tasks was below of the scores reported for hearing peers based on the NEPSY norms. Unlike the first study in which preimplant VMI scores were not significantly below normative data, the present results replicate earlier findings showing that visual-motor integration skills of deaf children are delayed compared to hearing children (Erden, Otman & Tunay, 2004; Tiber, 1985). When administered prior to implantation, it is possible that VMI and design copying tests are not sensitive enough to pick up differences between prelingually-deaf children and hearing peers. It is also possible that visual-motor integration skills display a slower developmental trajectory in prelingually-deaf children, compared to hearing children and, thus, delays in visual-spatial processing skills may only become apparent at later ages.

As in the first VMI study, longer periods of deafness prior to implantation were associated with greater delays on the Design Copying and Visual-Motor Precision. Children implanted at later ages showed lower Design Copying and Visual-Motor Precision standard scores than children implanted at earlier ages. Although the above correlations are not causal, they suggest that a period of auditory deprivation and language delay may lead to atypical development of non-verbal visual-spatial skills such as those assessed in the VMI tests. While recent neuroimaging work has begun to reveal mechanisms of auditory cortical plasticity underlying speech-perception and production outcomes (Lee, D. et al., 2001; Sharma, Dorman, & Spahr, 2002), little is currently known about how non-verbal processes such as visual-spatial coding and sensory-motor processes are affected by a period of profound deafness and delay in language. In a recent paper by H. Lee et al. (2005), increased pre-implant PET activity in frontal and parietal cortex, brain areas involved in behavioral control and visual-spatial processing, was found to be a predictor of post-implant speech perception scores.

One important finding that emerged from this study was that the Visual Motor Precision task was not correlated with the speech perception, vocabulary or Design Copying scores. The absence of a correlation between Visual-Motor Precision scores and backward digit span suggests that verbal working

memory was not strongly recruited during the Visual-Motor Precision task. The Visual Motor Precision test differs in several ways from the VMI and Design Copying tasks. First, Visual Motor Precision involves a tradeoff between speed and accuracy and therefore recruits controlled/executive attention and behavioral inhibition systems much more strongly than the Design Copying task. Further analyses of speed and error measures of the Visual-Motor Precision as a function of age at implantation revealed that children implanted earlier who had higher overall Visual-Motor Precision scores made fewer errors overall but completed the visual mazes more slowly than children who were implanted later.

These findings suggest that early auditory experience not only affects speech perception and language processing skills but it also affects the development of attentional and behavioral inhibition systems. Several investigators have reported that deaf children with CIs show more age-typical performance on visual-only tests of sustained attention than deaf children without CIs who use hearing aids (Quittner, Smith, Osberger, Mitchell, & Katz, 1994; Smith, Quittner, Osberger & Miyamoto, 1998). Sustained attention has also been shown to improve with length of CI use (Horn, Davis, Pisoni & Miyamoto, 2005b) .Furthermore, the ability of prelingually-deaf children with CIs to regulate and delay premature behavioral responses has been shown to increase with CI use and to be related to performance on several spoken-language measures (Horn et al., 2005). The findings obtained with the Visual Motor Precision task provide additional converging support for these earlier findings on the development of attention and behavioral regulation, processes that reflect the operation of cognitive control and executive function.

The studies carried out recently in our center by Horn et al. demonstrate that visual-motor integration skills in prelingually-deaf children are influenced by early auditory and linguistic experience. The findings suggest that early experience and activity affects the development of several basic elementary information-processing operations that are independent of the sensory domain. While the precise underlying neurobiological mechanisms behind these findings are still unclear, the results suggest that working memory, subvocal verbal rehearsal, and behavioral inhibition, neurocognitive processes typically associated with frontal lobe executive function may play important roles in cognitive control and self-regulation used in a wide range of behavioral tasks commonly used to assess speech and language outcomes in both hearing children and deaf children with CIs (see Hauser & Lukomski, in press).

The results reported by Horn et al. also demonstrate that several visual-motor integration tests, such as the Beery VMI, the NEPSY, Design Copying and Visual Motor Precision tests, can be used clinically to predict outcomes following implantation. These standardized neuropsychological tests, which can be easily administered to deaf children because they do not require auditory processing skills, should be considered as potential additions to assessment batteries used with this clinical population both pre- and post-implantation.

Cognitive Control and Executive Function. While the issues of variability and individual differences have been addressed by two previous NIH Consensus Conferences on Cochlear Implants in 1988 and 1995, very little progress has been made in identifying the neurobiological substrates and cognitive processes that are responsible for individual variation in speech and language outcomes. Many deaf children do not have only a hearing loss resulting from a congenital profound deafness. Other neurocognitive systems are also affected by a period of deafness and delay in language development and these may develop in an atypical manner in the absence of sound and auditory experience during early development, especially during the first few years of life.

When compared with findings obtained on behavioral tests with hearing children, our findings suggest that several aspects of executive function and frontal lobe activity may be disrupted or delayed and may underlie the differences we have observed in traditional outcome measures. Executive function is an umbrella term in neuropsychology and cognitive neuroscience that includes several different processing domains such as attention, cognitive control, working memory, and inhibition (see Hauser & Lukomski, in press).

Many cognitive neuroscientists believe that executive function involves using prior knowledge and experience to predict future events and modulate the current contents of immediate memory (Goldman-Rakic, 1988). There is general agreement that several different aspects of executive function play important roles in receptive and expressive language processes via top-down feedback and control of information processing activities in a wide range of behavioral tasks. The study of executive function and frontal lobe processes may provide new insights into the neurobiological and cognitive basis of individual differences following cochlear implantation.

BRIEF, LEAF and CHAOS Rating Scales of Executive Function. We are now engaged in a series of new studies to assess the contribution of executive function and self-regulation in the development of speech and language processes in deaf children following cochlear implantation. To obtain measures of executive function as they are realized in the real-world like home, school or preschool settings, outside the highly controlled conditions of the audiology clinic or research laboratory, we have been using a neuropsychological instrument called the BRIEF (Behavior Rating Inventory of Executive Function) (Psychological Assessment Resources, Inc, 1996). Three different forms of the BRIEF are available commercially with appropriate norms. One form was developed for preschool children (BRIEF-P: 2.0- 5.11 years); another for school-age children (BRIEF: 5-18 years) and finally one was also developed for adults (BRIEF-A: 18-90 years). The BRIEF family of products was designed to assess executive functioning in everyday environments.

The BRIEF and BRIEF-P, the forms we are using, consist of a rating form that is filled out by parents, teachers and daycare providers to assess a child's executive functions and self-regulation. These forms contain rating scales that measure specific aspects of executive function related to inhibition, shifting of attention, emotional control, working memory, planning and organization among others. Scores from these clinical subscales are then used to construct several indexes of behavioral regulation, inhibitory self-control, flexibility and metacognition. Each rating inventory also provides a global executive composite score.

The BRIEF has been shown in a number of recent studies to be useful in evaluating children with a wide spectrum of developmental and acquired neurocognitive conditions although it has not been used yet with deaf children who use cochlear implants (Gioia, Isquith, Kenworthy & Barton, 2002). From our preliminary work so far, we believe that this instrument may provide new measures of executive function and behavior regulation that are associated with conventional speech and language measures of outcome and benefit in this clinical population. Some of these measures can be obtained preimplant and therefore may be useful as behavioral predictors of outcome and benefit after implantation.

Our initial analysis of recent data obtained on the BRIEF from 15 hearing 5-8 year-old children and 12 deaf 5-10 year-old children with CIs revealed elevated scores in the CI group on several subscales (Conway et al., 2007b). The group means on the Behavioral Regulation Index (BRI), Metacognition Index (MCI) and the Global Executive Composite (GEC) scores were all higher for deaf children with CIs than hearing children although none of them fell within the clinically significant range.

Examination of the eight individual clinical subscales showed consistent differences in shifting, emotional control, initiation, working memory, planning and organization and organization of material. The elevated scores on the BRI suggest that a period of profound deafness and associated language delay before cochlear implantation not only affects basic domain-specific speech and language processes but also affects self-regulation and emotional control, metacognitive processes not typically considered to be sequela of deafness and sensory deprivation in this population (see Schorr, 2005). The BRIEF scores from this new study provide additional converging evidence that multiple processing systems are linked together in development and that disturbances resulting from deafness are not domain-specific and restricted only to hearing and processing auditory signals by the peripheral auditory system.

Analysis of the scores obtained on both the LEAF, which was developed to measure executive function in the context of learning environments, and the CHAOS, which was designed to screen for ADHD, and disruptive behavior symptoms, also revealed elevated scores on the clinical subscales for the children with CIs compared to the hearing comparison group. In particular, differences were observed in learning, memory, attention, speed of processing, sequential processing, complex information processing and novel problem solving subscales on the LEAF and attention, hyperactivity and opposition problems on the CHAOS. No differences were observed on the conduct disorder subscale of the CHAOS.

These additional results reflecting real-world behaviors demonstrate the involvement of several parallel information processing systems and neural circuits involved in learning, memory, attention and processing of complex sequential information. Deaf children with CIs show evidence of disturbances in cognitive and emotional control, monitoring behavior, self-regulation, planning and organization. These differences are not isolated domain-specific symptoms but reflect domain-general properties of an integrated system used in language and cognition linking brain function and behavior with the executive control processes that monitor and regulate on-going behavior and social functioning in novel environments where highly robust adaptive behaviors are routinely required.

GENERAL DISCUSSION

We have presented the results from a large number of studies carried out in our center covering a range of information processing domains. In this section, we provide a brief overview and summary of the major findings of these studies and suggest several conclusions about what these findings mean. We then offer several suggestions for how to understand and interpret these diverse findings in terms of both their direct clinical significance and more basic theoretical relevance for understanding and explaining the neurocognitive factors that are responsible for the large individual differences observed in conventional outcome measures of speech and language following cochlear implantation.

What do all of these diverse behavioral measures have in common? At first glance, the diverse pattern of differences observed across these tasks may seem diffuse and anomalous. However, more careful examination reveals they have links in common and show several important similarities with an extensive clinical literature on frontal lobe disturbances and executive dysfunction. These frontal lobe disturbances are associated with differences in controlled attention, monitoring of verbal information in working memory, functional integration, organization and coordination, self-regulation, inhibition, planning, and using prior knowledge and experience to predict future events and actions in the service of speech and language processing as well as other processing domains.

One of the hallmarks of research on CIs is the enormous variability and individual differences in outcome and benefit. Given this problem, which is observed universally at all implant centers around the world, how can we begin to identify the underlying neurobiological and cognitive factors and explain the

heterogeneity in speech and language outcomes? Are there a set of "core" attributes or common "defining features" or are there several different distinct subgroups of CI users? At this point in time, we cannot provide a definite answer to this question, but understanding the sources of variability in outcome has both clinical and theoretical significance and additional research using new methods and experimental techniques will provide answers to these questions.

Some of the best CI users overlap on specific behavioral measures with hearing children on the low end of a distribution of scores. In contrast, other children with CIs do more poorly and get little benefit from their CIs. At the present time, we do not know whether these individual differences lie on a continuum or whether there are specific subtypes of poor users and we do not know what neurocognitive processes and underlying neural circuits are responsible for these differences. Are the low performers simply poor on all outcome measures or is their performance restricted more selectively to only certain subtests and specific domains? These are important problems to explore because basic knowledge and understanding of the sources of variability in outcome will have several direct implications for diagnosis, treatment and assessment.

Theoretical and Clinical Issues. Are the problems observed with poor users "domain-" and "modality-specific," restricted to processing only speech and auditory signals? Or are their disturbances "domain-general" and "amodal" reflecting contributions of common basic elementary information processing operations shared by language and other information processing systems and neuropsychological domains regardless of processing domain or sensory modality. Our findings suggest that some deaf children with CIs have disturbances and delays with both "automatized" processes, ones typically carried out rapidly without conscious awareness or processing efforts, as well as "controlled" processing, operations that require active attention, processing resources and mental effort, working memory, cognitive control and executive function. Similar findings are discussed by Hauser and Lukomski (2008) and Marschark and Wauters (2008) both in press. Some children can adapt and overcome the first problem which is related to encoding and registration of early sensory information by using "controlled" conscious processes but other children may have more difficulty overcoming basic sensory limitations. Children who have delays or disturbances in both processing domains may be at much greater risk for doing poorly with their CI.

Functional Integration of Brain and Behavior. One of the major problems of past research efforts on CIs, especially research on variability and individual differences in outcome, is that the field of CIs has been and continues to be intellectually isolated from the mainstream of research in cognition and neural sciences and is narrowly focused on clinical issues surrounding efficacy and outcomes. CI researchers and clinicians have adopted an approach to hearing loss that ignores the role of functional connectivity and global systems-level integrative processes in speech and language.

There is now a growing consensus among speech scientists and psycholinguists that speech perception and spoken language processing do not take place in isolation and are heavily dependent on the contribution of multiple brain systems. All behavioral responses in any psychological task are a function of long sequences of processing operations. No part of the brain, even for sensory systems like vision and hearing, ever functions in isolation on its own without multiple connections and linkages to other parts of the brain and nervous system. As Nauta (1964) pointed out many years ago "It seems that if we try to discover the ways in which any part of the brain functions, it is only logical to try to find out in what way it acts within the brain as a whole… no part of the brain functions on its own, but only through the other parts of the brain with which it is connected" (p. 125). These observations apply equally well today in terms of research on cochlear implants.

Automatized and Controlled Processing. Our recent findings on deaf children with CIs suggest that in addition to the traditional demographic, medical and educational variables that have been found to predict some proportion of the variance in traditional audiological measures of outcome and benefit, there are several additional sources of variance that reflect the contribution of basic information processing skills commonly used in a wide range of language processing tasks, specifically, those which rely on rapid phonological encoding of speech and verbal rehearsal strategies in working memory and executive function. Thus, some proportion of the variability and individual differences in outcome following cochlear implantation is related to central auditory, cognitive and linguistic factors that reflect how the initial sensory information transmitted by the CI is subsequently encoded and processed and how it is used by the listener in specific behavioral tasks that are routinely used to measure speech and language outcomes and assess benefit.

Can we identify a common factor that links these diverse sets of findings together? A coherent picture is beginning to emerge from all of these results. At least two factors contribute to success with a CI. One factor is the development and efficient use of rapid automatized phonological processing skills. This is a significant contributor above and beyond the traditional demographic, medical and educational variables that have been found to be associated with outcome and benefit following cochlear implantation. Phonological analysis involves the rapid encoding and decomposition of speech signals into sequences of discrete meaningless phonetic segments and the assignment of structural descriptions to these sound patterns that reflect the linguistically significant sound contrasts of words in the target language.

For many years, both clinicians and researchers have considered open-set tests of spoken word recognition performance to be the "gold standard" of outcome and benefit in both children and adults who have received CIs. The reason open-set tests are viewed in this way is because they require several component processes including speech perception, verbal rehearsal, retrieval of phonological representations from short-term memory, and phonetic implementation strategies required for speech production, motor control and response output. All of these subprocesses rely on rapid highly automatized phonological processing skills for analysis and decomposition of the input signal in perceptual analysis and the reassembly and synthesis of these units into action sequences as motor commands and articulatory gestures for output and speech production. All of these open-set tests also load heavily on cognitive control processes and executive function. They require organization and coordination, planning, inhibition, attention, monitoring and manipulation of symbolic phonological representations in working memory and they make extensive use of past experiences and immediate context to predict, modulate and control future behavior.

When prelingually deaf children receive a CI as a treatment for their profound hearing loss, they do not simply have their hearing restored at the auditory periphery. After implantation, they receive novel stimulation to specialized cortical areas of their brain that are critical for the development of spoken language and, specifically, for the development of automatized phonological processing skills that are used to rapidly encode, process and reproduce speech signals linking up sensory and motor systems in new ways. Moreover, many different neural circuits in other areas of the brain also begin to receive inputs from auditory cortex and brainstem and these contribute to the global connectivity and integrative functions linking multiple brain regions in regulating speech and language processes in a highly coordinated manner.

The present set of findings permits us to identify a specific information processing mechanism, the verbal rehearsal process in working memory, that is responsible for the limitations on processing capacity (see also chapters by Marschark & Wauters and Hauser & Lukomski, in press). Processing

limitations are present in a wide range of clinical tests that make use of verbal rehearsal and phonological processing skills to rapidly encode, store, maintain and retrieve spoken words from working memory. These fundamental information processing operations are components of all of the current clinical outcome measures routinely used to assess receptive and expressive language functions. Our findings suggest that the variability in performance on the traditional clinical outcome measures used to assess speech and language processing skills in deaf children after cochlear implantation reflects fundamental differences in the speed of information processing operations such as verbal rehearsal, scanning of items in short-term memory and the rate of encoding phonological and lexical information in working memory.

Controlled Processing and Executive Dysfunction. A second factor uncovered in our research reflects differences in behavioral regulation, cognitive control and executive function, domain-general metacognitive processes that are slow, effortful and are typically thought to be under conscious control of the individual. One of the reasons we have focused our recent research efforts on executive function in deaf children with CIs is that executive functions are domain-general processes that are involved in regulating, guiding, directing and managing cognition, emotion and behavioral response and actions across diverse environments, especially novel contexts where active problem solving skills are typically required. Our recent findings suggest that the sequela of deafness and delay in language are not domain-specific and restricted to only hearing and auditory processing. Other neurocognitive systems display disturbances and these differences appear to reflect the operation of domain-general processes of cognitive control, self-regulation and organization.

Another reason for our interest in cognitive control processes in spoken language processing is that executive function develops in parallel with other aspects of neural development, especially development of neural circuits in the frontal lobe which are densely interconnected with other brain regions. The development of bidirectional connections among multiple brain regions suggests that the development of speech and spoken language processing may be more productively viewed within the broad context of development as an integrated functional system rather than a narrow focus on the development of hearing and the peripheral auditory system.

Moreover, large individual differences have been observed in the development of executive function within and across cognitive, emotional and behavioral domains. Thus, variability in outcome and benefit following implantation may not only reflect contributions from basic domain-specific sensory, cognitive and linguistic processes related directly to the development of hearing, speech and language function but may also reflect domain-general control processes that are characteristic of global cognitive control, emotional regulation and behavioral response and action.

Focusing new research efforts on executive function and frontal lobe disturbances in deaf children with CIs also provides a neurally-grounded conceptual framework for understanding and explaining a diverse set of behavioral findings on attention and inhibition, memory and learning, visual-spatial processing and sensory-motor function, traditional neurocognitive domains that have been studied extensively in other clinical populations that have acquired or developmental syndromes that reflect brain-behavior dysfunctions in these processing systems. Speech and language processing operations make extensive use of these neurocognitive domains and it seems entirely appropriate to include these in any future investigations seeking to understand and explain the basis of variability and individual differences in speech and language outcome following cochlear implantation.

Recent theoretical developments in cognitive neuroscience have established the utility of viewing the development and use of speech and language as embodied processes linking brain, body and world together as an integrated system. There is every reason to believe that these new theoretical views will

provide fundamental new insights into the enormous variability and individual differences in outcome and benefit following cochlear implantation in profoundly deaf children and adults.

Without knowing what specific biological and cognitive factors are responsible for the enormous individual differences in CI outcomes or understanding the underlying neurocognitive basis for variation and individual differences in performance, it is difficult to motivate and select a specific approach to habilitation and therapy after a child receives a CI. Deaf children who are performing poorly with their CIs are not a homogeneous group and may differ in numerous ways from each other, reflecting dysfunction of multiple brain systems associated with congenital deafness and profound hearing loss. Moreover, it seems very unlikely that an individual child will be able to achieve optimal benefits from his/her CI without researchers and clinicians knowing why a specific child is having problems and what particular neurocognitive domains and information processing systems underlie these problems.

References

Achenbach, T.M. & Rescorla, L.A. (2005). *Child Behavior Checklist (CBC)*. Lutz, FL: The Psychological Corp.

Ackerman, P.L., Kyllonen, P.C. & Roberts, R.D. (1999). *Learning and individual differences*. Washington, DC: American Psychological Association.

Archibald, L.M.D. & Gathercole, S.E. (2007). The complexities of complex memory span: Storage and processing deficits in specific language impairment. *Journal of Memory and Language, 57,* 177-194.

Bachara, G. & Phelan, W. (1980). Visual perception and language levels of deaf children. *Perceptual and Motor Skills, 51*, 272.

Baddeley, A., Gathercole, S. & Papagno, C. (1998). The phonological loop as a language learning device, *Psychological Review, 105*, 158-173.

Baddeley, A.D., Thomson, N. & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589.

Ball, G.F. & Hulse, S.H. (1998). Birdsong. *American Psychologist, 53,* 37-58.

Bavelier, D., Supalla, T., & Newport, E.L. (in press). Similar working memory capacity but different serial short-term memory span in signers and speakers: Theoretical and practical implications. In M. Marschark & P. Hauser (Eds.), *Deaf cognition: Foundations and outcomes*. New York: Oxford University Press.

Bayley, N. (1993). *Bayley Scales of Infant Development, 2nd ed.* San Antonio: Harcourt Assessment, Inc.

Beery, K. (1989). *The VMI developmental test of visual motor integration*, *3rd revision ed.* Cleveland: Modern Curriculum Press.

Bench, J., Kowal, A., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology, 13*, 108-112.

Bergeson, T. & Pisoni, D.B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. Calvert, C. Spence & B.E. Stein (Eds.), *Handbook of Multisensory Integration*. Cambridge: MIT Press.

Brooks, L. (1997). Non-analytical concept formation and memory for instances. In E. Rosch & B. Lloyd (eds.), *Cognition and categorization*, Pp. 169-211. Hillsdale, NJ: LEA.

Burkholder, R.A., & Pisoni, D.B. (2006). Working memory capacity, verbal rehearsal speed, and scanning in deaf children with cochlear implants. In P.E. Spencer & M. Marschark (Eds.), *Advances in the Spoken Language Development of Deaf and Hard-of-Hearing Children*, (pp. 328-357). Oxford University Press

Carpenter, P.A., Miyake, A. & Just, M.A. (1994). Working memory constraints in comprehension. In M.A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, (pp. 1075-1122). San Diego: Academic Press.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Cleary, M. & Pisoni, D.B. (2001). Sequence learning as a function of presentation modality in children with cochlear implants. Poster presented at *CID New Frontiers Conference,* St. Louis, MO.

Cleary, M., Pisoni, D.B. & Geers, A.E. (2001). Some measures of verbal and spatial working memory in eight- and nine-year-old hearing-impaired children with cochlear implants. *Ear & Hearing, 22*, 395-411.

Cleary, M., Pisoni, D.B. & Kirk, K.I. (2002). Working memory spans as predictors of spoken word recognition and receptive vocabulary in children with cochlear implants. *The Volta Review, 102,* 259-280.

Conway, C.M., Karpicke, J., Pisoni, D.B. (2007). Contribution of Implicit Sequence Learning to Spoken Language Processing: Some Preliminary Findings with Normal-Hearing Adults. *Journal of Deaf Studies and Deaf Education.*

Cowan, N. (1992). Verbal memory and the timing of spoken recall. *Journal of Memory and Language*, *31*, 668-684.

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press.

Cowan, N., Keller, T., Hulme, C., Roodenrys, S., McDougall, S., & Rack, J. (1994). Verbal memory span in children: Speech timing clues to the mechanisms underlying age and word length effects. *Journal of Memory and Language, 33*, 234-250.

Cowan, N., Wood, N.L., Wood, P.K., Keller, T.A., Nugent, L.D. & Keller, C.V. (1998). Two separate verbal processing rates contributing to short-term memory span. *Journal of Experimental Psychology: General, 127*, 141-160.

Dempster, F.N. (1981). Memory span: sources of individual and developmental differences. *Psychological Bulletin, 89,* 63-100.

Dillon, C. (2005). Phonological processing skills and the development of reading in deaf children who use cochlear implants. *Research on Spoken Language Processing, Technical Report No. 14*: Indiana University, Bloomington, IN.

Dunn, L. & Dunn, L. (1997). *Peabody picture vocabulary test, 3rd edition*. Circle Pines, MN: American Guidance Service.

Engle, R.W., Kane, M.J. & Tuholski, S.W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence and functions of the prefrontal cortex. In Miyake, A. & Shah, P. (Eds.) *Models of working memory: Mechanisms of active maintenance and executive control.* London: Cambridge Press.

Erden, Z., Otman, S. & Tunay, V. (2004). Is visual perception of hearing-impaired children different from healthy children? *International Journal of Pediatric Otorhinolaryngology, 68*, 281-5.

Ericsson, K.A., & Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits*. New York, NY: Cambridge University Press.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D.M.R., Gantz, B.J., Woodworth, G.G. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research, 40*, 183-199.

Garner, W.R. (1974). *The Processing of Information and Structure*. Potomac, MD: Lawrence Erlbaum.

Gathercole, S., & Baddeley, A. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*, 336-360.

Gathercole, S.E., Hitch, G.J., Service, E. & Martin, A.J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology, 33*, 966-979.

Geers, A., Brenner, C., & Davidson, L. (2003). Factors associated with development of speech perception skills in children implanted by age five. *Ear and Hearing, 24*, 24S-35S.

Geers, A., Nicholas, J., & Sedey, A. (2003). Language skills of children with early cochlear implantation. *Ear and Hearing, 24*, 46S-58S.

Gioia, G.A., Isquith, P.K., Guy, S.C., & Kenworthy, L. (2000). *BRIEF™: Behavior Rating Inventory of Executive Function.*

Gioia, G.A., Isquith, P.K., Kenworthy, L. & Barton, R.M. (2002). Profiles of everyday executive function in acquired and developmental disorders. *Child Neuropsychology, 8,* 121-137.

Goldman-Rakic, P.S. (1988). Topography of cognition: Parallel distributed networks in primate association cortex. *Annual Reviews of Neuroscience, 11,* 137-156.

Gupta, P. & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59*, 267-333.

Harnsberger, J.D., Svirsky, M.A., Kaiser, A.R., Pisoni, D.B., Wright, R. & Meyer, T.A. (2001). Perceptual "vowel spaces" of cochlear implant users: Implications for the study of auditory adaptation to spectral shift. *Journal of the Acoustical Society of America, 109,* 2135-2145.

Haskins, H. (1949). A phonetically balanced test of speech discrimination for children. Unpublished Master's Thesis, Northwestern University, Evanston, IL.

Hauser, P.C. & Lukomski, J. (in press). Development of deaf and hard of hearing students' executive functions. In M. Marschark & P. Hauser (Eds.), *Deaf cognition: Foundations and outcomes*. New York: Oxford University Press.

Hebb, D.O. (1961). Distinctive features of learning in the high animal. In J.F. Delafresnaye (Ed.) *Brain mechanisms and learning*. London and New York: Oxford University Press.

Horn, D., Davis, R., Pisoni, D. & Miyamoto, R. (2005a). Behavioral inhibition and clinical outcomes in children with cochlear implants. *Laryngoscope, 115*, 595-600.

Horn, D., Davis, R., Pisoni, D. & Miyamoto, R. (2005b). Development of visual attention skills in prelingually deaf children who use cochlear implants. *Ear and Hearing, 26*, 389-408.

Horn, D., Pisoni, D., Sanders, M. & Miyamoto, R. (2005). Behavioral assessment of pre-lingually deaf children prior to cochlear implantation. *Laryngoscope, 115*, 1603-1611.

Horn, D.L., Pisoni, D.B. & Miyamoto, R.T. (2006). Divergence of fine and gross motor skills in prelingually deaf children: Implications for cochlear implantation. *Laryngoscope, 116*, 1500-1506.

Horn, D.L., Fagan, M.K., Dillon, C.M., Pisoni, D.B. & Miyamoto, R.T. (2007). Visual-motor integration skills of prelingually deaf children: Implications for pediatric cochlear implantation. *The Laryngoscope, 117*, xxx-xxx.

Iverson, J.M. & Fagan, M.K. (2004). Infant vocal-motor coordination: Precursor to the gesture-speech system? *Child Development, 75,* 1053-1066.

Kaas, J. H., Merzenich, M.M., & Killackey, H.P. (1983). The reorganization of somatosensory cortex following peripheral nerve damage in adult and developing mammals. *Annual Review of Neuroscience,* 6, 325-356.

Karpicke, J. D., & Pisoni, D. B. (2004). Using immediate memory span to measure implicit learning. *Memory & Cognition, 32*, 956-964.

Kirk, K.I., Pisoni, D.B., & Miyamoto, R.T. (2000). Lexical discrimination by children with cochlear implants: Effects of age at implantation and communication mode. In Waltzman, S.B., & Cohen, N.L. (Eds.), *Cochlear Implants* (pp. 252-254). New York: Thieme.

Kirk, K.I., Pisoni, D.B. & Osberger, M.J. (1995). Lexical effect on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing, 16*, 470-481.

Konishi, M. (1985). Birdsong: From behavior to neuron. *Annual Review of Neuroscience, 8*, 125-170.

Konishi, M., & Nottebohm, R. (1969). Experimental studies in the ontogeny of avian vocalizations. In R.A. Hinde (Ed.), *Bird Vocalizations* (pp. 29-48). New York: Cambridge University Press.

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A Developmental Neuropsychological Assessment*. China: PsychCorp.

Kronenberger, W.G. (2006). *Learning Executive Attention Functioning (LEAF).* Indianapolis, IN.

Kronenberger, W.G., Dunn, D.W. & Giauque, A.L. (1998). *Conduct-Hyperactive-Attention Problem-Oppositional Scale (CHAOS)*. Indianapolis, IN.

Lee, D., Lee, J., Oh, S.H., Kim, H. et al. (2001). Cross-modal plasticity and cochlear implants. *Nature, 409*, 149-150.

Lee, H. et al. (2005). Preoperative differences of cerebral metabolism relate to the outcome of cochlear implants in congenitally deaf children. *Hearing Research, 203*, 2-9.

Lenneberg, E. (1967). *Biological foundations of language.* New York: John Wiley & Sons.

Locke, J., Bekken, K., McMinn-Larson, L. & Wein, D. (1995). Emergent control of manual and vocal-motor activity in relation to the development of speech. *Brain and Language, 51*, 498-508.

Marler, P., & Peters, S. (1988). Sensitive periods for song acquisition from tape recordings and live tutors in the swamp sparrow, Melospiza georgiana. *Ethology, 77*, 76-84.

Marschark, M. & Wauters, L. (in press). Cognitive foundations of language comprehension and learning by deaf students. In M. Marschark & P. Hauser (Eds.), *Deaf cognition: Foundations and outcomes.* New York: Oxford University Press.

McGarr, N.S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research, 26*, 451-458.

Melton, A.W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior, 2,* 1-21.

Miller, E.K. & Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Reviews in Neuroscience, 24,* 167-202.

Myklebust, H.R. & Brutten, M. (1953). A study of the visual perception of deaf children. *Acta Oto-laryngol, Suppl. 105. P. 126.*

Nauta, W.J.H. (1964). Discussion of 'Retardation and facilitation in learning by stimulation of frontal cortex in monkeys.' In J.M. Warren & K. Akert (Eds.), *The Frontal Granular Cortex and Behavior*. (pp. 125). New York: McGraw-Hill.

NIH. (1988). *Cochlear implants*. NIH Consensus Statement, May 4, Vol. 7.

NIH (1993). *Early identification of hearing impairment in infants and young children*. NIH Consensus Statement, March 1-3, Vol. 11.

NIH. (1995). *Cochlear implants in adults and children*. NIH Consensus Statement 1995 May 15-17; 13,1-30.

Osberger, M.J., Miyamoto, R.T., Zimmerman-Phillips, S., et al. (1991). Independent evaluations of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear and Hearing, 12*, 66-80.

Osberger, M., Robbins, A., Todd, S. & Riley, A. (1994). Speech intelligibility of children with cochlear implants. *Volta Review, 96*, 169-80.

Pelz, J. (in press). Visual gaze as a marker of deaf students' attention during mediated instruction. In M. Marschark & P. Hauser (Eds.), *Deaf cognition: Foundations and outcomes*. New York: Oxford University Press

Penny, C.G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*, 398-422.

Pisoni, D.B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J.W. Mullennix (Eds.) *Talker Variability in Speech Processing* (pp. 9-32). San Diego: Academic Press.

Pisoni, D.B. (2000). Cognitive factors and cochlear implants: Some thoughts on perception, learning, and memory in speech perception. *Ear and Hearing 21*, 70-78.

Pisoni, D.B. & Cleary, M. (2003). Measures of working memory span and verbal rehearsal speed in deaf children after cochlear implantation. *Ear and Hearing, 24*, 106S-120S.

Pisoni, D.B. & Cleary, M. (2004). Learning, memory, and cognitive processes in deaf children following cochlear implantation. In F.G. Zeng, A.N. Popper & R.R. Fay (Eds.), *Springer Handbook of Auditory Research: Auditory Prosthesis*, SHAR Volume X, 377-426.

Pisoni, D.B. & Davis, R.A.O. (2003). Sequence learning as a predictor of outcomes in deaf children with cochlear implants. In *Research on Spoken Language Processing Progress Report No. 26* (pp. 319-330). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D.B. & Geers, A. (2001). Working memory in deaf children with cochlear implants: Correlations between digit span and measures of spoken language processing. *Annals of Otology, Rhinology & Laryngology, 109*, 92-93.

Pisoni, D.B., Svirsky, M.A., Kirk, K.I., & Miyamoto, R.T. (1997). Looking at the "Stars": A first report on the intercorrelations among measures of speech perception, intelligibility, and language development in pediatric cochlear implant users. *Research on Spoken Language Processing Progress Report No. 21* (pp. 51-91). Bloomington, IN: Speech Research Laboratory.

Quittner, A., Smith, L., Osberger, M., Mitchell, T. & Katz, D. (1994). The impact of audition on the development of visual attention. *Psychological Science, 5*, 347-53.

Reynell, J. K. & Huntley. (1985). *M. Reynell Developmental Language Scales*. Windsor, UK: NFER-Nelson.

Riesen, A.H. (1975). *The developmental neuropsychology of sensory deprivation*. New York: Academic Press.

Rosen, V.M. & Engle, R.W. (1997). Forward and backward serial recall. *Intelligence, 25*, 37-47.

Ross, M., & Lerman, J. (1979). A picture identification test for hearing-impaired children. *Journal of Speech and Hearing Research, 13,* 44-53.

Rudel, R.G. & Denckla, M.B. (1974). Relation of forward and backward digit repetition to neurological impairment in children with learning disability. *Neuropsychologia, 12*, 109-118.

Ryugo, D.K., Limb, C.J. & Redd, E.E. (2000). Brain plasticity: The impact of the environment on the brain as it relates to hearing and deafness. In J.K. Niparko, et al. (Eds.), *Cochlear Implants: Principles and Practices,* (pp. 33-56) (eds.). Lippincott Williams & Wilkins, Philadelphia, PA.

Savelsbergh, G., Netelenbos, J. & Whiting, H. (1991). Auditory perception and the control of spatially coordinated action of deaf and hearing children. *Journal of Child Psychology and Psychiatry, 32*, 489-500.

Schorr, E.A. (2005). Social and emotional functioning of children with cochlear implants. Unpublished Master's Thesis, University of Maryland, College Park, MD.

Sharma, A., Dorman, M.F. & Spahr, A.J. (2002) A sensitive period for the development of the central auditory system in children with cochlear implants: Implications for age of implantation. *Ear and Hearing, 23*, 532-539.

Shepard, R.K. & Hardie, N. (2001). Deafness-induced changes in the auditory pathway: Implications for cochlear implants. *Audiology and Neuro-Otology, 6*, 305-318.

Siegel L., Saigal, S., Rosenbaum, P., Morton, R.A., Young, A., Berenbaum, S., & Stoskopf, B. (1982). Predictors of development in preterm and full-term infants: a model for detecting the at risk child. *Journal of Pediatric Psychology, 7*, 135-148.

Semel, Wiig, & Secord. (1995). *CELF.* San Antonio, TX: The Psychological Corp.

Smith, L., Quittner, A., Osberger, M. & Miyamoto, R. (1998). Audition and visual attention: the developmental trajectory in deaf and hearing populations. *Developmental Psychology, 34*, 840-850.

Sparrow, S., Balla, D., & Cicchetti, D. (1984). *Vineland Adaptive Behavioral Scales.* Circle Pines, MN: American Guidance Service.

Spencer, P. & Delk, L. (1985). Hearing-impaired students' performance on tests of visual processing: relationships with reading performance. *American Annals of the Deaf, 134*, 333-337.

Staller, S.J., Pelter, A.L., Brimacombe, J.A., Mecklenberg, D. & Arndt, P. (1991). Pediatric performance with the Nucleus 22-Channel Cochlear Implant System. *American Journal of Otology, 12,* 126-136.

Svirsky, M.A., Robbins, A.M., Kirk, K.I., Pisoni, D.B. & Miyamoto, R.T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science, 11*, 153-158.

Tait, M., Lutman, M.E. & Robinson, K. (2000). Preimplant measures of preverbal communicative behavior as predictors of cochlear implant outcomes in children. *Ear & Hearing, 21*, 18-24.

Taylor, K. (1999). Relationship between visual motor integration skill and academic performance in kindergarten through third grade. *Optometry and Vision Science 76*, 69-73.

Teoh, S.W., Pisoni, D.B. & Miyamoto, R.T. (2004a). Cochlear implantation in adults with prelingual deafness: I. Clinical results. *Laryngoscope, 114,* 1536-1540.

Teoh, S.W., Pisoni, D.B. & Miyamoto, R.T. (2004b). Cochlear implantation in adults with prelingual deafness: II. Underlying constraints that affect audiological outcomes. *Laryngoscope, 114,* 1714-1719.

Teuber, H.L. (1964). The riddle of frontal lobe function in man. In J.M. Warren & K. Akert (Eds.), *The Frontal Granular Cortex and Behavior*. (pp. 410-444). New York: McGraw-Hill.

Tiber, N. (1985). A psychological evaluation of cochlear implants in children. *Ear and Hearing, 6*, 48S-51S.

Tobey, E.A., Geers. A.E., Morchower, B., Perrin, J., Skellett, R., Brenner, C., & Torretta, G. (2000). Factors associated with speech intelligibility in children with cochlear implants. *Annals of Otology, Rhinology and Laryngology Supplement, 185*, 28-30.

Voress, J.K. & Maddox, T. (2003). *Developmental Assessment of Young Children (DAYC).* Lutz, FL: The Psychological Corp.

Watkins, M.J., Watkins, O.C. & Crowder, R.G. (1974). The modality effect in free and serial recall as a function of phonological similarity. *Journal of Verbal Learning and Verbal Behavior, 13*, 430-447.

Waltzman, S.B., Cohen, N.L., Gomolin, R.H., Shapiro. W.H., Ozdamar, S.R. & Hoffman, R.A. (1994). Long-term results of early cochlear implantation in congenitally and prelingually deafened children. *American journal of otology, 15,* 9-13.

Waltzman, S.B., Cohen, N.L., Gomolin, R.H., Green, J.E., Shapiro, W.H., Hoffman, R.A., & Roland, J.T., Jr. (1997). Open-set speech perception in congenitally deaf children using cochlear implants. *American Journal of Otology*, *18,* 342-9.

Wechsler. D. (1991). *Wechsler Intelligence Scale for Children, Third Edition* (WISC-III). San Antonio, TX: The Psychological Corporation.

Wiegersma, P. & Van der Velde, A. (1983). Motor development of deaf children. *Journal of Child Psychology and Psychiatry 24,* 103-111.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review, 9,* 625-636.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Perceptual Learning Under a Cochlear Implant Simulation[1]

Jeremy L. Loebach and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This research supported by NIH NIDCD R01 Research Grant DC00111, and NIH NIDCD T32 Training Grant DC00012 to Indiana University. I wish to thank Althea Bauernschmidt for her assistance in data collection and analysis, without her I would have had to work so much harder! I would also like to thank Luis Hernandez for providing technical assistance and advice in the design and implementation of the experimental procedures.

Perceptual Learning Under a Cochlear Implant Simulation

Abstract. Adaptation to the acoustic world following cochlear implantation does not typically include formal training or extensive audiological rehabilitation. Can cochlear implant (CI) users benefit from formal training, and if so, what type of training is best? This study used a pre/post-test design to evaluate the efficacy and generalization of training in normal hearing subjects listening to CI simulations (8-channel sinewave vocoder). Subjects were trained with words (simple or complex), sentences (meaningful or anomalous), or environmental stimuli, and then were tested using an open-set identification task. Subjects were trained on only one type of material but were tested on all materials. All groups showed significant improvement as a result of training, which successfully generalized to some, but not all stimulus materials. For easier tasks, all types of training generalized equally well. For more difficult tasks, training specificity was observed. Training on speech did not generalize to the recognition of environmental signals; however, training on environmental signals successfully generalized to speech. These data demonstrate that the perceptual learning of degraded speech is highly context-dependent and that the specific stimulus materials that a subject experiences during training have a substantial impact on generalization to new materials.

Introduction

Despite the recent advances in cochlear implant (CI) technology, a large amount of variability in outcome and benefit is consistently reported among CI users. Although differences in etiology, onset, and duration of deafness, age at implantation and physiological factors (electrode insertion depth, availability of viable neurons, etc.) can account for a portion of this variability (NIH, 1995), a considerable amount of variability remains unexplained. The absence of rigorous standardized training regimens confounds the issue at a fundamental level. The experiences of CI users may differ from the start, leading to differences in auditory perceptual learning during adaptation to their prostheses. Could the standardization of training establish a more stable foundation, and allow the dissociation of audiological factors from other neural and cognitive factors? Moreover, what type of training is most effective, and yields the most robust levels of generalization to new materials? The present study seeks to investigate the efficacy of different training regimens in normal hearing subjects using both speech and environmental stimuli that have been processed by a CI simulation.

Cochlear implantation can provide sufficient acoustic input to a deaf individual to allow the establishment of some form of hearing (NIH, 1995). Whereas early implants provided the hope of recovering some auditory ability, most recipients of modern implants have the expectation that they will recover oral communication skills, including the ability to talk on the telephone (Shannon, 2005). In the worst case, patients are expected to regain some awareness of sound (Clark, 2002), including the detection and recognition of environmental signals. While clinicians often cite this benefit as part of the rationale for implantation, the degree to which CI users can actually recognize and identify environmental signals is largely unknown (cf., Reed & Delhorne, 2005).

Research using acoustic simulations of CIs has met with great success. From the earliest simulations of Shannon and colleagues (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995), the effectiveness and utility of acoustic models of CIs has been apparent. The vocoder model of a CI simulates the limited number of spectral channels available in the electrode array by dividing the acoustic

signal using a series of band-pass filters. Band limited noise replaces the spectrum of each band to simulate the effect of wide-band electrical stimulation of each electrode. The amplitude envelope, which is derived from the original bands using a low pass filter, is then used to modulate the noise to simulate the temporal profile of electrical stimulation at each electrode. The result is a signal that acoustically simulates the spectrally degraded conditions that CI users may normally encounter.

The seminal work of Shannon and colleagues using the vocoder demonstrated that high levels of speech recognition persisted despite such radical spectral degradation (Shannon et al., 1995). Using signals with one, two, three or four spectral channels, Shannon demonstrated that when more spectral channels were available to the listeners, higher levels of perceptual identification were observed. A single channel provided sufficient information to allow moderately accurate closed-set recognition of English consonants and vowels (48% correct and 35% correct respectively), and as the number of channels increased so did recognition rates. For vowels and meaningful sentences, asymptotic performance (> 90% correct) was reached with three channels, whereas recognition rates for consonants continued to increase from three to four channels. Consonant recognition was far less robust than vowel recognition due to several factors. When consonants were classified according to the guidelines of Miller and Nicely (1955), perception of manner and voicing cues reached asymptote with just 2 spectral channels (> 90% correct identification), as compared to classification based on place of articulation, which never exceeded 60% correct even with four channels. These data demonstrate the robust nature of the human perceptual system, which can perform well even when spectral information is severely limited, so long as sufficient temporal information is preserved (Shannon et al., 1995).

Follow-up studies further refined the methodology, expanding the number of channels available and altering the carrier used. When the maximum number of spectral channels available was incrementally increased from 4 to 9, asymptotic performance was observed for closed-set vowels with 8 channels, and meaningful sentences with 5 channels (Dorman, Loizou & Rainey, 1997). Consonant identification reached asymptote with 6 channels, which was a result of increased accuracy in identifying place of articulation, which also reached asymptote at 6 channels (Dorman et al., 1997). Moreover, the type of carrier used did not appear to have an adverse effect on performance. In their original study, Shannon and colleagues used a noise vocoder, in which white noise was used to remove the spectral detail from each band (Shannon et al., 1995). The anecdotal reports of CI users, however, were not of hearing bursts of noise, but of hearing “beep tones” (Dorman et al., 1997), raising the question of whether noise is the most appropriate carrier to use (Dorman et al., 1997). Using a sinewave vocoder, which replaces the spectral detail of each band with a sinusoid anchored at the band center, Dorman and colleagues demonstrated that performance did not differ from that observed using the noise vocoder (Dorman et al., 1997). Moreover, the performance of CI users on consonants and vowels was similar to that of normal hearing subjects listening to six channel stimuli, demonstrating that the vocoder can successfully simulate the output of a CI in order to elicit equivalent levels of performance (Dorman & Loizou, 1998).

Although studies using the noise and sinewave vocoders have focused primarily on the identification of linguistic content (e.g., isolated consonants and vowels), the real world is composed of many other complex auditory events that are transmitted via the acoustic signal. Compared to speech, considerably less is known about the perception of environmental sounds, both in the clear and processed by vocoder models. Environmental signals are very useful for neuropsychological and cognitive evaluation because they can assess basic sensory and cognitive capabilities without the added dimensions of linguistic information and context. Although there may be some commonalities between the perceptual systems required for the identification of speech and environmental stimuli, the degree to which they operate independently is unknown. Some cross-modal priming has been observed for environmental

stimuli. When the acoustic presentation of an unprocessed environmental stimulus is paired with the orthographic presentation of the stimulus name during the study phase of an experiment, subjects are faster and more accurate at identifying the stimulus during the test phase as compared to if they saw the name presented without the sound (Chiu & Schacter, 1995). This priming effect is context specific, however. If the exemplar is sufficiently different from the test stimulus, the strength of priming is reduced (Chiu, 2000). For example, if the subject received one exemplar of an environmental stimulus during the study phase (such as the sound of a bird chirping), but was tested with a different exemplar (a different bird chirping), priming was significantly reduced. In speech, the stimulus-specific form can also be preserved in addition to the more abstract symbolic lexical form (Lachs, McMichael & Pisoni, 2003). Thus, at least at a surface level, it appears that environmental stimuli may be encoded in a similar manner to speech.

Although the processing of environmental signals may share some similarity with the neural and cognitive processes used to perceive speech, many of the acoustic (spectral and temporal) characteristics of speech are fundamentally different from environmental signals (see Stevens, 1980 for example). In a series of recent experiments investigating the perceptual identification of environmental signals, Gygi and colleagues trained subjects to identify 70 environmental stimuli in the clear using a three-letter code (Gygi, Kidd & Watson, 2004). They then processed the stimuli using a series of low, high, and band-pass filters and tested subjects over a period of nine days. Overall, they found that both speech and environmental stimuli may share a similar range of critical frequencies that are important for identification. The most important acoustic information for the recognition of environmental stimuli lies between 1200 and 2400 Hz, which is identical to the region identified as crucial for speech under the Articulation Index (Gygi et al., 2004). Moreover, even when the stimuli were low-pass or high-pass filtered at the extremes (300 and 8000 Hz), recognition remained higher than 50% correct (Gygi et al., 2004).

When the stimuli were processed using one and six-channel noise vocoders, the results were more variable. Naïve subjects in both groups showed significant improvement over a two-day period (1-channel: 13% correct on day 1 to 23% correct on day 2; 6-channel: 36% correct on day 1 to 66% correct on day 2), but performance was significantly higher for the 6-channel stimuli (Gygi et al., 2004). Not surprisingly, the stimuli that showed the greatest improvement were those that had broader harmonic structure and spectral detail (Gygi et al., 2004). However, these results should be considered with some caution because certain aspects of performance may be attributable to task familiarity. A group of subjects who were first trained to criterion on the unprocessed stimuli performed significantly better on the 1-channel stimuli than did the naïve subjects (Gygi et al., 2004), which could be attributable to increased experience with the three letter codes rather than to familiarity with the stimuli themselves. In addition, Gygi and colleagues used a closed set recognition task, which constrains the possible choices that subjects can make to those within a specified stimulus set. Subjects could have been systematically eliminating the possible alternatives as they became increasingly familiar with the test set.

Using a slightly different task, Shafiro demonstrated that the reliance on spectral and temporal information in the recognition of environmental stimuli processed with a noise vocoder may be different than is observed for speech (Shafiro, 2004). Sixty environmental stimuli were processed with 2, 4, 8, 16 and 32 channel vocoders, and presented to normal hearing subjects using a Latin square design, such that each subject only heard one version of a stimulus, but all band conditions were presented across all subjects. In general, improved closed set recognition (out of 60) was observed as the number of channels increased. With only 2 channels, performance was low (32% correct), but reached asymptote at 66% correct with 16 channels (Shafiro, 2004). Moreover, the performance depended on the stimulus itself: while some environmental stimuli showed increases in accuracy with the addition of more spectral

channels, others showed decreases (Shafiro, 2004). In particular, stimuli that relied more on spectral information (e.g., church bell, birds chirping) showed increases, whereas those that relied more on temporal information (e.g., clapping, footsteps) showed decreases. Thus, it appears that some environmental stimuli may show an altogether different pattern of spectro-temporal dependence as compared to speech signals.

Relatively few studies have examined the perception of environmental stimuli by CI users. Using a closed-set testing format, Tye-Murray and colleagues assessed the abilities of fourteen CI users to identify 36 environmental stimuli at 1, 9, 18 and 30 months post-implantation (Tye-Murray, Tyler, Woodward & Gantz, 1992). Overall, performance increased significantly over time, from about 32% correct at 1 month, to 38% correct at 9 months, and topping out at 42% correct at 18 months (Tye-Murray et al., 1992). These gradual changes were statistically significant, although far slower than the gains typically observed for speech. A more recent study by Reed and Delhorne (2005) compared environmental sound recognition and NU-6 word identification. Environmental stimuli were organized into four thematic lists of ten stimuli each, and subjects made closed set responses by clicking one of ten buttons presented on a computer screen. Performance of the eleven CI users differed across the four lists of environmental stimuli, with a mean identification score of 79% correct (Reed & Delhorne, 2005). Average performance on the closed set environmental stimulus identification was significantly better than performance on the open set word identification, which was only 39% correct (Reed & Delhorne, 2005). Subjects were divided into high performing and low performing groups based on the median score for word identification (34% correct). High performing subjects (> 34%) performed better at identifying environmental stimuli than did low performing subjects (Reed & Delhorne, 2005). The authors hypothesized that the differences in performance may be due to differences in exposure to environmental stimuli in their daily environment (Reed & Delhorne, 2005). However, it is unclear whether additional exposure or standardized training could increase the performance of the low performing subjects.

One common theme throughout the studies using vocoded signals is the issue of perceptual learning. Even though subjects can accurately identify speech processed by a vocoder, a period of adjustment is frequently required. In the original Shannon study, subjects received 8-10 hours of exposure to the synthesis condition in order to adapt to the stimuli and stabilize their performance (Shannon et al., 1995). Explicit training on the testing materials was used in the studies by Dorman and colleagues in order for subjects to "warm up" to the synthesis condition (Dorman et al., 1997; Dorman & Loizou, 1998). Although some type of auditory training is necessary when adapting to acoustic simulations of CIs, the best and most efficient form that maximizes perceptual learning and promotes robust generalization and transfer to other materials has not been adequately examined.

In a series of recent experiments, Davis and colleagues investigated the use of lexical information during adaptation to 6-channel noise vocoded sentences (Davis, Johnsrude, Hervais-Adelman, Taylor & McGettigan, 2005). Five experiments were conducted in order to examine the mechanisms of perceptual learning. In the first experiment, they assessed whether exposure to the stimulus materials without any feedback results in perceptual learning. Subjects were presented with a set of thirty sentences that were processed with the vocoder and asked to transcribe as much of each as possible. Open set identification increased significantly across the 30 sentences, from 32% correct keyword identification on the first ten sentences to 43% correct on the last 10 sentences. These gains can be attributed to perceptual attunement to vocoded speech, since subjects received no feedback.

The effectiveness of auditory feedback was assessed in Experiment 2. Like Experiment 1, subjects transcribed each sentence; however, after they made their response they were provided with auditory feedback. One group heard the "distorted" sentence followed by the unprocessed version

(DDC), whereas the other group heard the clear sentence followed by the repetition of the distorted version (DCD) in order to elicit stimulus pop out. Both groups showed significant gains as a result of training, increasing from 43% to 73% correct from the first to final 10 sentences for subjects in the DDC group, and from 50% to 77% correct for subjects in the DCD group. Although both groups showed equivalent gains, the group who received the DCD training experienced performed significantly better (Davis et al., 2005).

The typical CI user will not have access to the unprocessed version of a stimulus, however, so in Experiment 3, Davis and colleagues explored whether the addition of the orthographic version of the sentence enhances perceptual learning (Davis et al., 2005). Subjects either received feedback in the form of the repetition of the distorted version of the sentence paired with and without the written transcription. Subjects who were presented with the repetition of the distorted sentence alone showed significant improvement, increasing from 38% correct to 69% correct. Subjects who also received the orthographic form of the stimulus performed significantly better, improving from 50% to 77% correct, a level of performance identical to those subjects in experiment 2 who experienced stimulus pop out. Such comparable improvement suggests that presentation of the orthographic form of the sentence is just as effective as presentation of the original unprocessed acoustic version (Davis et al., 2005).

Although these gains are impressive, one potential factor that could contribute to the results of the first three experiments is the use of contextually constrained meaningful sentences. When Davis and colleagues controlled the amount of lexical information in the sentences, however, the amount of learning varied (Davis et al., 2005). Subjects who were trained with sentences comprised entirely of non-words improved with training, but performed significantly more poorly than those who were trained on meaningful sentences (Experiment 4). This experimental task may be more difficult, however, given that the materials are not valid English words. To examine these effects more in more detail, Davis and colleagues conducted a final experiment that systematically varied the amount of lexical information. Subjects were trained on meaningful sentences, semantically anomalous sentences (sentences where the function words are correctly placed, but the content words are unrelated), non-word sentences, or Jabberwocky sentences (anomalous sentences where the content words are replaced by non-words). All groups showed improvement over the training interval, and two distinct groups emerged based on performance. Subjects who were trained on meaningful and anomalous sentences performed identically to one another, and significantly better than those trained on non-word and Jabberwocky sentences. These findings suggest that access to the syntactic structure may be required in order to elicit effective levels of learning (Davis et al., 2005).

The results reported by Davis and colleagues raise several important questions. Although they demonstrated that feedback significantly influences performance, the type of feedback they used would not necessarily apply to the typical CI user. In an individual with electric hearing, there is never an opportunity for the presentation or repetition of the unprocessed stimulus. The finding that the subjects who received orthographic feedback paired with the vocoded version of the sentence performed just as well as those who received the clear version suggests that such feedback could be useful to CI users. In addition, subjects who did not receive explicit feedback showed significantly lower levels of performance overall, but still showed similar gains due to training.

In a more comprehensive study, Burkholder and colleagues (Burkholder, 2005; Burkholder, Svirsky & Pisoni, submitted 1; 2) demonstrated that the use of feedback consisting of the correct orthographic form of the sentence paired with the repetition of the vocoded stimulus produced significantly greater pre to post-test gains than receiving the unprocessed version alone. Moreover, subjects who were trained on the anomalous sentences showed identical pre to post-test gains as subjects

trained on meaningful sentences, but showed significantly greater benefits during generalization to new materials including environmental stimuli (Burkholder, 2005; Burkholder et al., submitted 1; 2). These data suggest that access to the syntactic structure of the sentence without relying on sentence meaning may provide a greater benefit, presumably because the listener is forced to reallocate attention to the acoustic-phonetic structure of the signal and rely on bottom-up processes for recognition. This point is underscored in the observation that training on speech stimuli successfully generalized to the identification of environmental stimuli.

One limitation of the studies by Burkholder and colleagues is that they only assessed the generalization of training with speech to environmental stimuli, but not the converse. If subjects are relying on the acoustic structure of the stimuli, one would predict that training on environmental stimuli should successfully generalize to speech, an issue that we address in the current work. In addition, no baseline identification data were collected for the environmental stimuli, so it is unknown if the subjects were performing significantly better at identifying the environmental stimuli than with no training at all. Moreover, although training with meaningful sentences appears to generalize to novel sentences, it is unknown whether this training generalizes to single words. Anomalous sentences can be conceptualized as a series of unrelated words connected by a permissible syntactic structure. If this is the case, then training on single word identification should generalize to anomalous sentences and vice versa. In addition, previous studies have shown that training on simple CV and CVCs may produce only modest gains in performance on sentence identification (Fu, Galvin, Wang & Nogaki, 2006). It is unclear whether the converse is true; that is, would training on sentences, both high and low in context, generalize to single words and CVCs?

As there are currently no standard rehabilitation protocols following cochlear implantation, understanding how the perceptual learning of spectrally degraded stimuli transfers to new materials is especially relevant. Evaluating the strengths and weaknesses of different training paradigms is critical for the development of rehabilitation strategies that maximize perceptual learning and promote robust generalization to new materials. The purpose of the present study, therefore, was to examine the effect of training on the recognition of speech and environmental stimuli processed by a sinewave vocoder. Specifically, we assessed the perceptual learning of CVCs, words, meaningful sentences, anomalous sentences and complex non-speech environmental stimuli using a pre/post-test design, and compared the generalization to different materials.

Method

Subjects

One hundred thirty normal-hearing adults from the IU community participated in the study. Of the 130 subjects, 95 were female, 34 were male, and one self reported being transgender. Subjects ranged in age between 18 and 60, with a mean age of 22.7 years. All subjects reported having uncorrected normal hearing and that English was the first language that they learned in infancy. Most subjects (n= 117) were monolingual; although a small number reported being fluent bi- (n= 11) or tri-lingually (n= 2). Subjects were given credit in their Introductory Psychology course for their participation (n= 34), or were paid at the rate of $10 per hour (n= 96).

Of the 130 subjects, five were excluded from the final data analysis. One subject was excluded after reporting that he/she could not hear the stimuli as speech. One subject was excluded due to a program malfunction. After the experiment, one subject revealed that they were not a native English speaker, and so their data were excluded. Two subjects were excluded after the decision was made that

they were not on task: one subject left many spaces intentionally blank and made frequent spelling errors that rendered the data impossible to score, and the other typed only gibberish (random keystrokes) rather than making a meaningful response to the stimuli.

Stimuli

Stimulus materials came from five different corpora that consisted of digital wave files of meaningful words, meaningful sentences, anomalous sentences, and environmental signals.

Modified Rhyme Test. The Modified Rhyme Test (MRT) corpus consisted of 300 words organized into fifty lists, where each list contains six rhymed variations on a common syllable (House, Williams, Hecker & Kryter, 1965). Within each list, the word initial or word final consonant is systematically varied to produce six items each differing only by a minimal pair (e.g., "bat", "bad", "back", "bass", "ban", "bath"). Stimuli consisted of ninety CVC words drawn from the MRT list, and their associated wav file recordings that were obtained from the PB/MRT Word Multi-Talker Speech Database in the Speech Research Laboratory at Indiana University, Bloomington. A female talker produced forty-two of the words, and a male talker, the remaining forty-eight.

Phonetically Balanced Words. The Phonetically Balanced corpus (PB) consisted of twenty lists of fifty monosyllabic words whose phonemic composition approximates the statistical occurrence in American English (e.g., "bought", "cloud", "wish", "scythe") (Egan, 1948). Stimuli consisted of ninety unique words drawn from lists 1-3 of the PB corpus so that no overlaps occurred with those selected from the MRT corpus. Wav file recordings were obtained from the PB/MRT Word Multi-Talker Speech Database in the Speech Research Laboratory at Indiana University Bloomington. Half of the stimuli were produced by a male talker, and the other half by a female talker.

Harvard/IEEE Sentences. The Harvard/IEEE Sentence database consisted of seventy-two lists of ten meaningful sentences (IEEE, 1969). These phonetically balanced (relative to American English) sentences contained five keywords embedded in a semantically rich meaningful sentence (e.g., "Her purse was full of useless trash", "The colt reared and threw the tall rider"). Stimuli consisted of twenty-five sentences drawn from lists 1-10 of the Harvard/IEEE Sentence database and their associated wav file recordings that were obtained from the speech corpus originally created by Karl and Pisoni (1994). A female talker produced fourteen sentences and a male talker produced the remaining eleven. Selection of these two talkers was based on their production of speech that was highly intelligible (90% correct keyword accuracy across the 100 sentences) as demonstrated by previous research (Bradlow, Toretta & Pisoni, 1996).

Anomalous Harvard/IEEE Sentences. Semantically anomalous sentences preserve the canonical syntactic structure of English, but have no meaning. The anomalous sentences from the corpus of Herman and Pisoni (2000) used the Harvard/IEEE sentence materials to create phonetically balanced meaningless sentences. The keywords from the 100 sentences in lists 11-20 were coded according to semantic category (noun, verb, adjective, adverb) and replaced with words from equivalent semantic categories from lists 21-70 (Herman & Pisoni, 2000). This operation created sentences that have legal syntactic structure in American English, but were semantically anomalous (e.g., "Trout is straight and also writes brass", "The deep buckle walked the old crowd"), thus precluding subjects from using semantic context to identify the keywords. Stimuli consisted of twenty-five anomalous sentences drawn from the anomalous Harvard/IEEE sentences corpus of Herman and Pisoni (Herman & Pisoni, 2000) and their associated wav file recordings. A female talker produced 13 of the sentences, whereas a male talker produced the remaining 12 sentences.

Environmental Stimuli. The environmental signal database of Marcell and colleagues consists of stimuli recorded from a wide variety of acoustic environments developed for use in neuropsychological evaluation and confrontation naming studies (Marcell, Bordella, Greene, Kerr & Rogers, 2000). The 120 stimuli in the corpus contain sounds from various acoustic events spanning a wide variety of categories: sounds produced by vehicles (e.g., automobile, airplane, motorcycle), animals (bird, dog, cow), insects (mosquito, crickets), non-speech sounds produced by humans (snoring, crying, coughing), musical instruments (piano, trumpet, flute), tools (hammer, vehicles), liquids (water boiling, rain) among others. These signals have been normed in a group of neurologically intact subjects on a variety of subjective (e.g., familiarity, complexity, pleasantness and duration) and perceptual measures (e.g., naming accuracy and naming response latency) (Marcell, Bordella, Greene, Kerr & Rogers, 2000). Stimuli consisted of ninety environmental stimuli and their associated wav file recordings obtained from a digital database published by the authors on the Internet (http://ww.cofc.edu/~marcellm/confront.htm). Stimulus selection from a variety of acoustic categories provided a wide representation of sound types and familiarity ratings.

Synthesis

Stimulus processing used a freeware program (Tiger CIS) developed for research that is available on the Internet (http://www.tigerspeech.com/). The software simulated an 8-channel CI using the CIS processing strategy. Stimulus processing involved two phases, an analysis phase, which divides the signal into bands and derives the amplitude envelope from each band and a synthesis phase, which replaces the frequency content of each band with a sinusoid that is modulated with the appropriate amplitude envelope. Analysis used band-pass filters to divide the stimuli into 8 spectral channels between 200 and 7000 Hz in steps with corner frequencies based on the Greenwood function (24 dB/octave slope). Envelope detection used a low pass filter with an upper cutoff at 400 Hz with a 24 dB/octave slope. Following the synthesis phase, the modulated sinusoids were combined and saved as 22 kHz 16 bit windows PCM wav files. Normalization of the wav files to a standard amplitude (65 dB RMS) using a leveling program (Level v2.0.3 Tice & Carrell, 1998) ensured that stimuli were equal in intensity across all materials, and that no peak clipping occurred.

Materials

Data collection used a custom script written for PsyScript, and implemented on four Apple PowerMac G4 (512 Mb RAM) computers running OS 9.2.2, and four 15 inch color Sony LCD monitors (1024x768 pixels, 75 Hz refresh). Audio signals were presented over four sets of Beyer Dynamic DT-100 headphones, calibrated with a voltmeter to a 1000 Hz tone at 70 dBv SPL using a voltage/intensity conversion table for the headphones. Sound intensity was fixed within PsyScript in order to guarantee consistent sound presentation across subjects.

Procedures

All methods and materials were approved by the Human Subjects Committee and Institutional Review Board at Indiana University Bloomington. Informed consent was established before beginning the experiment, and subjects were given a short subject information form asking for basic background information (basic background, demographic and contact information) and inquiring as to any prior hearing, speech, or language problems.

Multiple booths in the testing room accommodated up to four subjects at the same time. Subjects were informed that the stimuli they would hear were processed by a computer and that while they may have difficulty understanding them at first, they would quickly adapt. On screen instructions preceded each block to orient the subject to the materials and requirements of the upcoming task. Before the presentation of each audio signal, a fixation cross, presented at the center of the screen for 500 milliseconds alerted the subject as to the upcoming trial. The fixation cross was erased, and the sound file was presented at the next vertical retrace. Following stimulus offset, a dialog box appeared on the screen prompting subjects to type in what they heard. There were no time limits for responding. Subjects performed at their own pace, and were allowed to rest between each trial as needed. The experimental session lasted on average 45 minutes. All subjects received written and verbal debriefing after the experiment.

Training

Each training condition consisted of seven blocks. Stimuli were pre-randomized, and organized into separate lists for presentation in each training condition. Although the stimuli used in each block varied as a function of training materials, the same basic block design was consistent throughout all conditions (Fig. 1). Each condition began with a pre-test (block 1), which assessed the subjects' ability to identify the materials before training began. At this point, subjects are naïve to the stimulus processing, and had received no familiarization or adaptation. During the training sessions, subjects heard a stimulus, and then responded in the dialog box that appeared on the computer screen. Following their response, subjects received feedback in the form of the repetition of the processed auditory stimulus paired with the written form of the stimulus on the computer screen (the transcription of the word or sentence, or the descriptive label of the environmental stimuli) irrespective of whether their previous response was correct. An intervening generalization block occurred between the training block (block 2) and the post-test block (block 4). During the post-test, subjects heard a selection of old materials from the pre-test and post-test, as well as new materials from the same category. The post-test materials were selected to assess the effects of explicit training (using training materials), familiarity without explicit training (pre-test materials) and novelty (previously unheard materials). The remaining three blocks were generalization blocks testing the effects of training.

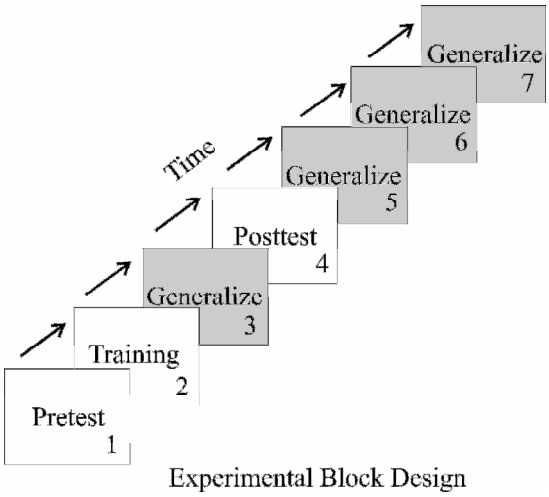


FIGURE 1. Block design of the experimental trials for all training groups.

MRT Word Training. During the pre-test, listeners were presented with twenty MRT words. Training consisted of fifty novel MRT words. An intervening generalization block occurred in block 3 to prevent habituation to the stimuli, and consisted of twenty-five anomalous sentences. The post-test in block 4 presented a total of 60 MRT words, twenty of which were drawn from the pre-test materials, twenty from training, and twenty were novel stimuli with which subjects had no previous experience in the experiment. The remaining three blocks consisted of generalization to 25 meaningful sentences (block 5), 50 novel PB words (block 6) and 60 environmental signals (block 7).

PB Word Training. PB training utilized an identical design to the MRT training, except that PB words consisted of the pre-test, training, and post-test materials and block 6 consisted of generalization to 50 novel MRT words.

Harvard/IEEE Sentence Training. In order to balance for the relative effect of words transcribed across sentences, fewer sentences were selected. The pre-test block consisted of four Harvard/IEEE sentences (20 key words); the training block consisted of ten novel Harvard/IEEE sentences (50 key words). Block 3 was an intervening generalization block, consisting of 50 MRT words. The post-test in block 4 utilized 12 Harvard/IEEE sentences, 4 selected from the pre-test, four from the post-test and four novel sentences (60 keywords). The remaining three blocks tested the effects of generalization to new materials. Block 5 consisted of 25 anomalous sentences, block 6 of 50 PB words and block 7 of 60 environmental signals.

Anomalous Sentence Training. Anomalous sentence training utilized an identical design to the Harvard/IEEE sentence training, except that the pre-test, training and post-test materials consisted of Anomalous sentences, and block 6 consisted of generalization to 25 novel Harvard/IEEE sentences.

Environmental Stimulus Training. Like the MRT and PB training, training on environmental training stimuli began with a pre-test consisting of twenty environmental signals and training consisting of fifty novel environmental signals. An intervening generalization block occurred in block 3 in order to prevent habituation to the stimuli and consisted of twenty-five Anomalous sentences. The post-test in block 4 presented a total of 60 environmental signals, twenty of which were drawn from the pre-test materials, twenty from training, and twenty were novel stimuli with which subjects had no previous experience in the experiment. The remaining three blocks consisted of generalization to 50 MRT words (block 5), 25 Harvard/IEEE sentences (block 6), and 50 novel PB words (block 7).

Analysis and Scoring

A supervised spellchecker corrected the more obvious spelling errors and standardized spelling across subjects by changing homophones into a standard spelling. An automated macro searched for target/response matches using a pre-ordained target list, the result of which was then hand checked by a trained research assistant. Responses that were morphologically related to the target were scored as incorrect. PB and MRT words were scored based on whether the entire word was correct, whereas anomalous and meaningful sentences were scored for keywords correct (5 keywords per sentence).

Environmental stimuli were checked using a similar procedure, except more options were included in the target list given the complexity of the stimuli. Scoring rules were modified slightly from those originally used by Marcell and colleagues (Marcell, Bordella, Greene, Kerr & Rogers, 2000) given the nature of the degradation. Animal and insect sounds were scored as correct if the subject identified the target agent (e.g., cow), the sound the agent made if it did not have multiple possible agents (e.g., moo), or the linking of the two (e.g., cow mooing). Responses were considered incorrect if the subject

failed to disambiguate the perceived agent from multiple agents (e.g., 'whistling' was an incorrect response for 'birds' given that human 'whistling' was a viable target, however 'tweet' and 'chirping' were considered correct). Failure to specify agent, or incorrectly specifying agent was scored as an incorrect response (e.g., for 'seal' the response 'seal barking' is correct, but the response 'barking' is incorrect given that the agent is not specified and could refer to a dog). Correct identification of musical instruments required accurate identification of the instrument. The generic response of 'music' was scored as incorrect, given that the instructions explicitly told subjects that this was not a valid response option. Multiple instruments from a given class were considered as viable options so long as they afforded a common action (e.g., the responses 'viola' and 'violin' were considered correct options for the target 'violin', however 'string' and 'guitar' were incorrect responses given that the action affords the use of a bow, whereas the action afforded by the latter response requires plucking).

Non-speech sounds produced by humans were considered correct if they correctly identified the sound given that the agent was unambiguous (e.g., 'child coughing' has the possible correct response options of 'child coughing', 'coughing' or 'cough'). 'Scream' on the other hand was correct if subjects identified the target 'scream' or some variant supposing a human agent. 'Monkey screaming' was incorrect given the misidentification of the agent. Liquid sounds were considered correct if the subject identified the agent or the action, and allowed for multiple specific sources as appropriate (e.g., 'water boiling' had the possible correct options of 'boil', 'bubble', 'bubbling' or 'bong').

For each training condition, responses were averaged across subjects for each block. Within-subjects analyses compared performance across blocks of a given training condition. Paired samples t-tests were used to assess the effects of training by comparing pre and post-test performance. Post-test scores were balanced by only averaging the responses to the materials on which subjects were not explicitly trained, to avoid biasing the findings. The differences in performance on the various post-test materials (items from pre-test, training and novel lists) were assessed with a one-way ANOVA and post hoc Tukey tests. Scores were organized in a column, and coded to reflect the source (pre-test, training or novel). Other paired t-tests were conducted to assess the effects of context (Anomalous sentences vs. Harvard/IEEE sentences) and complexity (PB words vs. MRT words). A correlational analysis examined the relationship between performance across blocks to assess whether performance on one type of material was correlated with performance on another. Between subjects comparisons assessed the effects of training on materials across training conditions using one-way Analysis of Variance and post-hoc Tukey tests.

Results

Within Group Comparisons

MRT Training. Overall, initial performance of the 25 subjects who received training on the MRT materials started out very poor, but increased following training (Fig. 2). Percent correct recognition increased from 5.8 % correct at pre-test to 37.5% after training, demonstrating a gain of nearly 32 percentage points. A paired t-test indicated that the effect of training was highly significant ($t(1, 24)$=13.576, p<0.001). Comparison of the various post-test materials (data not shown) demonstrated that subjects performed best on stimuli from the training list (materials on which they were explicitly trained), followed by stimuli from the pre-test list (materials with which they were familiarized but not trained) and finally stimuli from the novel list (MRT materials that did not appear before the post-test). A one-way ANOVA revealed a significant main effect of source material, demonstrating that subjects performed differently on materials from the pre-test, training and novel lists ($F(2, 74)$=18.967, p<0.001). Post hoc Tukey tests revealed that subjects performed significantly better on the materials that they heard

during training (58% correct) than on pre-test (40.4% correct) or novel materials (34.6% correct, both $p<0.001$), demonstrating a significant effect of feedback, and indicating good retention of training. Subject performance did not differ on the materials drawn from the pre-test and novel lists ($p=0.313$), suggesting that explicit training promotes more of a benefit than exposure alone.

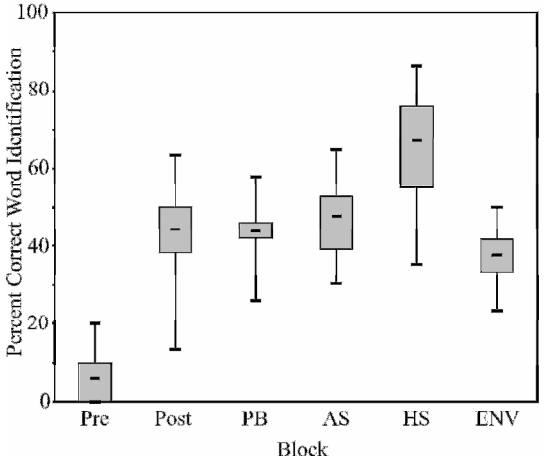


FIGURE 2. Box plot displaying the perceptual accuracy scores as a function of experimental block for the 25 subjects trained to identify the MRT stimuli. Boxes encompass the middle 50% of the data, and horizontal lines indicate the average score for that block. Pre-test scores reflect the baseline performance on the MRT words before training, when subjects were naïve to the processing condition. Post-test scores contain only the responses to MRT stimuli on which subjects did not receive explicit training (see text). MRT and PB words were judged as correct if the subject typed the entire word correctly. Harvard/IEEE (HS) and Anomalous (AS) sentence scores reflect the percent of key words correctly typed. Environmental stimuli (ENV) scores reflect the correct identification of the sound (see text).

Overall, subjects performed best on the Harvard/IEEE sentences (67.0% correct), followed by anomalous sentences (47.7% correct), PB words (43.7% correct) and Environmental stimuli (37.6% correct). A paired t-test revealed a significant effect of sentence context on recognition. Subjects performed significantly better on the Harvard/IEEE sentences than on the anomalous sentences ($t(1,24)=18.327$, $p<0.001$). The difference between the scores for the meaningful and anomalous sentences suggests that the addition of context leads to improvement by almost 20%. A paired t-test comparing performance on the MRT and PB words also indicates a difference in performance, with subjects performing significantly better on PB materials than on MRT ($t(1,24)=3.928$, $p=0.001$). This may be due to differences in the difficulty of the words used in the MRT and PB lists, since the MRT words include only minimal pairs.

Correlations of the performance across blocks revealed several significant results. Performance at post-test was significantly correlated with performance on each measure except for environmental stimuli (MRTpost-test vs. PB $r=0.766$, MRTpost-test vs. HS $r=0.672$, MRTpost-test vs. AS $r=.576$, all $p<0.01$). Similar relationships were observed for the PB words (PB vs. HS $r=.654$, PB vs. AS $r=.552$), and anomalous and Harvard/IEEE sentences (AS vs. HS $r=.905$). It is interesting to note that performance on isolated words was most strongly correlated with performance on other words, followed by meaningful and anomalous sentences, and that sentences were most strongly correlated with other sentences followed by PB and MRT words.

PB Training. Subjects trained on the PB words started out better than those subjects trained on the MRT words. Performance at pre-test was 23.4%, but increased to 46.2% correct following training (Fig. 3). A paired samples t-test indicated that subjects performed significantly better at post-test as compared to pre-test ($t(1,24)$=7.134, p<0.001). Examination of the post-test materials (data not shown) revealed that subjects performed best on stimuli on which they were explicitly trained (55.4% correct), followed by novel PB words (48.2% correct) and words on which they were previously exposed, but not explicitly trained (44.2% correct). A one-way ANOVA revealed a significant main effect of list ($F(2, 74)$=5.484, p=0.006). Post hoc Tukey tests revealed that subjects performed significantly better on materials from the training list (p=0.005) than on materials from the pre-test list, but no difference was observed when compared with the materials drawn from the novel list (p=0.097). More importantly, subject performance did not differ

As observed for the MRT training condition, subjects performed best on the Harvard/IEEE sentences (66.2% correct), followed by the Anomalous sentences (48.2% correct), MRT words (42.8% correct) and Environmental stimuli (35.2% correct). A paired t-test revealed that subjects performed significantly better on the Harvard/IEEE sentences than on the anomalous sentences ($t(1,24)$=12.214, p<0.001). Subtraction of the scores for the anomalous sentences from those for the Harvard/IEEE sentences reveals a 20% gain from context. Subjects' performance did not differ significantly between the PB words and MRT blocks ($t(1,24)$=1.855, p=.076) although a slight numerical trend was observed favoring PB words.

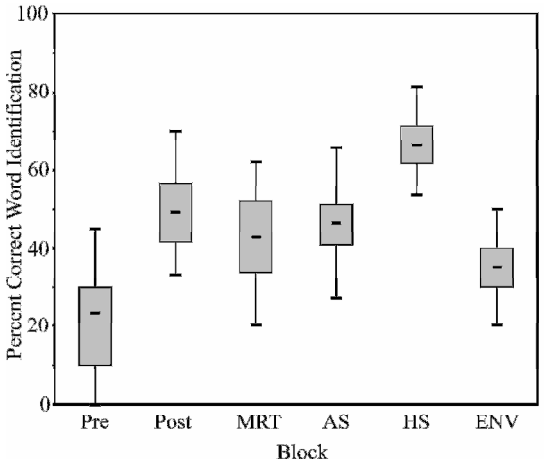


FIGURE 3. Box plots displaying the perceptual accuracy scores as a function of experimental block for the 25 subjects trained to identify the PB stimuli.

Performance on the PB words during the post-test was significantly correlated with performance in the MRT block (r=.658, p<0.001), Harvard/IEEE sentences (r=.539, p=0.005), but not for the Anomalous sentences or Environmental stimuli. Performance on the Harvard/IEEE sentences was significantly correlated with performance on Anomalous sentences (r=.609, p=0.001) and MRT words (r=.598, p=0.02). MRT performance was also correlated with performance on Anomalous sentences (r=.515, p=0.008) and Environmental stimuli (r=.554, p=0.004). As observed in the MRT training group, performance on words (PB or MRT) was most strongly correlated with performance on other words, and performance on sentences was most strongly correlated with performance on other sentences.

between materials drawn from the pre-test and novel list (p=0.477), indicating that training generalized to new words of the same class, and that performance was not contingent on having heard the word before.

Anomalous Sentence Training. Figure 4 shows subject performance on the Anomalous sentence training condition. Performance was good at pre-test (33.6% correct) but increased significantly following training (61.7% correct, $t(1,24)$=11.713, p<0.001). Examination of the post-test materials (data not shown) revealed a significant main effect of source ($F(2, 74)$=14.115, p<0.001), and post hoc Tukey tests confirmed that subjects performed significantly better on the materials from the training list (78.2% correct) than on materials from either the pre-test list (61.6% correct, p<0.001) or novel list (61.8% correct, p<0.001). No differences in performance were observed on the materials from the pre-test and novel lists (p=0.998).

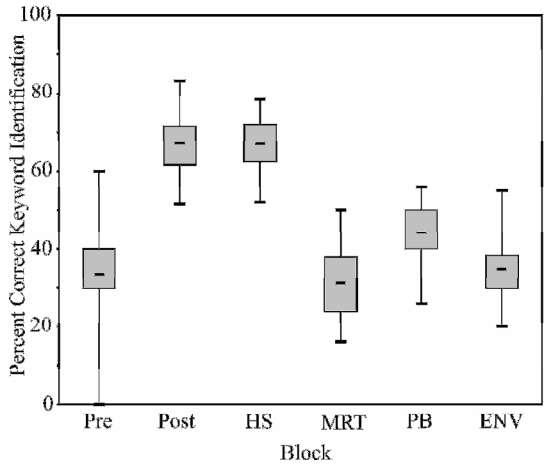


FIGURE 4. Box plots displaying the perceptual accuracy scores as a function of experimental block for the 25 subjects trained to identify the Anomalous sentences.

As observed previously, subjects performed best on the Harvard/IEEE sentences (67.04% correct), followed by the anomalous sentences (61.7% correct), PB words (44.1% correct), environmental stimuli (34.9% correct) and MRT words (31.2% correct). Performance on the Harvard/IEEE sentences was significantly higher than on the Anomalous sentences ($t(1,24)$=3.406, p=0.002), and subtraction of the scores on these blocks revealed only a small 5% gain from context, suggesting that the large gains due to context observed in the MRT and PB training were ameliorated with explicit training on the anomalous sentences. Subjects also performed significantly better on PB as compared to MRT words ($t(1,24)$=6.140, p<0.001), as observed previously. Performance on the Anomalous sentences was correlated only with performance on Harvard/IEEE sentences (r=.63, p=0.001). The only other significant correlation observed was between PB words and Environmental stimuli (r=.446, p=0.025). All other correlations were not significant.

Harvard/IEEE Sentence Training. Performance on the Harvard/IEEE sentence post-test significantly increased from pre (40% correct) to post-test (63.9% correct, $t(1,24)$=7.041, p<0.001). Subject performance varied across the post-test materials, and an ANOVA analysis revealed a significant main effect of source ($F(2, 74)$=114.043, p<0.001). Subjects performed significantly better on materials from the training list (97% correct) than on those from the pre-test (71.6% correct) and novel (56.2%

correct) lists (all $p<0.001$). Subjects also performed significantly better on the materials drawn from the pre-test list as compared to the novel list ($p<0.001$). This is likely due to the high contextual salience of the sentences, because this pattern was not observed for the Anomalous sentence training group.

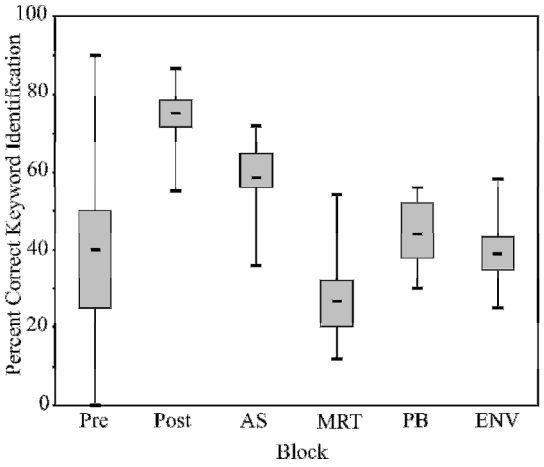


FIGURE 5. Box plots displaying the perceptual accuracy scores as a function of experimental block for the 25 subjects trained to identify the Harvard/IEEE sentences.

Figure 5 shows that subjects performed best on the Harvard/IEEE sentences, followed by the Anomalous sentences (58.6% correct), PB words (44.1% correct), Environmental stimuli (39% correct) and MRT words (26.7% correct). A paired t-test revealed that subjects performed significantly better on the Harvard/IEEE sentences than on the Anomalous sentences ($t(1,24)$=3.328, p=0.003). The gain from context was only approximately 5%. Subjects also performed significantly better on the PB words as compared to the MRT words ($t(1,24)$=10.332, $p<0.001$). Performance on the Harvard/IEEE sentences was significantly correlated with performance on Anomalous sentences (r=.551, p=0.004), followed by PB words (r=.512, p=0.009) and MRT words (r=.398, p=0.49). Anomalous sentences were most strongly correlated with performance on PB words (r=.733, $p<0.001$) and MRT words (r=.623, p=0.001). Performance on PB words was significantly correlated with performance on MRT words (r=.587 p=0.002) and Environmental stimuli (r=.568, p=0.003).

Environmental Stimulus Training. Performance on the Environmental stimuli also showed a significant benefit from explicit training (Fig. 6). Subjects showed significant improvement between pre (38.2% correct) and post-test (46.4% correct, $t(1,24)$=2.804, p=0.01). An analysis of the post-test materials (data not shown) revealed a significant main effect of source ($F(2, 74)$=8.717, $p<0.001$). Subjects performed best on stimuli from the novel list (53.2% correct), followed by materials from the training list (50% correct) and pre-test (39.6% correct). Subjects performed significantly better on materials from both novel and training lists than on materials from the pre-test list (p=0.009 and $p<0.001$ respectively) but did not differ from one another (p=0.617).

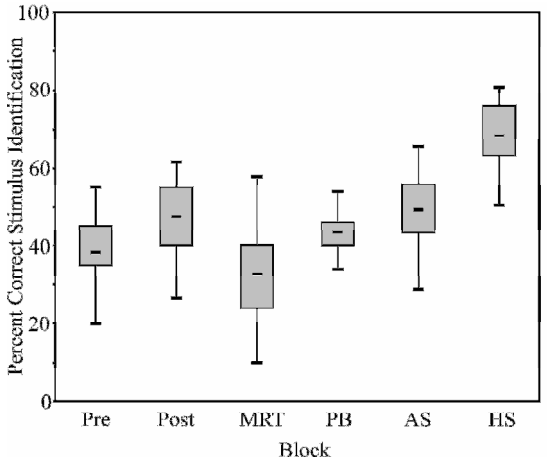


FIGURE 6. Box plots displaying the perceptual accuracy scores as a function of experimental block for the 25 subjects trained to identify the Environmental stimuli.

Overall, subjects performed best on Harvard/IEEE sentences (68.5% correct), followed by the Anomalous sentences (49.31% correct), Environmental stimuli, PB (43.4% correct) and MRT words (32.8% correct). Subjects received an approximated gain of 19% from context ($t(1,24)$=13.772, p<0.001). Subjects also performed significantly better on PB words as compared to MRT words ($t(1,24)$=3.830, p=0.001). Performance on the Environmental stimuli was not significantly correlated with any other material, but as observed earlier, Harvard/IEEE sentences were significantly correlated with Anomalous sentences (r=.69, p<0.001).

Across Group Comparisons

To assess the effect of training on the source materials, the recognition accuracy scores for a given set of materials were compared across training conditions and to the scores at pre-test. Comparison with the post-test scores (which did not contain the materials repeated from pre-test) assessed whether the type of training significantly affected performance, and whether training on a specific set of materials produces better and more robust generalization than another.

MRT Words. Figure 7 displays the across group performance on the MRT words. A one-way ANOVA using Training Materials as the between subjects factor main effect of training materials ($F(5, 149)$=37.495, p<0.001). Post hoc Tukey tests revealed that subjects performed significantly better than the pre-test regardless of the type of material that they were trained upon (all p<0.001). This is not surprising, given the poor baseline performance (5.8% correct). Although any type of training produced a benefit, MRT and PB training produced greater benefits than any other material (37.5% correct, and 42.8% correct respectively). That performance did not differ between the MRT and PB trained groups (p=0.477) suggests that training on words, regardless of their origin, produces equivalent benefit when recognizing other single words. Training on Anomalous sentences, Harvard/IEEE sentences and Environmental stimuli also produced significant gains over baseline, but were the poorest of all conditions (31.2% correct, 26.7% and 32.8% correct respectively). Moreover, performance did not differ between these three groups (all p>0.319). Interestingly, subjects trained on the Anomalous sentences and Environmental When the scores were grouped by material type, however, subjects who received training on words (MRT and PB) performed significantly better than subjects trained on sentences (p<0.001) or environmental stimuli (p=0.027).

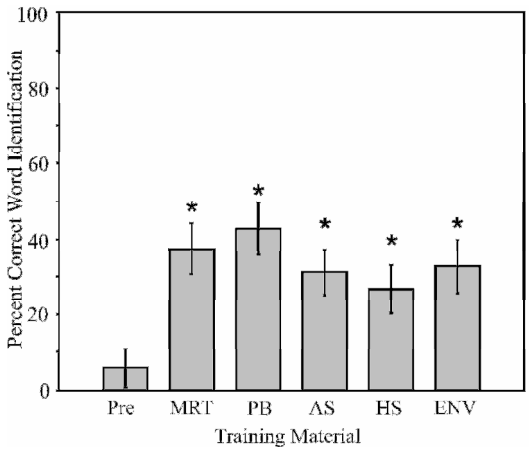


FIGURE 7. Bar graph displaying perceptual accuracy scores at identifying MRT stimuli as a function of training. Training condition is indicated along the x-axis. Pre and post-test scores are for the subjects who were explicitly trained on the MRT stimuli. The remaining bars indicate subjects' performance on the MRT generalization block of their respective training sessions. Post-test scores contain only the responses to stimuli on which subjects did not receive explicit training (see text). Asterisks indicate when performance was significantly greater than baseline ($p<0.05$).stimuli performed as well as subjects trained on the MRT stimuli ($p=0.281$ and $p=0.610$ respectively).

PB Words. Training produced a significant impact on subjects performance on the PB materials (Figure 8), and a one-way ANOVA using Training Materials as the between subjects factor indicated a significant main effect materials ($F(5, 149) = 24.86$, $p<0.001$). Compared to baseline, post-test performance is significantly higher as a result of training on PB materials (23.4% as compared to 46.2%, $p<0.001$). Overall, it did not matter what type of training subjects received, as performance was significantly higher than pre-test for all training conditions (MRT training 43.4% correct $p<0.001$, AS training 44.1% correct $p<0.001$, HS training 44.1% correct $p<0.001$, ENV training 43.68% correct $p<0.001$). The main effect for training condition is carried entirely by the gains in performance relative to the pre-test, as there were no significant differences between performance across the five training conditions (all $p>0.867$). This indicates that when identifying words that are highly discriminable, training with any type of material will provide an equivalent benefit.

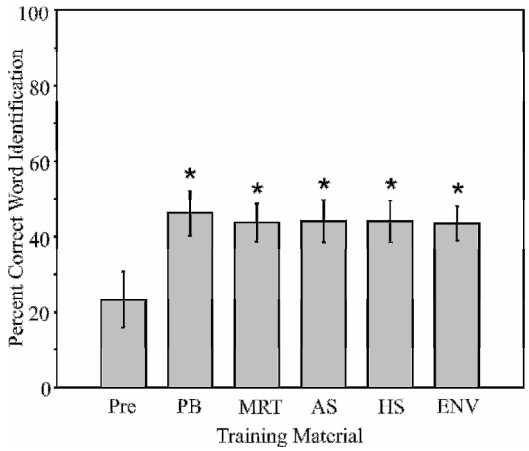


FIGURE 8. Bar graph displaying perceptual accuracy scores at identifying PB stimuli as a function of training. Training condition is indicated along the x-axis. Pre and post-test scores are for the subjects who were explicitly trained on the PB stimuli. The remaining bars indicate subjects' performance on the PB generalization block of their respective training sessions. Post-test scores contain only the responses to stimuli on which subjects did not receive explicit training (see text).

Anomalous Sentences. The performance on the anomalous sentences across training conditions is shown in Figure 9. A one-way ANOVA revealed a significant main effect of training ($F(5, 149) = 22.986$, $p<0.001$). Comparison with the pre-test revealed that all types of training produced significant increases in performance relative to the baseline (33.6% correct, all $p<0.001$). No differences in performance were observed between subjects who received explicit training on the Anomalous sentences (61.7% correct) and to those who were trained on the meaningful Harvard/IEEE sentences (58.6% correct, $p=0.902$). In contrast, subjects who received training on the PB, MRT and Environmental stimuli showed significantly less gain in performance as compared to subjects trained on the Anomalous (all $p<0.001$) or Harvard/IEEE (all $p<0.004$) sentences. Training on MRT (47.7% correct), PB (46.5% correct) and Environmental stimuli (47.7% correct) provided equivalent benefit when recognizing the Anomalous sentences, however (all $p>0.0998$).

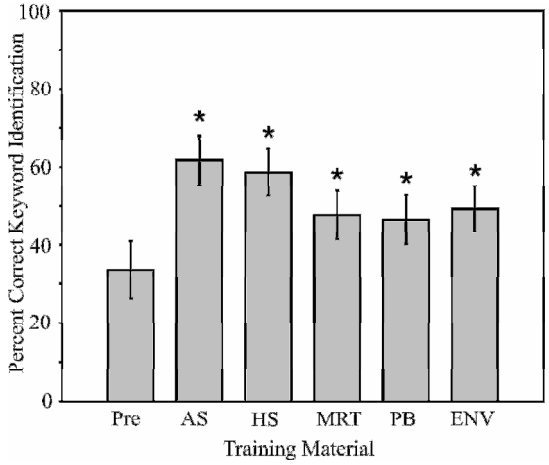


FIGURE 9. Bar graph displaying perceptual accuracy scores at identifying Anomalous sentences as a function of training. Training condition is indicated along the x-axis. Pre and post-test scores are for the subjects who were explicitly trained on the Anomalous sentences. The remaining bars indicate subjects' performance on the AS generalization block of their respective training sessions. Post-test scores contain only the responses to stimuli on which subjects did not receive explicit training (see text).

Harvard/IEEE Sentences. The comparison of performance on the Harvard/IEEE sentences across training conditions is shown in Fig. 10. A one-way ANOVA revealed a significant main of training condition on performance ($F(5, 149) = 22.444$, $p<0.001$). The comparison of each of the training conditions to the Harvard/IEEE sentence pre-test revealed that subjects performed significantly better than the baseline (40% correct) regardless of the type of training they received (all $p<0.001$). As was the case for the PB materials, the training effect is carried entirely by the gains in performance relative to the pre-test, as there were no significant differences between performance across the five training conditions (MRT 67.0% correct, PB 66.2% correct, HS 63.9% correct, AS 67.0% correct, ENV 68.5% correct, all $p>0.719$).

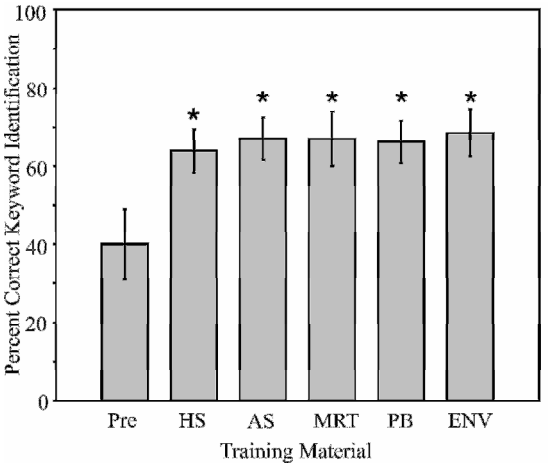


FIGURE 10. Bar graph displaying perceptual accuracy scores at identifying Harvard/IEEE sentences as a function of training. Training condition is indicated along the x-axis. Pre and post-test scores are for the subjects who were explicitly trained on the Harvard/IEEE sentences. The remaining bars indicate subjects' performance on the HS generalization block of their respective training sessions. Post-test scores contain only the responses to stimuli on which subjects did not receive explicit training (see text).

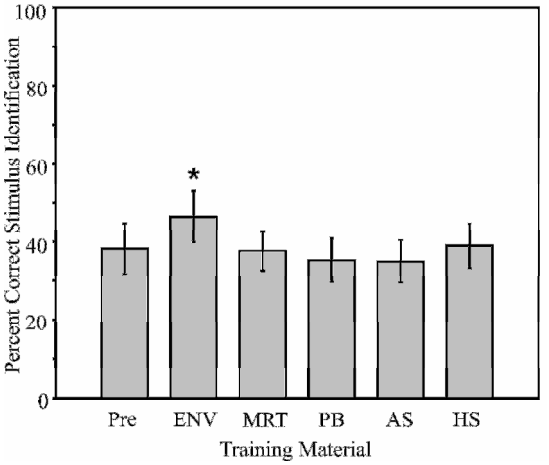


FIGURE 11. Bar graph displaying perceptual accuracy scores at identifying the Environmental stimuli as a function of training. Training condition is indicated along the x-axis. Pre and post-test scores are for the subjects who were explicitly trained on the Environmental stimuli. The remaining bars indicate subjects' performance on the ENV generalization block of their respective training sessions. Post-test scores contain only the responses to stimuli on which subjects did not receive explicit training (see text).

Environmental Stimuli. The effect of training on the recognition of the Environmental stimuli is shown in Figure 11. A one-way ANOVA revealed a significant main effect of training group on performance ($F(5, 149) = 5.847$, $p<0.001$). Unlike the training effects observed for the other stimulus materials, subjects only showed gains relative to baseline (38.2% correct) when they were explicitly trained on the Environmental stimuli (46.4% correct, $p=0.013$). Subjects trained on all other materials failed to show any differences as compared to baseline (MRT 37.6% correct $p=1.00$; PB 35.2% correct

p=0.822; HS 34.9% correct p=0.999; AS 34.9% correct p=0.764). Moreover, performance was significantly higher for those subjects explicitly trained on the Environmental stimuli as compared to all other groups (all p<0.03). Training on MRT, PB, AS and HS materials provided equivalent levels of generalization to the Environmental stimuli (all p>0.557). Since these values did not differ from the baseline, however, it suggests that training on the speech materials is equally ineffective when transferring to environmental stimuli. In effect, when asked to identify environmental stimuli, training with speech materials is as effective as not receiving any training at all.

Discussion

Overall, the specific type of materials used during the training portion of the experiment had a significant impact on performance. Across all training conditions, subjects showed significant pre to post-test improvement, demonstrating that for each set of training materials, subjects were able to utilize the feedback to improve their identification accuracy. Generalization effects were not uniform across materials. Subjects showed encoding specificity, performing best on the materials on which they were explicitly trained. Subjects who were trained on words (PB or MRT) performed significantly better when identifying MRT stimuli than the other groups, and subjects who were trained on sentences (anomalous or meaningful) performed significantly better when identifying Anomalous sentences than the other groups. This suggests that when the task demands were high, subjects performed better when they were trained on stimuli of the same general class (e.g., training on words generalized significantly better to other words, sentences generalized significantly better to other sentences), demonstrating transfer of appropriate processing. The opposite effect was observed for the "easier" materials: subject performance did not differ across training groups on the PB words and Harvard/IEEE sentences. This suggests that when the task demands are less difficult, such as when identifying high frequency words and meaningful sentences, all forms of training are equivalent.

One intriguing finding from this study was the asymmetry in training that was observed for the environmental stimuli. Subjects trained on environmental stimuli performed significantly better than baseline on all speech materials, suggesting that training on complex non-speech stimuli produces robust generalization to speech. The inverse, however, was never observed: training on speech consistently failed to produce performance that differed from the environmental baseline. Thus, it appears that training on complex non-speech materials leads to improved performance on speech materials, but training on speech materials does not produce gains in the perception of complex non-speech abilities. Increased attentional sensitivity to the spectral and temporal characteristics of the environmental stimuli may have enhanced subjects' abilities at utilizing similar spectral information that is important to speech.

The present findings are similar to those of Gygi and colleagues, who found that the most important information for recognition of environmental stimuli occupies an identical frequency range as that for speech (Gygi et al., 2004). If the important information for environmental stimuli overlaps with that of speech, then training subjects to better utilize the spectro-temporal information in this frequency region more efficiently should foster generalization to speech, as we report here. Training on speech alone may not be sufficient to foster generalization to environmental stimuli, since the spectro-temporal information to which subjects are utilizing may be more broadly distributed for these stimuli. Additionally, some environmental stimuli may be inherently more identifiable than others based on their spectro-temporal profiles (Shafiro, 2004; Burkholder, 2005; Burkholder et al., submitted 1; 2). The interaction between the number of spectral bands needed for successful recognition that was found by Shafiro (2005) was somewhat divergent from that typically observed for speech. Some environmental stimuli were most recognizable with fewer bands, and recognition actually decreased with the addition of bands (Shafiro, 2005). This suggests that some environmental stimuli may not be as readily identifiable

when processed by a vocoder. Moreover, given that the amount of acoustic information differs across acoustic environments and task demands, the spectral resolution of the current generation of CIs may be insufficient to provide significant benefit under all listening situations (Shannon, Fu & Galvin, 2004; Shannon, 2005). This possibility warrants further investigation.

The finding that training on speech does not generalize to environmental stimuli conflicts with the earlier findings of Burkholder and colleagues (Burkholder, 2005; Burkholder et al., submitted 1; 2), who reported that training on speech did generalize to environmental stimuli. However, Burkholder did not use a pre-post test design, so the baseline performance levels for environmental stimuli were not known. In the present study, although training on speech materials produced performance levels for environmental stimuli that were greater than zero, they did not exceed the baseline values. This suggests that subjects in the Burkholder et al study (submitted 1, 2) may not have performed any differently after training than subjects who were totally naïve to the stimulus processing conditions.

One methodological difference between the present study and earlier studies using environmental stimuli is the use of open set testing procedures in all conditions. The majority of the earlier studies used closed-set forced-choice testing procedures. Gygi and colleagues reported closed-set identification scores of up to 66% correct using 6-channel noise vocoded stimuli (Gygi et al., 2004). Shafiro found that although closed-set performance reaches asymptote with 16 channels (66%), large stimulus specific effects were observed (Shafiro, 2004). Moreover, Reed and Delhorne (2005) found that CI users show higher levels of closed-set performance still (79% correct). Under open set testing average performance after training (46% correct) was substantially lower than the performance observed in the previous studies. Given that the closed set procedures necessarily limit subjects to a certain set of responses, open set testing allows subjects to record their actual impressions of the stimuli in a way that would be more appropriate to real world listening environments (see Clopper, Pisoni & Tierney, 2006 for a more complete account).

A methodological question is also raised here. Although many previous studies have not demonstrated substantive differences for the perception of speech as processed by a noise and sinewave vocoder (Dorman et al., 1997), other studies have found that for non-speech tasks, performance is actually better for sinewave vocoded speech (Gonzales & Oliver, 2005). Gender and talker identification were significantly better for stimuli processed using a sinewave vocoder than when processed using a noise vocoder (Gonzales & Oliver, 2005). The authors suggest that the sinewave carriers may have introduced less distortion, thus preserving more accurate and robust detail in the amplitude envelopes that could be useful to the listener. A comparison of the two methods revealed more residual periodic information in the sinewave vocoder processed signal as compared to the noise vocoder processed signal, forming the basis for their claim (Gonzales & Oliver, 2005). It may be the case that a sinewave vocoder may produce better, more robust results for studies using music and environmental stimuli than would a noise vocoder: for stimuli that carry more salient spectral information, less distortion and better preserved periodicities in the envelope may translate to heightened recognition. Whether performance on these types of stimuli differs from performance of CI users remains an open question.

The asymmetry in training that was observed in the present study suggests that the ability to utilize the residual spectro-temporal information in the vocoded signals may enhance the ability to perceive unfamiliar speech signals under these difficult listening conditions. Surprenant and Watson (2001) reported a significant correlation between subjects' ability to discriminate non-speech stimuli based on spectro-temporal cues and their identification of speech in noise. The authors suggested that common higher order acoustic processes may contribute to both speech and non-speech processing capabilities. This could account for the substantial differences in performance of subjects who receive

hearing aids, and CIs alike: auditory sensitivity at a peripheral level may not be the sole cause of variability; rather the inability to utilize and manipulate such information at higher levels may supersede the benefits of an acoustic prosthesis (Surprenant & Watson, 2001). This relationship may not be completely bidirectional, however, given our findings that training on environmental stimuli generalizes to speech, but training on speech does not generalize to environmental stimuli.

Moreover, recent neuroimaging studies investigating the encoding of environmental stimuli have suggested that similar cortical regions may be involved during the processing of environmental stimuli and speech sounds (Lewis, Wightman, Brefczynski, Phinney, Binder & DeYoe, 2004). These cortical regions include the canonical auditory areas required for the recognition of sound (primary auditory cortex), the identification of auditory speech stimuli (superior temporal gyrus, posterior superior temporal sulcus, pSTS), semantic processing and accessing of lexical information during sound, picture and action naming (posterior medial temporal gyrus, pMTG) (Lewis et al., 2004). These cortical areas (the pMTG and pSTS in particular) showed bilateral activation in response to environmental stimuli, but tend to be left lateralized during speech perception tasks (Lewis et al., 2004). This difference may partially explain the asymmetry that we observed for training with environmental stimuli and speech. Perhaps training with environmental stimuli activated cortical regions implicated in the processing of speech stimuli, leading to efficient generalization to speech. Due to different task demands, training with speech may have utilized additional lateralized cortical regions which would not necessarily facilitate generalization to environmental stimuli. Additionally, other recent neuroimaging studies have demonstrated that the functional connectivity between cortical regions may be differentially altered due to task demands when identifying speech (Obleser, Wise, Dresner & Scott, 2007). This may facilitate generalization in one case (environmental stimuli to speech), but not the other (speech to environmental stimuli).

Our findings also replicate and extend the recent studies conducted by Davis et al (2005) and Burkholder et al (submitted 1, 2). Training using orthographic feedback paired with a repetition of the processed version of the sentence produced keyword correct identification scores (71% correct) that were nearly identical to those observed by Davis in the last block of training (75% correct). We also found that training on anomalous sentences produced excellent generalization to meaningful sentences, as was reported previously by both Davis et al. (2005) and Burkholder et al. (submitted 1, 2). Thus, access to syntactic structure without relying on sentence context enhances general sentence recognition. Our extension to include single PB words and CVCs also provides support for this conclusion: training on all materials produced excellent generalization to the meaningful Harvard/IEEE sentences. The results observed for training on environmental stimuli suggest that learning to recognize the acoustic form of a stimulus enhances selective attention to spectro-temporal information, and bottom up perceptual encoding processes.

The present study also replicates the findings of Fu and colleagues, who showed that giving CI users explicit training on CV and CVCs does indeed produce gains in sentence intelligibility (Fu et al., 2006). The similar patterns of performance observed with normal hearing subjects listening to acoustic simulations of a CI provides further support for the utility of the vocoder as an effective model of electric hearing. By studying the perceptual learning of CI simulated speech in normal hearing listeners, we can simultaneously learn about the neural and behavioral mechanisms that underlie speech and language processing in general, and expand our knowledge about effective rehabilitation and training programs to assist newly implanted individuals. By formalizing training paradigms that utilize a wide variety of stimulus materials, we may be able to provide CI users with tools that will bootstrap onto a variety of tasks and difficult listening conditions above and beyond those on which they were trained (i.e. increase "carry-over" effects). Given the substantial variability in performance among CI users that cannot be

attributed to individual differences in etiology and duration of deafness, the question remains as to how differences in post-implantation experience contribute to outcome and benefit. Providing explicit instruction as to the important information in the signal may help to account for a portion of this variability, thereby allowing us to disentangle the role of experience and provide a more objective assessment of the CI user success.

In summary, we demonstrated that the type of stimulus materials used during perceptual learning affects generalization to new materials. Although all forms of training provided some benefit, generalization of training was not uniform. When the task was easy, such as was the case when identifying contextually rich, meaningful sentences or highly discriminable isolated words, all five training conditions provided equivalent benefits. When the task was difficult, such as was the case when identifying low discriminable CVCs or sentences without the benefit of context, subjects who were trained on materials of a similar nature to those on which they were being tested performed significantly better. However, the addition of environmental signals revealed a unique asymmetry: training on environmental signals generalized to the recognition of speech, but training on speech did not generalize to environmental signals. This pattern of performance suggests that a wide variety of stimulus materials should be used during training to maximize perceptual learning and promote robust generalization to novel acoustic signals.

References

Bradlow, A.R., Toretta, G.M. and Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*, 255-272.

Burkholder, R.A. (2005). Perceptual learning of speech processed through an acoustic simulation of a cochlear implant. *Research on Spoken Language Processing Technical Report No. 13,* Bloomington, IN: Speech Research Laboratory, Indiana University.

Burkholder, R.A., Pisoni, D.B. and Svirsky, M.A. (submitted 1). Transfer of auditory perceptual learning with spectrally reduced speech to speech and nonspeech tasks: Implications for cochlear implants. *Ear and Hearing.*

Burkholder, R.A., Pisoni, D.B. and Svirsky, M.A. (submitted 2). Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant. *Journal of Experimental Psychology: Human Perception and Performance.*

Chiu, C.-Y.P., and Schacter, D.L. (1995). Auditory priming for nonverbal information: Implicit and explicit memory for environmental sounds. *Consciousness and Cognition, 4,* 440-458.

Chiu, C.-Y.P., (2000). Specificity of auditory implicit and explicit memory: Is perceptual priming for environmental sounds exemplar specific? *Memory and Cognition, 28(7),* 1126-1139.

Clark, G.M. (2002). Learning to understand speech with the cochlear implant, In Fahle, M, and Poggio, T. Eds. *Perceptual Learning,* pp. 147-160. Boston: MIT press.

Clopper, C.G., Pisoni, D.B. and Tierney, A.T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology, 17(5),* 331-349.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K. and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise vocoded sentences. *Journal of Experimental Psychology, 134(2),* 222-241.

Dorman, M.F., Loizou, P.C. and Rainey, D. (1997). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. *Journal of the Acoustical Society of America, 102(1),* 2993-2996.

Dorman, M. and Loizou, P. (1998). The identification of consonants and vowels by cochlear implants patients using a 6-channel CIS processor and by normal hearing listeners using simulations of processors with two to nine channels. *Ear and Hearing, 19,* 162-166.

Egan, J.P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955-991.

Fu, Q.-J., Galvin, J., Wang, X. and Nogaki, G. (2006). Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoustic Research Letters Online, 6(3),* 106-111.

Gonzales, J. and Oliver, J.C. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *Journal of the Acoustical Society of America, 118,* 461-470.

Gygi, B., Kidd, R.R. and Watson, C.S. (2004). Spectral-temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America, 115(3),* 1252-1265.

Herman, R. and Pisoni, D.B. (2000). Perception of elliptical speech by an adult hearing-impaired listener with a cochlear implant: some preliminary findings on coarse-coding in speech perception. *Research on Spoken Language Processing Progress Report No. 24* (pp. 87-112) Bloomington, IN: Speech Research Laboratory, Indiana University.

House, A.S., Williams, C.E., Hecker, M.H.L. and Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37*, 158-66.

IEEE. (1969). IEEE recommended practice for speech quality measurements. (IEEE Report No. 297).

Karl, J.R. and Pisoni, D.B. (1994). Effects of stimulus variability on recall of spoken sentences: A first report. *Research on Spoken Language Processing Progress Report No. 19* (pp. 145-193) Bloomington, IN: Speech Research Laboratory, Indiana University.

Lachs, L., McMichael, K. and Pisoni, D.B. (2003). Speech perception and implicit memory: Evidence for detailed episodic encoding of phonetic events. In J. Bowers and C. Marsolek (Eds.), *Rethinking implicit memory*. (pp. 215-235). Oxford: Oxford University Press.

Lewis, J.W., Wightman, F.L., Brefczynski, J.A., Phinney, R.E., Binder, J.R. and DeYoe, E.A. (2004). Human brain regions involved in recognizing environmental stimuli. *Cerebral Cortex, 14(9),* 1008-1021.

Marcell, M.M., Borella, D., Greene, M., Kerr, E., and Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology, 22(6),* 830-864.

Miller, G.A. and Nicely, P. (1955). An Analysis of Perceptual Confusions among some English Consonants, *Journal of the Acoustical Society of America, 27(2),* 338-352.

National Institutes of Health. (1995). Cochlear implants in adults and children. *NIH Consensus statement, 13(2),* 1-29.

Obleser, J., Wise, R.J.S., Dresner, M.A., and Scott, S.K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience, 27(9),* 2283-2289.

Reed, C.M and Delhorne, L.A. (2005). Reception of environmental sounds through cochlear implants. *Ear and Hearing, 26(1)*, 48-61.

Shafiro, V. (2004). Perceiving the sources of environmental sounds with a varying number of spectral channels, Unpublished doctoral dissertation. CUNY.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303-304.

Shannon, R.V., Fu, Q.-J. and Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngologica Supplementum, 552,* 1-5.

Shannon, R.V. (2005). Speech and music have different requirements for spectral resolution. *International Review of Neurobiology, 70,* 121-134.

Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America, 68(3),* 836-842.

Surprenant, A.M. and Watson, C.S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal hearing listeners. *Journal of the Acoustical Society of America, 110(4),* 2085-2095.

Tye-Murray, N., Tyler, R., Woodward, G. and Gantz, B. (1992). Performance over time with a Nucleus and Ineraid cochlear implant. *Ear and Hearing, 13(3),* 200-209.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Multiple Routes to Perceptual Learning[1]

Jeremy L. Loebach, Tessa Bent, and Althea Bauernschmidt

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This research supported by NIH NIDCD R01 Research Grant DC00111, and NIH NIDCD T32 Training Grant DC00012 to Indiana University. The authors wish to thank Larry Phillips for his assistance in data collection. We would also like to thank Luis Hernandez for providing technical assistance and advice in the design and implementation of the experimental procedures.

Multiple Routes to Perceptual Learning

Abstract. A listener's ability to utilize indexical information in the speech signal can enhance their performance on a variety of speech perception tasks. It is unclear, however, whether such information plays a similar role for spectrally reduced speech signals, such as those experienced by individuals with cochlear implants. The present study compared the effects of training on linguistic versus indexical tasks when adapting to cochlear implant simulations. Listening to sentences processed with an 8-channel sinewave vocoder, three groups of subjects were trained on a transcription task (Transcription), a talker identification task (Talker ID) or a gender identification task (Gender ID). Pre- to post-test comparisons demonstrated that training produced significant improvement for all groups. Moreover, subjects from the Talker ID and Transcription training groups performed similarly at post-test and generalization, and significantly better than the subjects from the Gender ID training group. These data suggest that training on an indexical task that requires high levels of attention can provide equivalent benefit to training on a linguistic task. When listeners selectively focus their attention on the extra-linguistic information in the speech signal, they still extract linguistic information, the degree to which they do so, however, appears to be task dependent.

Introduction

The acoustic speech stream contains two different sources of information: linguistic information, which carries the meaning of the utterances, and indexical information, which specifies the characteristics of the speaker's voice (e.g. gender, age, dialect) (Ladefoged & Broadbent, 1957). How these two types of information interact during speech processing is largely unknown. Does the listener encode linguistic and indexical information in independent streams via different perceptual mechanisms, or are they encoded and processed together? The present study addressed this question by investigating how selectively focusing the listener's attention on linguistic or indexical information during training affects adaptation to spectrally degraded speech. Using sentences that had been processed by a cochlear implant (CI) simulator, we investigated how different types of training affected both perceptual learning and generalization to new sentences, talkers, and more severely spectrally degraded conditions. We found that the amount of attention required during the training task modulated the relative gain and strength of perceptual learning. Training on Talker ID, an indexical task that required a higher degree of attentional control and focus on the acoustic information in the signal, elicited more robust generalization than training on Gender ID.

Indexical Information Enhances Linguistic Processing

Indexical characteristics of talkers are important for successful interpersonal communication. A talker's particular realizations of acoustic-phonetic parameters will ultimately determine their intelligibility (Bond & Moore, 1994; Bradlow, Toretta & Pisoni, 1996; Cox, Alexander & Gilmore, 1987; Hood & Poole, 1980). Adaptation to talker idiolect is a natural part of speech perception, and adult listeners are constantly adjusting their internal categories to accommodate new talkers. Such perceptual learning, which can be defined as long-term changes in the perceptual system based on sensory experience that will influence future behaviors and responses (Goldstone, 1998; Fahle & Poggio, 2002), may play a central role in adaptation to novel talkers. When a listener is explicitly trained to classify an ambiguous sound in a word in which it does not belong (such as the word "vacation" produced with a /Z/ versus a /S/), category boundaries for words containing the sound will be adjusted to accommodate the

new pronunciation (Eisner & McQueen, 2005). This result only holds if the talker used during training is included in the test set, however (Eisner & McQueen, 2005). In this case, the phonemic distinction is relatively isolated, and listeners do not generalize to new talkers.

In addition, familiarity with a talker's voice can enhance speech perception under difficult listening conditions (Nygaard, Sommers & Pisoni, 1994). Listeners trained to identify talkers by name demonstrated better word identification accuracy than listeners who were unfamiliar with the test talkers (Nygaard et al., 1994; Nygaard & Pisoni, 1998). Two distinct types of subjects were observed: "good" learners, who exceeded 70% correct talker identification and "poor" learners, who did not (Nygaard & Pisoni, 1998). "Poor" learners performed significantly worse on word and sentence identification after training than did the "good" learners, suggesting that it is not the mere exposure to the talkers that is enhancing word identification accuracy, but rather the ability to store and utilize the acoustic information that characterize the talker's voice. When taken together, these data demonstrate the presence of significant interactions between the linguistic and indexical channels of information in speech, and suggest that the two may indeed be coded in the same stream.

Listeners can adapt not only to specific talkers but given the appropriate exposure also show talker-independent adaptation to talkers from a variety of special populations whose speech deviates from normal, native talker norms. For example, when first confronted with a non-native speaker, many listeners may have difficulty understanding them, but with exposure, they quickly learn to adapt to their speaking patterns (Bradlow & Bent, in press; Clarke & Garrett, 2004; Weil, 2001). Similarly, a beneficial effect of experience on speech intelligibility has been shown for listeners with extensive experience listening to speech produced by talkers with hearing impairments (McGarr, 1983), computer manipulated speech (Schwab, Nusbaum & Pisoni, 1985; Greenspan, Nusbaum & Pisoni, 1988; Dupoux & Green, 1997; Pallier, Sebastian-Gallés, Dupoux, Christophe & Mehler, 1998), and noise-vocoded speech (Davis, Johnsrude, Hervais-Adelman, Taylor & McGettigan, 2005). Critically, this benefit extends to new talkers, or to new speech signals created using the same types of signal degradation.

Furthermore, adaptation to a talker's idiolect may not be completely talker specific, however, if the training contrasts are lexically contrastive in the language and have a greater degree of potential generalizabillity (Kraljic & Samuel, 2006). When exposed to words containing an ambiguous sound between /d/ and /t/ in which the voicing distinction is blurred (e.g., "crocatile" or "cafederia"), subjects show robust generalization to novel utterances containing the ambiguous phoneme produced by novel talkers. Moreover, perceptual learning generalizes to a novel consonant set including an ambiguous /b/ - /p/ in which the voice onset time boundary is similarly blurred. These data suggest that when the phonemic distinction is important to more phonemes than are used in the training set (as is the case for the voicing distinction), generalization will be robust and occur independent of talker.

Compared to the literature on the perceptual learning of naturally produced speech, the explicit perceptual learning and generalization of spectrally reduced speech has received little attention. Previous research using sinewave speech has demonstrated that subjects trained to identify talkers from sentences containing three sinewave analogs of the formant frequencies show robust generalization when asked to identify these same talkers from naturally produced versions of the sentences (Remez, Fellowes & Rubin, 1997; Sheffert, Pisoni, Fellows & Remez, 2002). The effect is not bidirectional, however, since training with naturally produced speech does not generalize to sinewave speech (Sheffert et al., 2002). These data suggest that talker identification of sinewave speech may be utilizing different acoustic information than is normally used for talker identification of naturally produced speech. Since sinewave speech is derived from natural speech, some acoustic cues will be shared in both types of stimuli, promoting generalization from sinewave to naturally produced speech. When trained on naturally produced speech, however, the

listener may rely on other acoustic cues that are not preserved in the sinewave analogs, and as such, generalization to the sinewave utterances does not occur (Sheffert et al., 2002). Thus, it appears that the listener is opportunistic, relying on whatever acoustic cues are available in the signal in order to identify the talker. Although the findings with sinewave speech demonstrate that talker identification training with spectrally reduced speech generalizes to talker ID tasks for naturally produced speech, these studies have not assessed whether training on talker identification generalizes to word or sentence recognition under conditions of severe spectral degradation as has been shown for naturally produced speech (Nygaard & Pisoni, 1998).

The type of training a listener receives during adaptation to spectrally degraded speech affects the extent of perceptual learning and transfer to new materials. Feedback promotes more rapid adaptation to CI simulated speech than no feedback (Davis, et al., 2005). Moreover, the type of feedback that is given can also modulate the speed of perceptual learning. The most effective feedback includes the processed audio stimuli paired with the orthographic representation (Burkholder, 2005). Additionally, training with complex non-speech environmental stimuli promotes transfer and generalization to speech materials (Loebach & Pisoni, under review). Whether training on indexical tasks will generalize to speech perception under CI simulations, however, is unknown.

Indexical Information in Cochlear Implants

Although cochlear implants have been successful in providing the profoundly hearing impaired with access to the acoustic signal, a large amount of variability remains among cochlear implant users. While the age at onset of deafness, duration of auditory deprivation and etiology of deafness all influence outcomes after implantation, these factors do not account for all intra-subject variability (NIH, 1995). Moreover, research with CI users has focused almost exclusively on speech perception, leaving the perception of other types of acoustic signals (e.g., meaningful environmental sounds) unexplored. Although ideally individuals will achieve high levels of speech perception in quiet and noise, not all CI users will receive such a benefit. At a minimum, the individual is expected to gain some awareness of sound, including environmental stimuli (Clark, 2002).

For linguistic tasks, acoustic simulations of cochlear implants have provided a useful tool for determining what acoustic information is necessary for speech perception. Early work demonstrated that sufficient linguistic information is conveyed via acoustic simulations of a cochlear implant processor and electrode array to allow the identification of single consonants, vowels and sentences (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995). Designed to simulate different numbers of active electrodes in the intracochlear array, these simulations have demonstrated that successful speech perception is largely dependent on number of acoustic channels. Under quiet listening conditions, normal hearing subjects reach asymptote for sentences containing eight channels (Dorman, Loizou & Rainey, 1997), although more channels are needed when listening in noise (Dorman, Loizou, Fitzke & Tu, 1998). Furthermore, normal hearing subjects listening to 6 channel simulations perform similarly to cochlear implant users (Dorman & Loizou, 1998). Although limited spectral information is sufficient for high levels of consonant, vowel and sentence perception, other tasks may require substantially more spectral information. Acoustic stimuli that contain complex acoustic spectra, such as music, may require well over thirty channels to be perceived accurately (Shannon, Fu & Galvin, 2004; Shannon, 2005).

Compared to perception of linguistic information in the speech signal, considerably less is known about the perception of indexical information both in CI users, and in normal hearing subjects listening to CI simulations. Cleary and Pisoni (2002) demonstrated that prelingually deafened children with cochlear implants have more difficulty discriminating talkers based on their voices than do normal hearing

children. Moreover, considerable variability existed across subjects: over half of the children who had cochlear implants could not discriminate talkers at a level greater than chance, while those who could discriminate talkers performed comparably to the normal hearing children (Cleary, Pisoni & Kirk, 2005). When considered as a group, all children with cochlear implants required larger pitch deviations between talkers in order to distinguish them, and showed more pronounced difficulty in talker discrimination when the sentences varied across talkers than did normal hearing children (Cleary et al., 2005).

While talker discrimination may rely on acoustic details that are not well conveyed by a cochlear implant processor, gender discrimination may utilize primarily temporal cues. Normal hearing subjects listening to CI simulations require more spectral channels to accurately discriminate the gender of talkers than to identify vowels from a closed set response set (Fu, Chinchilla & Galvin, 2004). As the number of channels increase from four to thirty-two, percent correct gender identification increased approximately linearly. Moreover, a tradeoff between spectral and temporal information was observed for gender discrimination: fewer spectral channels are required when more precise temporal information is preserved. CI users' performance was roughly comparable to normal hearing subjects listening to four or eight band simulations. Moreover, accuracy depends on the individual voices of the talkers who are used in the study. If the differences between male and female talkers are large, normal hearing subjects and CI users utilize temporal information to classify the speakers' gender based on their fundamental frequency (Fu, Chinchilla, Nogaki & Galvin, 2005). Thus, it appears that CI users may be relying primarily on temporal pitch information to distinguish talkers, a strategy that becomes ineffective when the difference between male and female fundamental frequencies decreases (Fu et al., 2005).

The performance on gender identification tasks is also dependant on the method of synthesis. While speech perception accuracy does not differ for noise and sinewave vocoders (Dorman et al., 1997), gender discrimination is more accurate with sinewave than noise vocoders (Gonzalez & Oliver, 2005). Compared to noise vocoders, subjects listening to sinewave vocoders require fewer channels to reach asymptote on the gender identification task.

Gender identification and talker discrimination, however, require different types of processing compared to talker identification. The acoustic cues that allow the listener to discriminate male from female talkers or to decide if two sentences are produced by the same or different talkers may be much coarser than those required to identify a speaker from their voice alone. Vongphoe and Zeng (2005) trained normal hearing subjects and CI users to identify ten talkers and compared talker and vowel identification accuracy. Normal hearing subjects listening to sinewave vocoded vowels achieved high levels of talker identification accuracy, particularly with stimuli containing more spectral channels (e.g., 32 channels). Cochlear implant users performed significantly worse than the normal hearing listeners. For vowel recognition, however, performance by CI users approximated the normal hearing subjects listening to 8-channel vocoders. The differences in performance of the CI users on the vowel and talker identification tasks led the authors to conclude that the subjects may be utilizing different processing strategies during linguistic and indexical tasks (Vongphoe & Zeng, 2005).

One possible confound in the study, however, comes from the overlap in the fundamental frequencies of the talkers voices. When considered on a talker-by-talker basis the predominant source of errors in talker identification was not between adult male and adult female talkers, but from confusions between the voices of adult females, girls and boys (Vongphoe & Zeng, 2005). Given that the dominant confusions were between talkers with higher pitched voices, the conclusion that linguistic and indexical tasks may utilize two independent processes may be premature. When boys and girls are excluded from the analysis, the CI users resemble the normal hearing subjects listening to 8-channel simulations, as they did in the vowel identification task. Rather than concluding that two separate processes are involved,

these data may suggest that when the listener must make fine spectral distinctions, such as is required to distinguish talkers who share a similar range of vocal pitch, both CI users and normal hearing subjects listening to CI simulations perform comparably due to similar processes.

The Present Study

Understanding how linguistic and indexical information interact in speech perception may provide new insight into possible training methodologies for newly implanted individuals. Given that there are no standardized training and rehabilitation protocols available to CI users, the source of the variability in benefit and outcome are further confounded with experience. Would listeners benefit from explicit training after implantation, and if so, what type of training is most appropriate? Given that most previous research has focused exclusively on linguistic tasks (Fu, Galvin, Wang & Nogaki, 2005), it is unknown whether training on nonlinguistic tasks will also promote robust generalization and transfer. Moreover, does the level of attention required to perform the training task modulate the amount of learning that is observed following training?

The present study compared how training on a linguistic versus indexical task affected listeners' ability to accurately perceive words in sentences. Using sentences processed with an 8-channel sinewave vocoder, normal hearing subjects were trained to identify either the gender or identity of six talkers, or transcribe their speech. Pre- to post-test comparisons of transcription accuracy scores assessed the effectiveness of training. Given the results of previous studies, we hypothesized that subjects trained on talker identification would perform better than those who were trained on gender identification. Moreover, we predicted that training on talker identification would match or exceed the performance of subjects trained on sentence transcription due to increased perceptual attention required to learn to identify the talkers from such severely spectrally degraded stimuli.

Method

Subjects

Seventy-eight normal-hearing young adults participated in the study (60 female, 18 male; mean age 21 years). All subjects were native speakers of American English. Most (n = 69) were monolingual, with only nine reporting being fluent speakers of more than one language. Subjects were recruited from the Indiana University community, and either received monetary compensation for their participation ($10 per session) or course credit in an Introductory Psychology class (1 credit per session). Of the seventy-eight subjects tested, six were excluded from the final data analysis (two failed to return for the generalization session, one failed to return in a timely manner, and three due to program errors). Of the 72 remaining subjects, 43 returned for the follow up portion of the experiment.

Stimuli

Stimuli consisted of 212 meaningful (116 high predictability (HP), 48 low predictability (LP)), and 48 anomalous (AS) SPIN sentences (Kalikow, Stevens & Elliott, 1977; Clopper, Carter, Dillon, Hernandez, Pisoni, Clarke, Harnsberger & Herman, 2002). SPIN sentences are phonetically balanced for phoneme occurrence in English, and contain between five and eight words, the last of which is the keyword to be identified. In the HP sentences, the final word is highly constrained by the preceding semantic context (e.g., "A bicycle has two wheels."), whereas in the LP sentences the preceding context is uninformative (e.g., "The old man talked about the lungs."). The AS sentences retain the overall format of their meaningful counterparts, except that all words in the sentence are semantically unrelated,

resulting in a sentence that preserves proper syntactic structure, but is semantically anomalous (e.g., "The round lion held a flood."). A passage of connected speech (Rainbow Passage; Fairbanks, 1940) was used during the familiarization portion of the experiment. Wavefiles of the materials were obtained from the Nationwide Speech Corpus (Clopper, 2004). Materials were produced by 8 speakers (4 male, 4 female) from the midland dialect.

Synthesis

Stimulus processing was conducted in Tiger CIS (http://www.tigerspeech.com/) and simulated an 8-channel cochlear implant using the CIS processing strategy. Stimulus processing involved two phases, an analysis phase, which divided the signal into bands and derived the amplitude envelope from each band; and a synthesis phase, which replaced the frequency content of each band with a sinusoid that was modulated with its matched amplitude envelope. Analysis used band-pass filters to divide the stimuli into 8 spectral channels between 200 and 7000 Hz with corner frequencies based on the Greenwood function (24 dB/octave slope). Envelope detection used a low pass filter with an upper cutoff at 400 Hz and a 24 dB/octave slope. Subsets of the materials to be used in the generalization phase were processed with four and six channels, to further reduce the amount of information in the signal. All stimuli were saved as 22 kHz sampling rate 16-bit windows PCM wav files, and normalized to 65 dB RMS (Level v2.0.3, Tice & Carrell, 1998) to ensure that stimuli were equal in intensity across all materials, and that no peak clipping occurred.

Procedures

All methods and materials were approved by the Human Subjects Committee and Institutional Review Board at Indiana University Bloomington. For data collection, a custom script was written for PsyScript and implemented on four Apple PowerMac G4 each with a 15-inch color LCD monitor. Audio signals were presented over Beyer Dynamic DT-100 headphones, calibrated with a voltmeter to a 1000 Hz tone at 70 dBv SPL. Sound intensity was fixed within PsyScript in order to guarantee consistent sound presentation across subjects. Multiple booths in the testing room accommodated up to four subjects at the same time. Before the presentation of each audio signal, a fixation cross was presented at the center of the screen for 500 milliseconds to alert the subject to the upcoming trial. Following stimulus offset, the subject was prompted to make their response. A 1000 millisecond interval separated each trial. For the transcription trials, a dialog box was presented on the screen prompting subjects to type in what they heard. For talker identification, subjects clicked on the one box (out of six) that contained the name of the talker that produced the sentence. For gender identification, subjects clicked on a box labeled "female" or "male". There were no time limits for responding, and subjects pressed a button to advance to the next trial. Subjects performed at their own pace, and were allowed to rest between blocks as needed. The experimental session lasted approximately 40-60 minutes.

Training. Training took place over two sessions. The materials and tasks varied across blocks, but the same block structure was used for all groups, and all stimuli were randomized within each block. Session 1 began with two pre-test blocks in order to establish a baseline level of performance before training (Table 1). In block 1, subjects transcribed 30 unique LP sentences, and 30 unique AS sentences in block 2. In these blocks, the subjects simply transcribed the sentences, and received no feedback.

In the familiarization phase (Block 3) subjects passively listened to the Rainbow passage produced by each of the six talkers in order to familiarize them with the voices and synthesis condition, and teach them the appropriate labels that would be used during training. Although subjects in all three training groups heard the same materials, they were required to make different responses during training.

During familiarization, subjects in the Talker ID group were presented with the passage paired with the name of the talker who produced it (Jeff, Max, Todd, Beth, Kim, Sue). Subjects were informed that they would be asked to identify the talkers by name, and to listen carefully for any information that would help them learn to recognize the talkers' voice. Subjects in the Gender ID group heard the same passages, but paired with the appropriate gender label (Male or Female) for each talker. These subjects were informed that they would be asked to identify the gender of the talkers, and to listen carefully for any information that would help them learn to recognize each talker's gender. Subjects in the Transcription group heard each passage presented along with the name of the talker who produced it (Jeff, Max, Todd, Beth, Kim, or Sue), but were informed that they would be asked to transcribe sentences produced by each talker, and to listen carefully in order to better understand the degraded signals.

Blocks 1-2	Block 3	Blocks 4-6
Pre-test	Familiarization	Training
Transcribe: 30 LP and 30 AS sentences	Passively listen: Rainbow passage	Transcribe, ID Talker, or ID Gender: 150 HP sentences

Table 1. Session 1 assessed the pre-test transcription abilities of subjects before training, familiarized them with the talkers and materials, and initiated training. The tasks that subjects performed and the materials that were presented in each block of Session 1 are listed in the table.

The training blocks (4, 5 and 6) consisted of 150 HP sentences. Each talker produced the same 25 sentences, so that subjects would hear six versions of each sentence in order to learn characteristics of the individual voices. During the training trials, subjects were presented with a sentence and asked to make a response appropriate for their training group. Subjects in the Talker ID group were asked identify the correct talker by clicking one of six buttons on the computer screen labeled with the talkers' names. After the subject indicated their response, a red circle appeared around the name of the correct talker as feedback. Subjects in the Gender ID group responded by clicking one of two buttons on the computer screen that contained the appropriate gender label. After the subject indicated their response, a red circle appeared around the correct gender of the talker as feedback. Subjects in the Transcription training group were asked to type what they thought the talker said, and received the correct transcription of the sentence as feedback. For all training groups, feedback was provided regardless of the accuracy of the subject's response.

Session 2 (Table 2) was completed within 3 days of session 1, and began with a repetition of the familiarization phase (block 7) in which subjects again heard the rainbow passage produced by each talker. The purpose of this block was to re-familiarize the listener with the voices and labels, since at least 24 hours had passed since the first training session. Two training blocks followed, consisting of 90 HP sentences. Again, subjects received feedback regardless of their performance.

Generalization and transfer of training were tested in blocks 10, 11 and 12, and subjects were asked to transcribe novel materials that they had not heard earlier during the experiment. In block 10, the transfer of training to more severe spectral degradation was assessed using 36 unique HP sentences, half of which were processed with 4-channel sinewave vocoder, and the other half with a 6-channel sinewave vocoder. Generalization of training to novel materials by familiar talkers was assessed in block 11 with 18 AS and 18 LP sentences processed with the same 8-channel vocoder used during training. In block 12,

transfer of perceptual learning to novel talkers was assessed using 20 unique HP sentences produced by two new talkers (1 male, 1 female). Following generalization, two post-test blocks (13 and 14) assessed the relative gains in performance due to training. In block 13 subjects transcribed a selection of twelve AS sentences from pre-test block 2, whereas in block 14 subjects transcribed a selection of twelve LP sentences selected from pre-test block 1.

Block 7	Blocks 8-9	Block 10	Block 11	Block 12	Blocks 13-14
Familiarization	Training	Generalization: More degraded	Generalization: Novel materials	Generalization: Novel talkers	Post-test
Passively listen: Rainbow passage	Transcribe, ID Talker, or ID Gender: 90 HP sentences	Transcribe: 18 HP (4-band) 18 HP (6-band) sentences	Transcribe: 18 AS &18 LP sentences	Transcribe: 20 HP sentences	Transcribe: 12 AS & 12 LP sentences (from pre-test)

Table 2. Session 2 featured a continuation of training, followed by tests of generalization to new materials and the post-test (both transcription tasks). The tasks that subjects performed and the materials that were presented in each block of Session 2 are listed in the table.

Retention. One month after the initial training sessions, subjects returned for a third session to assess long-term retention of training (Table 3). During the retention test, subjects transcribed the same materials from generalization and post-test blocks 10 through 14. The purpose of this retention session was to assess how well perceptual learning was maintained over time, and to discern whether training differentially affected the long-term retention of training.

Block 15	Block 16	Block 17	Blocks 18-19
Generalization: More degraded	Generalization: Novel materials	Generalization: Novel talkers	Post-test
Transcribe: 18 HP (4-band) 18 HP (6-band) sentences	Transcribe: 18 AS & 18 LP sentences	Transcribe: 20 HP sentences	Transcribe: 12 AS & 12 LP sentences (from pre-test)

Table 3. Session 3 occurred 1 month after session 2, and tested subjects abilities to transcribe the materials that they experienced in session 2 to assess the stability of training over time. The tasks that subjects performed and the materials that were presented in each block of Session 3 are listed in the table.

Analysis and Scoring. Keyword accuracy scores were based on the final word in each sentence. Common misspellings and homophones were counted as correct responses, but words with added or deleted morphemes were counted as incorrect. Perceptual learning during training was assessed by comparing performance across the five training blocks. Pre- to post-test comparisons provided an assessment of the relative gains from training across the three training groups. Comparison of performance at pre- and post-test to performance on new materials provided an assessment of generalization of training to novel stimuli. Generalization was said to have occurred if performance was significantly higher than the pre-test and greater than or equal to that at post-test. Comparison of pre- and

post-test performance to performance on new talkers provided an assessment of transfer of training to novel talkers. Comparisons of performance on the 4-band and 6-band stimuli provided an assessment of how well training transferred to more severely degraded stimuli. Comparison of performance in session 2 with performance in session 3 provided an estimate of long-term retention of training. A measurement of savings was calculated for each type of material by dividing performance in session 2 by that in session 3 and normalizing to one (e.g., $4\text{-band}_{savings} = 1 - (4\text{-band}_2/4\text{-band}_3)$). This provided an estimate of how robust perceptual learning was over time.

Results

Perceptual Learning during Training

Accuracy on the training tasks varied by training group (Figure 1). Subjects in the Gender ID and Transcription training groups performed near ceiling and subjects from the Talker ID group performed just above chance.

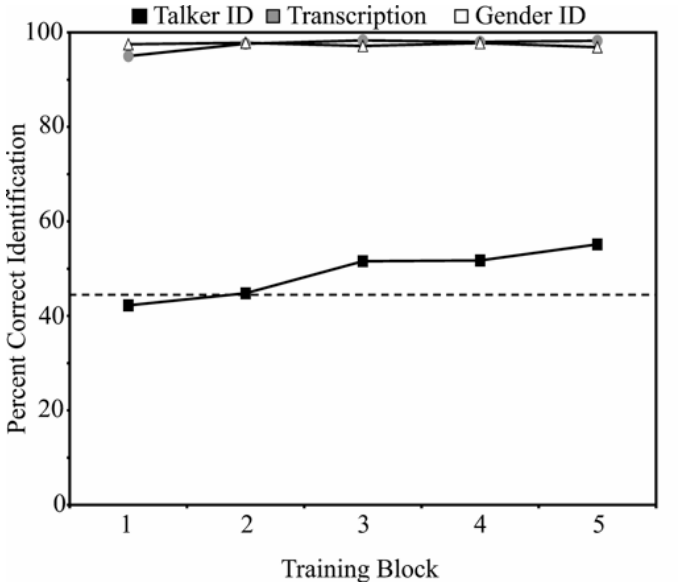


FIGURE 1. Perceptual learning across the five training blocks. The dashed horizontal line indicates the level of performance that subjects must exceed in order to be considered significantly different from chance in the talker identification condition. Subjects trained to transcribe the sentences (Transcription) appear as filled circles. Subjects trained to identify the gender of the talker (Gender ID) appear as filled triangles. Subjects trained to identify the talkers by their voices (Talker ID) appear as filled squares.

Subjects in the Transcription training group performed extremely well across all five training blocks. In block 1, subjects correctly identified 95% of the keywords and performance reached ceiling in block 2 (98% correct) and remained at ceiling for the last three training blocks. A univariate ANOVA revealed a significant main effect of Block ($F\ (4, 190) = 6.441$, $p < 0.001$), indicating that subjects showed improvement across training blocks. Post hoc Bonferonni tests revealed that subject performance in block 1 was significantly lower then performance in all other blocks (all $p < 0.009$). Performance in blocks 2 through 5 did not differ from one another (all $p > 0.88$). A trend toward a main effect for Talker

Gender was observed (F *(1, 190)* = 3.156, p = 0.077), with female speech being transcribed more accurately than male speech.

Subjects' accuracy in the Gender ID training condition was also extremely high across all five training blocks. Subjects' ability to identify the gender of the talkers was at ceiling (>95%) in all training blocks. Main effects for Block (F *(4, 190)* = .228, p = 0.922), and Talker Gender (F *(1, 190)* = 1.324, p = 0.251) were not observed, indicating that subject performance did not vary across blocks, and was equal for male and female talkers.

Performance of the Talker ID group was considerably more variable across subjects. Since inter-gender confusions (identifying male talkers as female, or female talkers as male) were rare, occurring less than 2 percent of the time, a more conservative level of chance was used (1 out of 3 rather than 1 out of 6). According to the binomial probability distribution, performance must be at least 44.46% correct to significantly exceed chance. Most subjects (n = 26) were able to identify talkers at a level greater than chance beginning in block 2 and showed improvement as training progressed (Block 1: 42.2%, Block 2: 44.8%, Block 3: 51.6%, Block 4: 51.7%, Block 5: 55.1%). A univariate ANOVA revealed a significant main effect of Block (F *(4, 250)* = 9.428, $p < 0.001$) with subject performance improving significantly between blocks 1 and 5 ($p < 0.001$). A significant main effect of Talker Gender was also observed (F *(1, 250)* = 39.509, $p < 0.001$), with subjects identifying female talkers (54%) more accurately than male talkers (44%).

Performance after Training

Pre- to Post-test Comparisons. Overall, the type of training a subject received determined how well they performed at post-test; however, all subjects showed significant gains in sentence transcription accuracy due to training (Figure 2). For the subjects in the Transcription training group, performance increased from 51% correct for the meaningful sentences at pre-test to 77% correct at post-test. Similar gains were observed for the anomalous sentences increasing from 60% correct at pre-test to 75% at posttest. A univariate ANOVA revealed a significant main effect of Materials (F *(3, 152)* = 32.136, $p <$ 0.001), with subjects performing significantly better on anomalous than meaningful sentences at pre-test but not at posttest ($p < 0.001$ and p = 0.977, respectively). This effect is likely due to exposure, since the anomalous sentence pre-test always came after the meaningful sentence pre-test. This difference of 9% is within the normal range of gains expected from merely being exposed to the stimuli without engaging in explicit training, as documented by Davis and colleagues (2005). Furthermore, a significant main effect of Talker Gender (F *(1, 152)* = 5.939, p = 0.016) was observed. Subjects were significantly more accurate at transcribing the speech of female talkers than male talkers.

Training on Gender identification successfully transferred to sentence transcription (Figure 2). For meaningful sentences, performance increased from 45% at pre-test to 69% correct at post-test. Similar gains were observed for the anomalous sentences increasing from 56% correct at pre-test to 69% at posttest. An ANOVA revealed a significant main effect of Materials (F *(3, 152)* = 25.959, $p < 0.001$), indicating that performance varied according to the type of materials subjects were asked to transcribe. At pre-test, subjects performed significantly better on the anomalous sentences than the meaningful sentences (p = 0.006), but were identical at posttest (p = 1.00). A significant main effect was observed for Talker Gender (F *(1, 152)* = 10.222, p = 0.002), again indicating that subjects were significantly more accurate at transcribing the speech of female talkers than male talkers.

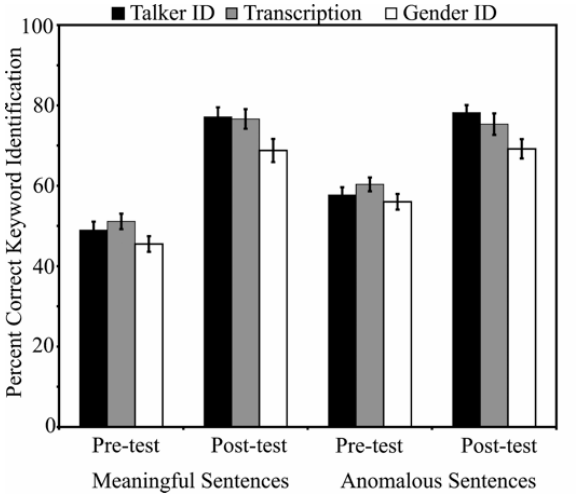


FIGURE 2. Percent correct keyword identification scores for subjects trained on talker identification (Talker ID), gender identification (Gender ID) or sentence transcription (Transcription) on the pre- and post-test materials.

For subjects in the Talker ID group, a significant main effect of Materials was also observed (F *(3, 200)* = 69.555, $p < 0.001$). For meaningful sentences (Figure 2), subjects improved significantly from pre- (48% correct) to post-test (75% correct, $p < 0.001$). Similar findings were observed for the anomalous sentences, with performance increasing significantly from 56% correct to 79% correct ($p < 0.001$). As was observed for subjects in the Transcription and Gender ID groups, performance was significantly better on anomalous sentences than meaningful sentences at pre-test ($p = 0.003$), but identical at post-test ($p = 0.652$). A significant main effect of Talker Gender was also observed (F *(1, 200)* = 72.664, $p < 0.001$) indicating that the materials produced by female talkers were correctly transcribed significantly more accurately than those produced by male talkers.

Univariate ANOVAs comparing the scores of all three training groups revealed that pre-test performance did not differ across training groups for the anomalous (F *(2, 126)* = 1.356, $p = 0.262$) or meaningful sentences (F *(2, 123)* = 2.569, $p = 0.081$) indicating that subjects in all groups performed at a comparable level before training began. Differences in performance emerged at post-test, for both the meaningful (F *(2, 126)* = 3.656, $p = 0.029$) and anomalous sentences (F *(2, 126)* = 4.234, $p = 0.017$). In both cases, subjects in the Gender ID training group performed less accurately than subjects in the Talker ID training ($p = 0.036$, $p = 0.013$) and Transcription training groups ($p = 0.075$, $p = 0.156$).

Generalization to New Materials. Overall, training successfully generalized to the transcription of novel sentences produced by familiar talkers (Figure 3). Transcription training successfully generalized to new meaningful sentences produced by the familiar talkers (85.7%). A univariate ANOVA revealed that there was a significant main effect of Session (F *(3, 126)* = 96.629, $p < 0.001$), and Bonferonni tests indicated that subjects performance was significantly better for novel meaningful materials than at pre-test ($p < 0.001$) or post-test ($p = 0.014$). A similar finding was observed for the new anomalous sentences (F *(2, 114)* = 25.974, $p < 0.001$), and subjects performed significantly better on the novel anomalous sentences (79.1%) than at pre-test ($p < 0.001$) but not at post-test ($p = 0.175$). Additionally, a significant main effect of Talker Gender was observed for both meaningful (F *(1, 126)* = 6.741, $p = 0.010$) and anomalous sentences F *(1, 114)* = 5.462, $p = 0.021$), indicating that female talkers were again transcribed more accurately than male talkers.

Subjects trained to identify talker gender showed robust generalization to new meaningful (76.7%; *F (3, 126)* = 55.096, $p < 0.001$) and anomalous sentences (74.4%; *F (2, 114)* = 17.593, $p < 0.001$). For both anomalous and meaningful sentences, performance on the new materials was significantly higher than pre-test (both $p < 0.001$) and did not differ from post-test (both $p > 0.09$). A significant main effect of Talker Gender was observed for the new meaningful sentences (*F (1, 126)* = 10.058, $p = 0.002$), but not new anomalous sentences (*F (1, 114)* = 2.746, $p = 0.10$).

Subjects trained on Talker ID also showed robust generalization to new meaningful (84.7%; *F (3, 200)* = 136.095, $p < 0.001$) and anomalous sentences (81.1%; *F (2, 150)* = 58.199, $p < 0.001$). For both training groups, performance on the new materials was significantly more accurate than pretest (all $p < 0.001$) and was greater than (meaningful sentences $p < 0.001$) or equal to (anomalous sentences $p = 1.00$) performance at post-test. A significant main effect of Talker Gender was observed for both new meaningful sentences (*F (1, 200)* = 38.217, $p < 0.001$) and anomalous sentences (*F (1, 150)* = 30.201, $p < 0.001$), and subjects were more accurate in transcribing the female talkers than the male talkers.

Comparison of performance on the meaningful sentences across all three groups using a univariate ANOVA revealed a significant main effect of Training (*F (2, 126)* = 6.403, $p = 0.002$). Subjects in the Transcription group performed nearly identically to subjects in the Talker ID group ($p = 0.932$), and both groups performed significantly better than the subjects in the Gender ID group ($p = 0.015$ and $p = 0.003$ respectively). In addition, a trend toward a significant main effect of Talker Gender was observed (*F (1, 126)* = 3.724, $p = 0.056$) indicating that female talkers were transcribed with more accuracy than male talkers. Although a main effect of Training was not observed for the novel anomalous sentences (*F (2, 126)* = 2.795, $p = 0.067$), a trend was observed for subjects in the Transcription training group to perform better than subjects in the Gender ID training group ($p = 0.073$). A significant main effect of Talker Gender was also observed (*F (1, 126)* = 18,769, $p < 0.001$) with subjects transcribing female talkers more accurately than male talkers.

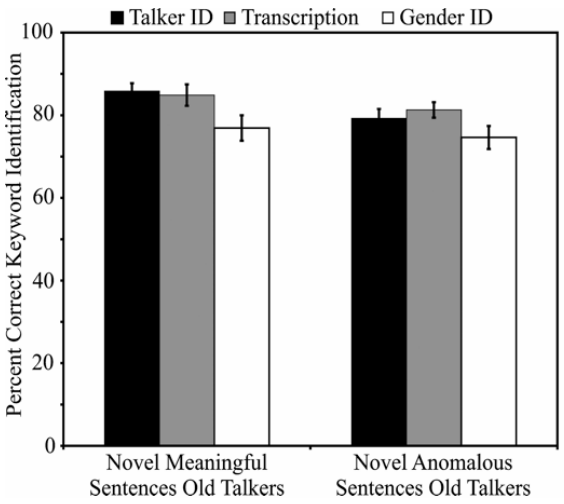


FIGURE 3. Percent correct keyword identification scores on the new anomalous and meaningful sentences produced by familiar talkers (session 2, block 11) for subjects trained on talker identification (Talker ID), gender identification (Gender ID) or sentence transcription (Transcription).

Transfer of Training to Increased Spectral Degradation. Subjects showed a graded response to stimuli that were more severe spectral degraded (Figure 4). Overall, subjects were more accurate at transcribing sentences in the 6-band processing condition (Transcription: 83.1%; Gender ID: 78.6%; Talker ID: 88.9%) than sentences in the 4-band processing condition (Transcription: 51.7%; Gender ID: 56.4%; Talker ID: 61.9%). A univariate ANOVA reveled a significant main effect of Processing for all groups (Transcription ($F\ (1, 76)$ = 69.104, $p < 0.001$); Gender ID ($F\ (1, 76)$ = 29.731, $p < 0.001$); Talker ID ($F\ (1, 100)$ = 120.846, $p < 0.001$)), indicating that subjects performed significantly better on the 6-band sentences than the 4-band sentences. The main effect of Talker Gender was not significant for the Transcription training group ($F\ (1, 76)$ = .066, p = 0.798), or the Gender ID training group ($F\ (1, 76)$ = 2.248, p = 0.138), indicating that subjects performed equally well on male and female speech. Subjects in the Talker ID training group, however, did show a significant main effect of Talker Gender ($F\ (1, 100)$ = 9.094, p = 0.003), indicating that they transcribed the speech of female talkers more accurately than male talkers.

Comparison of the performance on the 4-band processed sentences across training groups using a univariate ANOVA revealed a significant main effect of Training ($F\ (2, 126)$ = 4.44, p = 0.014). Subjects in the Transcription training group performed significantly better than subjects in the Talker ID group (p = 0.01), but did not differ from talkers in the Gender ID group (p = 0.399). Subjects in the Talker ID training group performed similarly to subjects in the Gender ID group (p = 0.359). The main effect of Talker Gender was not significant ($F\ (1, 126)$ = .933, p = 0.336). Comparison of performance on the 6-band stimuli across training groups also revealed significant main effect of Training ($F\ (2, 126)$ = 4.702, p = 0.001). Subjects in the Transcription group performed as well as subjects in the Talker ID group (p = 0.465), but significantly better than subjects in the Gender ID group (p = 0.008). Subjects in the Gender ID group performed as well as subjects in the Talker ID group (p = 0.213). A significant main effect of Talker Gender was observed ($F\ (1, 126)$ = 8.273, p = 0.005), and subjects were significantly more accurate at transcribing the speech of female talkers than male talkers.

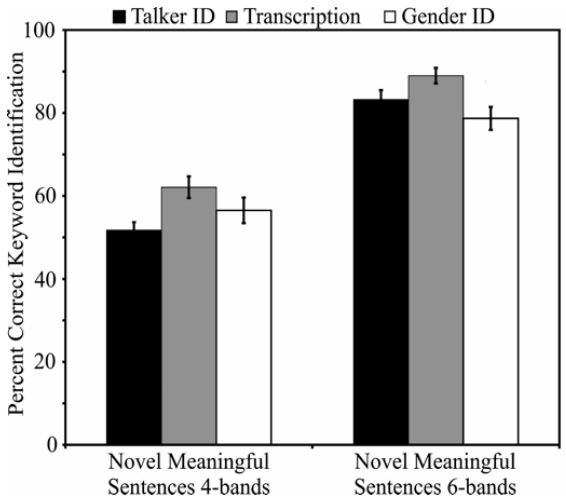


FIGURE 4. Percent correct keyword identification scores for subjects trained on talker identification (Talker ID), gender identification (Gender ID) or sentence transcription (Transcription) on the meaningful sentences produced by familiar talkers but processed to have more severe spectral degradation (Block 10).

Transfer of Training to Novel Talkers. Transcription of novel sentences produced by unfamiliar talkers was equivalent to or better than transcription of meaningful sentences produced by familiar talkers (Transcription 92.3% correct; Gender ID 85% correct, Talker ID 93% correct). For all training groups, performance on new talkers was significantly higher than pre-test and post-test (both $p < 0.001$) suggesting that talker familiarity may not necessarily enhance transcription accuracy on CI simulations as compared to other types of spectral degradation (e.g. noise). Moreover, training-induced differences in performance were also observed (Figure 5), and a significant main effect of Training was again noted ($F\ (2, 126) = 6.874$, $p < 0.001$). Subjects from the Transcription and Talker ID training groups performed the same ($p = 0.951$), and significantly better than subjects in the Gender ID group ($p = 0.004$ and $p = 0.005$, respectively).

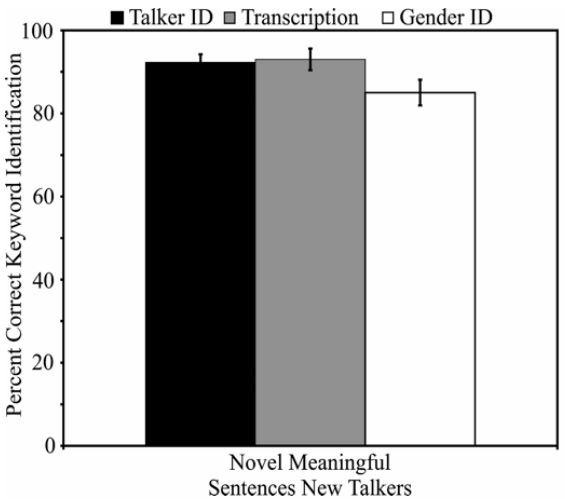


FIGURE 5. Percent correct keyword identification scores for subjects trained on talker identification (Talker ID), gender identification (Gender ID) or sentence transcription (Transcription) on the meaningful sentences produced by novel talkers (block 12).

Retention of Training.

Of the 72 subjects who participated in sessions 1 and 2, 43 returned for retention testing in session 3. Since fewer subjects overall participated in session 3, data were matched such that the analyses only compared the performance of subjects who attended all three sessions. A one-way ANOVA comparing performance in session 2 with that in session 3 revealed all subjects in the Transcription training group improved their performance on the 4-band stimuli ($F\ (1, 34) = 9.092$, $p = 0.015$) from 57% in session 2, to 73% in session 3. In session 3, performance on all other materials (6-band, new anomalous sentences, new meaningful sentences, new talkers, post-test anomalous, post-test meaningful) did not change from session 2 (all $p > 0.1$). Subjects in the Gender ID training group also showed significant gains on the 4-band stimuli in session 3 ($F\ (1, 46) = 4.713$, $p = 0.035$), improving from 61% in session 2 to 70% in session 3. Improvements were also observed for the new meaningful sentences ($F\ (1, 46) = 6.595$, $p = 0.014$), which increased from 79% in session 2 to 87% in session 3. Performance on all other materials (6-band, new anomalous sentences, new talkers, post-test anomalous, post-test meaningful) did not change from session 2 (all $p > 0.1$). Subjects in the Talker ID group showed significant improvement on the 4- (51 to 66%, ($F\ (1, 58) = 14.236$, $p < 0.001$) and 6-band stimuli (85 to 93%, ($F\ (1, 58) = 5.353$, $p = 0.024$). Performance on all other materials (new anomalous sentences, new

meaningful sentences, new talkers, post-test anomalous, post-test meaningful) did not change from session 2 to session 3 (all $p > 0.1$).

It is important to note that these retention tests included the same materials that appeared in the post-tests and generalization tests, so these measures are purely designed to show whether training is stable over time rather than to assess generalization to novel materials or conditions. To this end, a measure of savings was employed that divided the performance in session 2 by the performance in Session 3 and subtracting the result from one (e.g., 1- (Post-test$_3$/Post-test$_2$)) in order to determine the percent gain or loss that subjects received for each type of material (Figure 6). Across all materials, subjects in the Gender ID group showed the largest gain from Session 2 to Session 3 (increasing overall by 54%) followed by subjects in the Talker ID group (38%) and subjects in the Transcription training group (12%). The largest gains for all groups were observed for the 4-band vocoded stimuli, demonstrating that previous exposure to the more severely spectrally degraded materials tended to improve performance most at retention.

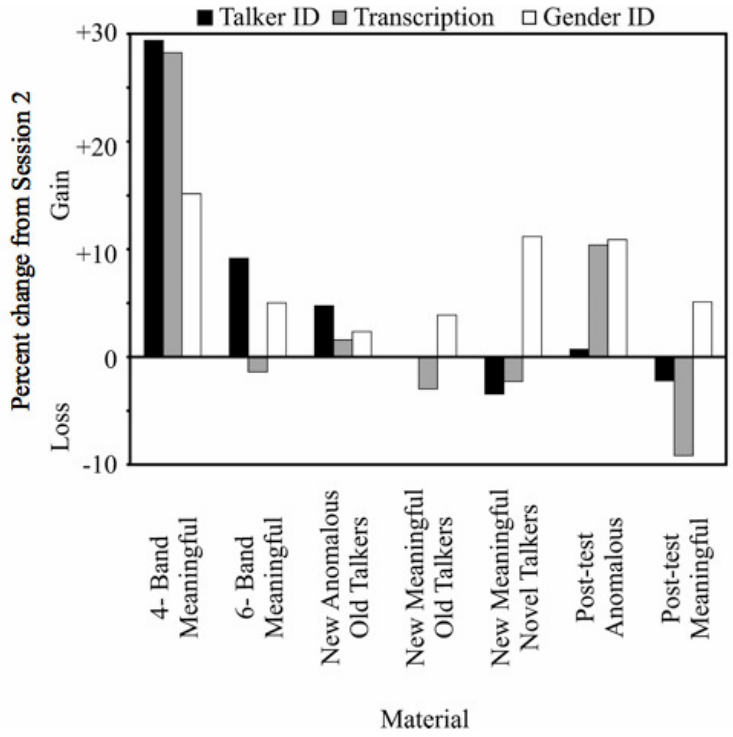


FIGURE 6. Percent gain or loss across groups as a function of testing materials. The amount of savings was calculated by dividing performance in Session 2 by performance in Session 3 and subtracting one from the result.

Talker ID Training: Subgroups.

An additional finding of the present study emerged when first assessing subject performance on the Talker ID training task. As noted earlier, most (n = 26) subjects could be trained to successfully identify talkers at a level greater than chance (44.3%). There was an additional subset of subjects, however, who could not, and were excluded from the analysis for the Talker ID group. Unlike the good

learners, these poor learners (n = 5) were never able to identify talkers at a level greater than chance in any of the blocks (Block 1: 30.8%, Block 2: 35.4%, Block 3: 36.3%, Block 4: 34.7%, Block 5: 32.9%), as indicated by a univariate ANOVA ($F\ (4, 40) = 0.05$, $p = 0.628$). A significant main effect of Talker Gender was found, however ($F\ (1, 40) = 9.941$, $p = 0.003$), revealing that female talkers were correctly identified significantly more often (38%) than male talkers (30%), as was the case in the good learners.

Furthermore, subjects who could not identify the talkers at a level exceeding chance performed significantly more poorly on the transcription tasks than the subjects who were proficient at talker identification. A series of one-way ANOVAs revealed that performance did not differ at pre-test for either the meaningful ($p = 0.105$) or anomalous sentences ($p = 0.310$). After training, however, a significant main effect of Group was observed for all materials (all $p < 0.003$), indicating that although subjects performed the same at pre-test, their performance increased at a different rate depending on how well they could perform the training task. Such a result is not likely to be caused by inattention, or laziness on the part of the participants in the poor learning group, since the transcription errors they made were phonologically related to the target words, and response omissions were no more prevalent than in the good learning group. Rather, it appears that the ability to detect and utilize acoustic information important for the indexical training task is related to the ability to extract acoustic information important for recognizing the linguistic content of utterances.

Discussion

The present study compared training that selectively focuses the listener's attention on the indexical information in the speech signal to training that focuses entirely on the linguistic content. Although all three types of training in this experiment produced significant pre- to post-test gains in performance, talker identification and sentence transcription training appeared to provide the largest and most robust overall improvement (Figure 2). Generalization to new materials and talkers was equivalent for the talker identification and transcription trained subjects, both of whom performed better than the subjects trained on gender identification (Figure 3 and Figure 5). Generalization to materials that were more spectrally degraded showed a mixed pattern of results (Figure 4). For stimuli that were more severely spectrally degraded (4- and 6- band), subjects trained on sentence transcription performed best, subjects trained on gender identification performed least accurately and subjects trained on talker identification displayed an intermediate level of performance. No effect of talker familiarity was observed. Subjects performed as well or better on the new talkers than they did on the old talkers, suggesting that the benefit of talker familiarity may not be as robust under cochlear implant simulations as compared to other forms of degradation (e.g., noise). However, baseline intelligibility for these talkers has not been established so it is possible that the talkers in the "new talker" condition were intrinsically more intelligible than the "old talkers" used in the training blocks.

Two main conclusions can be drawn from these data. The first is that training on an indexical task yields equivalent results to traditional linguistic training using transcription tasks if the task demands are high enough to require sustained attention. Evidence for this comes from the across group comparisons of post-test and generalization scores for the subjects in the Talker ID group, who performed similarly to the subjects in the Transcription training group, but significantly better than the subjects in the Gender ID training group (Figure 2). Compared to gender identification (which was at ceiling in the first training block), talker identification training is a difficult task under cochlear implant simulations, requiring high levels of attention and focus. Moreover, when a listener is exposed to a speech signal that is meaningful in their native language they cannot help but to process it as such. Even though subjects' attention in the Talker and Gender ID tasks were not directed toward the linguistic

information in the signal, presumably they still processed the linguistic content of the sentences automatically.

The second main finding is that the benefit of exposure may be determined by whether the subject can successfully access the acoustic information in the speech signal. Subjects in the Talker ID group, who had to make fine acoustic distinctions among voices, performed significantly better than subjects in the Gender ID group. Moreover, the subjects from the Talker ID group who could not learn to identify the talkers at a level greater than chance performed significantly worse on sentence transcription than subjects who could identify the talkers. Taken together, these findings suggest that the access and attention to fine acoustic details learned during talker identification training may enhance a listener's ability to extract linguistic information.

Differences in Task Demands and Attentional Resources

The data from the present study suggest that interactions between attentional demands and task difficulty may play a large role in determining the amount of benefit that a subject will receive from training. Talker identification under a CI simulation is considerably more difficult than under normal acoustic conditions. The acoustic information that specifies the voice of the talker in the natural signal appears to be significantly degraded when processed through a cochlear implant speech processor, whereas the acoustic information needed to successfully identify the gender of a talker under 8 channel CI-simulation is relatively intact. Thus, the task demands placed on a listener are significantly higher in a talker identification task than those in a gender identification task. Subjects in the Talker ID training group, while performing significantly greater than chance, only achieved an average score of 55% correct talker recognition on the final day of training; subjects in the gender ID training group were at ceiling from the first training block. These results suggest that the identifying characteristics of a talker's voice may rely on detailed spectral cues within specific frequency regions. Such cues are not well preserved in a cochlear implant. Gender identity cues, on the other hand, may rely more on spectral information across a wider range of frequencies and the relative spectral weighting of information in each frequency band in the vocoder may allow listeners to perform more accurately.

The differences in the availability of acoustic information may have produced differences in task demands. More attention is required when making fine-grained distinctions between talkers' voices, and comparably less is required to distinguish genders. These differences in attentional requirements may explain the differences in post-test gains and strength of generalization. Subjects who were required to perform a more demanding task during training performed better in the post-test and generalization phase than subjects who performed a less demanding task. Additionally, talker identification may require the utilization of cues from many different aspects of phonological structure (i.e., prosody, stress patterns, speaking rate, etc.), which are apparent in longer speech samples and require sustained attention for a longer period of time as compared to cues for gender identity. After the experiment, some subjects in the talker identification group said that they focused their attention on distinctions in overall speaking patterns and pronunciation habits in order to distinguish the talkers. As such, listening attentively to longer samples of speech may have resulted the perception of more of the linguistic information in the signal. If subjects in the gender identification group could make a decision more rapidly based on lower level acoustic cues, they may not have attended to the signal as long, and may not have received as much of a benefit from the mandatory linguistic processing.

One might expect subjects in the Talker and Gender ID training conditions to perform worse on the post-test and generalization tests than subjects in the Transcription training condition due to subjects performing fundamentally different tasks than they were trained on. Subjects in the Transcription group,

however, performed similarly to those in the Talker ID group, suggesting that training on two different tasks can produce an equivalent benefit. For subjects in the talker identification group, perceptual learning transferred from training to testing even though they were performing a different task in each condition. Subjects in the Transcription group help to establish what levels of generalization should be expected, since they performed the same task during both training and testing. Subjects performed similarly in the talker identification condition, but significantly more poorly in the gender identification condition. This finding suggests that additional attentional demands during training may help to overcome the differences in the tasks.

Although the differences in performance were only observed in the short term, the equivalence of performance across the three groups at the retention session could simply be a factor of the familiarity with the materials. Subjects were tested on the same materials that they were exposed to during the first testing session rather than on novel materials from the same talkers. It could be the case that performance on a true generalization and retention test consisting of completely novel materials may distinguish group performance across the different training conditions. Due to the limitation of available stimulus materials, however, this could not be assessed by the present experiment.

Access to the Acoustic Information in the Signal

It is important to note that not all subjects were able to learn the talkers' voices over the five training blocks. Although the vast majority of subjects (84%) could learn to identify the talkers at a level greater than chance, several subjects could not. Although transcription scores at pre-test were comparable for both groups, the subjects who could not learn to identify the talkers by voice performed significantly worse on sentence transcription in the post-test and generalization blocks. Moreover, these differences could not be attributed to inattention or disinterest, since transcription errors were largely phonologically relevant and demographic variables such hearing insult or speech pathology problems did not reveal any abnormalities.

Additionally, previous research supports the proposal that the ability to learn to identify talkers by voice can predict transcription accuracy for speech samples produced by these talkers. In the original study demonstrating the transfer of talker identification training to word identification accuracy, Nygaard, Sommers, and Pisoni (1994) reported that not all of their subjects were able to learn to identify talkers by voice. Subjects who could successfully identify the talkers by voice showed higher recognition accuracy scores for words produced by familiar talkers as compared to novel talkers. The subjects who could not learn to identify the talkers by voice did not show such a difference. Taken together, the present finding suggest that it is not the mere exposure to a talker or a synthesis condition that is responsible for the gains observed after training, but rather the ability to access and utilize the acoustic information required to recognize the talkers by voice.

The findings of the present study also replicate those of Cleary and colleagues (2005) who examined talker discrimination in a group of pediatric cochlear implant users. Children listened to pairs of sentences and decided whether the two sentences were produced by the same or different talkers. Considerable variability was observed among the children with CIs, but those who were more proficient at talker discrimination also showed increased accuracy on a word identification task (Cleary et al., 2005). Taken together with the findings of Cleary and colleagues and Nygaard and colleagues, these data provide strong evidence for the interaction of lexical and indexical information, and suggest that the two streams may indeed be encoded and processed together.

There was no clear effect of talker familiarity on the recognition of speech processed by a CI simulation; subjects were as accurate at transcribing the speech produced by novel talkers as they were at transcribing speech produced by talkers used during training. The lack of a talker familiarity effect using CI simulated speech may not be completely anomalous, however. Barker (2006) showed a similar pattern of results for adult CI users trained to identify the voices of six talkers. In her study, CI users showed no differences in transcription accuracy performance for familiar versus unfamiliar talkers at a signal to noise ratio of +10 dB SNR. At 59% correct, talker ID accuracy scores for her fifteen CI users were nearly identical to our results with normal hearing subjects listening to 8-channel sinewave vocoders. Although she used a control group of normal hearing subjects, they performed the talker identification training with the unprocessed speech stimuli, so a direct comparison is inappropriate. Taken together, these data suggest that although indexical information regarding talker identity is preserved in electric hearing (as well as in acoustic simulations thereof), the talker familiarity effects that are observed for natural speech may differ in fundamental ways from those for cochlear implant simulations or individuals with CIs.

Behavioral and Clinical Implications

The findings from the present study suggest that there are multiple routes to the perceptual learning of speech. Although most studies utilize traditional methods of training that exclusively focus the listener's attention on the symbolic linguistic content encoded in the speech signal (e.g., Fu et al., 2005), other routes can yield similar outcomes and benefits. The crucial factor seems to be the amount of attention that is required of the subject, and the degree to which performance can be improved. Tasks that require significant amounts of controlled attention to the indexical properties of the signal can be just as effective as tasks that rely exclusively on attention to the linguistic content of the message. This finding has important implications for training and rehabilitation strategies for individuals who receive cochlear implants. The benefit observed in the current study for non-traditional training methods suggests that a variety of stimulus materials could be utilized to maximize outcome. Instruction on how to auditorily distinguish individual voices may provide the CI user with a more stable foundation for voice recognition that can be generalized to new talkers in new situations. Additionally, including a variety of stimulus materials and challenging perceptual tasks may promote interest in training, and protect against boredom and fatigue that can occur when only a single task is used.

Although the overall goal of cochlear implantation has been to restore receptive auditory capacity to the severely hearing-impaired individual, there are many other nonlinguistic aspects to hearing on which a CI user could experience benefit. Sound localization, the detection and identification of environmental stimuli and the enjoyment of music are all aspects of normal hearing that have not been well investigated in cochlear implant populations. Since all of these tasks require attention to acoustic information encoded in the signal that is nonlinguistic, greater variety in training tasks and materials may yield more robust results, many of which may transfer to speech perception and language processing tasks. If the goal of cochlear implantation is to provide the listener with access to the acoustic world, we should begin focusing training on achieving on such a goal. By limiting training to linguistic tasks, we may be undermining the robust adaptive abilities of CI users by depriving them of the full benefit that they may one day enjoy. Speech is not isolated from the rest of the acoustic world in which we live. A decision needs to be made as to whether the goal of cochlear implantation is only to provide access to the speech signal or to replace hearing, and directed measures need to be taken to achieve these goals accordingly.

References

Barker, B.A. (2006). An examination of the effect of talker familiarity on the sentence recognition skills of CI users. Unpublished Doctoral Dissertation, University of Iowa.

Bond, Z.S. & Moore, T.J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication, 14(4),* 325-337.

Bradlow, A.R. & Bent, T. (In Press). Perceptual adaptation to non-native speech. *Cognition.*

Bradlow, A.R., Torretta, G.M. & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20,* 255-272.

Burkholder, R.A. (2005). Perceptual learning of speech processed through an acoustic simulation of a cochlear implant. Unpublished Doctoral Dissertation, Indiana University, Bloomington.

Clark, G.M. (2002). Learning to understand speech with the cochlear implant. In Fahle, M and Poggio, T. (Eds), *Perceptual Learning*, Pp. 147-160. Cambridge: MIT Press.

Clarke, C.M. & Garrett, M.F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116(6),* 3647-3658.

Cleary, M., Pisoni, D.B. & Kirk, K.I. (2005). Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, and Hearing Research, 48,* 204-223.

Cleary, M. & Pisoni, D.B. (2002). Talker discrimination by prelingually deaf children with cochlear implants: Preliminary results. *Annals of Otoaryngology, Rhinoogy and Laryngology. 111(5-2, Supplement. 189),* 113-118.

Clopper, C.G., Carter, A.K., Dillon, C.M., Hernandez, L.R., Pisoni, D.B., Clarke, C.M., Harnsberger, J.D., and Herman, R. (2001), The Indiana Speech Project: An overview of the development of a multi-talker multi-dialect speech corpus. In *Research on Spoken Language Processing Progress Report No. 25* (pp. 367-380). Bloomington, IN: Speech Research Laboratory, Indiana University.

Clopper, C.G. (2004). Linguistic experience and the perceptual classification of dialect variation. Unpublished Doctoral Dissertation, Indiana University, Bloomington.

Cox, R.M., Alexander, G.C. & Gilmore, C. (1987). Development of the Connected Speech Test (CST). *Ear and Hearing, 8(5) Supplement,* 119s.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K. & McGettigan, C. (2005) Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134(2),* 222-241.

Dorman, M.F., Loizou, P.C. & Rainey, D. (1997). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding/ *Journal of the Acoustical Society of America, 102(1),* 2993-2996.

Dorman, M. & Loizou, P. (1998). The identification of consonants and vowels by cochlear implants patients using a 6-channel CIS processor and by normal hearing listeners using simulations of processors with two to nine channels. *Ear and Hearing, 19,* 162-166.

Dorman, M., Loizou, P., Fitzke, J. & Tu, Z. (1998). The recognition of sentences in noise by normal hearing listeners using simulations of cochlear implant signal processors with 6-20 channels. *Journal of the Acoustical Society of America, 104(6),* 3583-3585.

Dupoux, E. & Green, K.P. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23,* 914-927.

Eisner, F. & McQueen, J.M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics, 67(2),* 224-238.

Fairbanks, G. (1940). *Voice and Articulation Drillbook*. New York: Harper and Row.

Fahle, M. & Poggio, T. (Eds.). (2002). *Perceptual Learning.* Cambridge: MIT Press.

Fu, Q-J., Chinchilla, S. & Galvin, J.J. (2004). The role of spectral and temporal cues on voice gender discrimination by normal-hearing listeners and cochlear implant users. *Journal of the Association for Research in Otolaryngology, 5,* 253-260.

Fu, Q-J., Chinchilla, S., Nogaki, G. & Galvin, J.J. (2005). Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *Journal of the Acoustical Society of America, 118,* 1711-1718.

Fu, Q-J., Galvin, J.J., Wang, X. & Nogaki, G., (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoustic Research Letters Online, 6(3),* 106-111.

Goldstone, R.L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585-612.

Gonzales, J. & Oliver, J.C. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *Journal of the Acoustical Society of America, 118(1),* 461-470.

Greenspan, S.L., Nusbaum, H.C. & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14,* 421-433.

Hood, J.D. & Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, *19(5),* 434-55.

Kalikow, D.N., Stevens, K.N. & Elliot, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337-1351.

Krajlic, T. & Samuel, A.S. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin Review, 13(2),* 262-268.

Ladefoged, P. & Broadbent, D.E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29(1),* 98-104.

Loebach, J.L. & Pisoni, D.B (Under Review). Perceptual learning of spectrally degraded speech. *Journal of the Acoustical Society of America.*

Luo, X., and Fu, Q. J. (2005). Speaker normalization for Chinese vowel recognition in cochlear implants. *IEEE Transactions on Biomedical Engineering, 52,* 1358-1361.

McGarr, N.S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research, 26,* 451-458.

National Institutes of Health. (1995). Cochlear implants in adults and children. *NIH Consensus statement, 13,* 1-29.

Nygaard, L.C., Sommers, M.S. & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5(1),* 42-46.

Nygaard, L.C. & Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60(3),* 355-376.

Pallier, C., Sebastian-Gallés, N., Dupoux, E., Christophe, A. & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory and Cognition, 26(4),* 844-851.

Remez, R.E., Fellowes, J.M. & Rubin, P.E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 23(3),* 651-666.

Schwab, E.C., Nusbaum, H.C. & Pisoni, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27,* 395–408.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303-304.

Shannon, R.V., Fu, Q.-J. & Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngocia Supplementum, 552,* 1-5.

Shannon, R.V. (2005). Speech and music have different requirements for spectral resolution. *International Review of Neurobiology, 70,* 121-134.

Sheffert, S.M., Pisoni, D.B., Fellowes, J.M. & Remez, R.E. (2002). Learning to recognize talkers from natural, sinewave and reversed speech samples. *Journal of Experimental Psychology, 28(6),* 1447-1469.

Vongphoe, M. & Zeng, F.G. (2005). Speaker recognition with temporal cues in acoustic and electric hearing. *Journal of the Acoustical Society of America, 118,* 1055-1061.

Weil, S.A. (2001). Foreign accented speech: Adaptation and generalization. Unpublished Master's Thesis, Ohio State University.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Language Identification from Visual-Only Speech[1]

Rebecca E. Ronquest, Susannah V. Levi, and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We would also like to thank Luis Hernandez for assistance with programming, and Manuel Díaz-Campos, Althea Bauernschmidt, and Vidhi Sanghavi for their help in running subjects.

Language Identification from Visual-Only Speech

Abstract. The goal of the present investigation was to examine how observers identify English and Spanish from visual-only displays of speech. First, we replicated the recent findings of Soto-Faraco et al. (2007) with Spanish and English bilingual and monolingual observers using a different methodology. We found that prior linguistic experience affected response bias, but not sensitivity (Experiment 1). Additional experiments investigated the cues that observers used to carry out the language identification task. Participants were able to reliably identify languages when video clips were temporally-reversed, suggesting that prosody provides cues to language identity (Experiment 2). The contribution of lexical information to language identification was also investigated in Experiment 3. Participants' ability to identify stimulus direction (i.e., forwards vs. backwards) confirmed their sensitivity to differences in naturalness (Experiment 4). Taken together, the results of these four experiments indicate that prior linguistic experience, prosody, and perceived naturalness influence visual-only language identification

Introduction

A large body of research has demonstrated that speech perception is multimodal in nature. In addition to the auditory properties of speech, the visual signal carries important information about the phonetic structure of the message that affect the perception of the speech signal (c.f. Summerfield, 1987; Massaro, 1987). The visual aspects of speech have been shown to enhance or alter the perception of the auditory speech signal not only for listeners with hearing impairment, but for normal-hearing listeners as well (c.f., Campbell & Dodd, 1980; Summerfield, 1987; Lachs, 1999; Lachs, Weiss, & Pisoni, 2002; Kaiser, Kirk, Lachs, & Pisoni, 2003). In their seminal study of audio-visual speech perception, Sumby and Pollack (1954) demonstrated that the visual properties of speech carry important information about the linguistic content of the signal. They found that including the visual signal along with the auditory signal allowed listeners to better understand speech at less favorable signal-to-noise ratios. When the auditory signal became more degraded, the visual aspects of speech were more important, and increased the intelligibility of the speech signal.

The contribution of visual information to speech perception is also illustrated by the McGurk Effect, in which visual information alters the perception of the speech signal. McGurk and MacDonald (1976) found that when observers were presented with mismatched auditory and visual information, they perceived a sound that was not present in either sensory modality. For example, a visual velar stop /g/ paired with an auditory bilabial stop /b/ was perceived as /d/. Thus, the information carried by the visual signal not only enhances speech perception, as found by Sumby and Pollack (1954), but can override and alter the perception of auditory information, yielding a novel percept, as in the McGurk effect.

More recently, studies in the field of L2 acquisition have shown that the inclusion of visual information, along with the auditory signal aids in the acquisition of non-native contrasts. Hardison (2003) examined the acquisition of the English /l/-/ɹ/ contrast by native Japanese and Korean speakers. Participants were trained to identify these sounds under either auditory-only or auditory-visual presentation conditions. Learners who were trained in the auditory-visual condition showed better identification of /l/ and /ɹ/ in the post-test than those participants who were trained in auditory-only conditions. Hardison (2003) concluded that facial gestures enhance the discrimination of L2 targets in

difficult phonetic environments, and that visual cues to speech can be an additional source of information for L2 learners.

Similar studies have found that the contribution of visual information to speech perception, and the manner in which it is utilized, is also affected by an observer's native language and past experience with a second language. Hazan, Sennema, & Faulkner (2002) reported that visual information can facilitate L2 learners' perception of sounds that are contrastive in the L2, but do not contrast in the native language. For example, English contrasts the bilabial stop /b/ with the labiodental fricative /v/, whereas Spanish does not contain the latter phoneme. Hazan et al. (2002) found that Spanish learners of English who could perceive the contrast in the auditory-only condition also perceived the difference in the visual-only condition. In contrast, learners at early stages of acquisition who demonstrated higher rates of confusion between /b/ and /v/ auditorily did not benefit from the addition of the visual presentation. Hazan et al. (2002) concluded that learners at later stages of acquisition are sensitive to both the acoustic and visual cues associated with the non-native /b/-/v/ contrast, whereas less experienced learners do not gain any significant benefits from visual cues until the contrast has been acquired auditorily.

In a related study, Werker, Frost, and McGurk (1992) found that the percentage of "visual-capture" (i.e., when the visual signal overrides the auditory signal) responses in a McGurk-type task was affected by the participants' native language and L2 experience. L1 and L2 speakers of French and English were presented with an auditory-visual stimulus that consisted of conflicting auditory and visual information; auditory /ba/ was paired with visual /ba, va, ða, da, ʒa, and ga/. Werker and colleagues found that beginning and intermediate L2 learners of English demonstrated significantly less visual capture of the interdental place of articulation /ð/ than did more proficient speakers of English. The beginning and intermediate learners of English generally reported "hearing" /ta/ or /da/, thus assimilating the interdental place of articulation with the closest French phoneme (/t/ or /d/). In contrast, the native English speakers, bilinguals, and advanced English learners were more influenced by the visual stimulus, and demonstrated a higher percentage of /ða/ responses. Werker et al. (1992) concluded that the ability to lip-read in a language is highly dependent upon experience with that language.

The studies reviewed above indicate that the visual information carried in the speech signal contributes substantially to speech intelligibility and that linguistic experience affects the manner in which the visual information is processed. Although previous research on visual speech perception and speech-reading has focused primarily on examining participants' ability to identify specific segments or words in a particular language, whether languages can be discriminated or identified based on the information in the visual signal alone has not been directly examined until recently. Two recent studies by Soto-Faraco and colleagues (2007) and Weikum and colleagues (2007) investigated visual-only language discrimination in both adult and infant observers, respectively. Soto-Faraco et al. assessed the ability of monolingual and bilingual observers to discriminate Spanish and Catalan from visual-only displays of speech. Two groups of bilinguals (Spanish dominant, Catalan dominant) and three groups of monolinguals (Spanish, Italian, and English) took part in the task. Bilingual participants exhibited higher rates of discrimination than monolingual Spanish speakers. The English and Italian monolingual speakers were not successful at the task, suggesting that knowledge of at least one of the languages is necessary for visual-only discrimination. Soto-Faraco et al. concluded that prior experience with the specific languages is one of the primary factors contributing to successful discrimination. They suggested that a number of different aspects of the stimuli facilitated discrimination, such as the length of the utterance, and the number of distinctive segments or words present in the stimulus. A similar study with infants showed that 4 to 6 month olds can discriminate between French and English in visual-only displays, but that by 8 months, this ability is limited to bilingual infants (Weikum et al., 2007).

Soto-Faraco et al. suggested that future investigations should examine observers' ability to discriminate or identify languages that are less closely related than Spanish and Catalan. In the present study, we sought to corroborate Soto-Faraco et al.'s earlier findings with a pair of languages that differ in prosody using a different task. Spanish and English were chosen in this study because they differ in terms of prosody, or rhythmic structure (e.g., Pike, 1946; Grabe & Low, 2002); Spanish is considered a syllable-timed language, whereas English is considered a stress-timed language. Syllable-timed languages exhibit more even spacing of syllables in an utterance (Pike, 1946), measured by variability of vowel durations (Grabe & Low, 2002). Thus, the duration of vowels is more regular for syllable-timed languages. In contrast, successive vowel durations in stress-timed languages are more variable. For example, English exhibits extensive vowel reduction and shortened duration of unstressed vowels. In terms of visual correlates of speech, the vocalic gestures (i.e., vocal aperture) in Spanish are more regular, while the gestures in English are more varied. Thus, differences in the rhythmic properties of speech should be perceivable from visual information alone.

In Experiment 1, we replicated the initial findings reported by Soto-Faraco et al. with Spanish-English bilingual talkers and both monolingual and bilingual Spanish-English observers using a two-alternative forced-choice identification paradigm. Soto-Faraco et al. concluded that participants attend to a combination of lexical and segmental cues to discriminate languages in visual-only conditions, but they were unable to determine the exact properties that their participants relied on to discriminate the two languages used in their study. A second goal of the present investigation was to examine in more detail the specific types of cues that observers may use to identify a language from visual-only displays of speech. Experiments 2-4 manipulated several aspects of the visual signal to examine participants' reliance on prosodic cues and lexical information in visually-presented displays of speech.

The first experiment demonstrated that observers can reliably identify the language being spoken from a visual-only stimulus. Experiments 2A and 2B investigated the role of prosodic information in visual-only language identification. The third experiment examined whether participants used lexical information from visual-only displays of speech, by asking them to judge the lexicality of a stimulus. The fourth experiment assessed whether observers could reliably identify the direction (forwards or backwards) of video clips presented in both English and Spanish.

Experiment 1: Visual-only Language Identification

Methods

Stimulus Materials. The stimulus materials in Experiment 1 consisted of a series of visual-only video clips of 40 English and 40 Spanish sentences (see Appendix 1). One male and one female talker were recorded using Behringer B1 Studio Condenser microphone and a Panasonic AG-DVX100 video recorder. All recordings were made in a sound attenuated IAC booth in the Speech Research Laboratory at Indiana University. Both talkers were bilingual speakers of Spanish and English. The male talker was a native of Venezuela and the female talker was a native of Puerto Rico. Both talkers acquired English during early adolescence and had lived in the United States for at least 6 years at the time of recording.

Participants. Four groups of participants were recruited for Experiment 1: monolingual English speakers (N=16), monolingual Spanish speakers (N=12), English-dominant bilinguals (N=16), and Spanish-dominant bilinguals (N=12). The monolingual English observers were all undergraduate students at Indiana University who reported minimal knowledge of Spanish. The monolingual Spanish observers were all residents of Caracas, Venezuela, who reported that they did not speak or have knowledge of English. The Spanish-dominant bilinguals and English-dominant bilinguals were all

graduate students at Indiana University who reported that they were proficient speakers of both Spanish and English, and had some experience teaching college-level Spanish. Age of L2 acquisition for these bilinguals ranged from birth to 19 years of age. None of the participants reported a history of a speech or hearing disorder at the time of testing. All participants received $10 for taking part in the study.

Procedure. The stimuli were presented to the bilingual and monolingual English-speaking participants on an Apple Macintosh G4 computer. The monolingual Spanish speakers completed the experiment on an Apple Macintosh iBook G3 notebook computer in Caracas, Venezuela. PsyScript version 5.1 was used for stimulus presentation. Participants' responses were recorded with a button box for the language identification task. The entire experiment took approximately one hour to complete.

The visual-only language identification task consisted of two blocks of 40 video clips of short meaningful sentences in Spanish and English (see Appendix A). Each block consisted of 20 English sentences and 20 Spanish sentences spoken by both the male and female talkers. The stimuli were blocked by talker gender and counterbalanced across participants. After seeing each video clip, participants were asked to decide if the person in the video was speaking English or Spanish. No feedback was provided.

Data Analysis. In a two alternative forced-choice (2AFC) identification task, percent correct scores are influenced by both sensitivity and bias. For this reason, non-parametric measures of sensitivity (A') and bias (B") were calculated for each participant to obtain robust measures of performance (Grier, 1971). Both of these measures use the proportion of hits and false alarms to determine how sensitive the participants are to the differences in the signal and to quantify the extent to which they are biased toward one response alternative over another. In Experiments 1 and 2, a response of "English" to English stimuli was considered a "hit"; a response of "English" to Spanish stimuli was considered a "false alarm."

Sensitivity (A') is measured on a scale of 0.0-1.0, with 0 indicating no ability to discriminate differences in the signal and 1.0 indicating perfect discrimination. A value of 0.5 on the sensitivity scale indicates chance performance. Bias (B") is measured on a scale of -1.0 to 1.0. In Experiments 1 and 2, negative bias scores denote a tendency to respond "English" when presented with a stimulus, and positive values indicate a tendency to respond "Spanish." A score of zero indicates no response bias.

Results

To determine if participants' sensitivity was above chance performance (above 0.5 on the sensitivity scale) a one-sample t-test was conducted. As shown in Figure 1, the sensitivity measures for all four groups of subjects were significantly above chance (monolingual English $t(15) = 17.72$, $p < .001$; English-dominant bilinguals $t(15) = 28.30$, $p < .001$; monolingual Spanish $t(11) = 20.93$, $p < .001$; Spanish-dominant bilinguals $t(11) = 9.03$, $p < .001$). Thus, all participants were able to reliably identify the visual stimulus materials as English or Spanish. A one-way ANOVA was conducted on the A' scores with participant group as a between-subjects factor. The results of this analysis were not significant, demonstrating that all four groups performed comparably, and that observers' ability to identify a stimulus as Spanish or English did not depend on their native language or prior language experience.

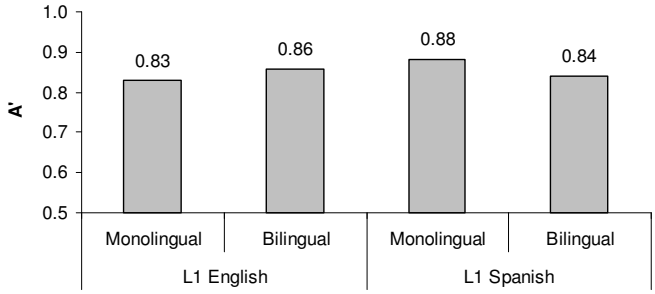


Figure 1. Mean sensitivity (A') for all four participant groups for Experiment 1.

The mean bias (B") scores for all four participant groups are shown in Figure 2. A one-sample t-test of B" scores showed that only the group of English-dominant bilinguals showed a response bias that differed significantly from 0.0 ($t(15) = -3.77$, $p = .002$); the English-dominant bilinguals had a strong tendency to choose the "English" response options, whereas the other three groups of participants did not demonstrate a significant bias. A one-way ANOVA was conducted on the B" scores in to analyze differences between response bias and participant group. The main effect of participant group was significant ($F(3,52) = 5.95$, $p = .001$). Post-hoc Tukey tests revealed that the English-dominant bilinguals had a response bias that was significantly different from the other three participant groups (English-dominant bilinguals compared to English monolinguals $p = .03$; Spanish-dominant bilinguals $p = .001$; Spanish monolinguals $p = .03$). While all participant groups showed a tendency to respond with their native language, the bias was strongest for the group of English-dominant bilinguals.

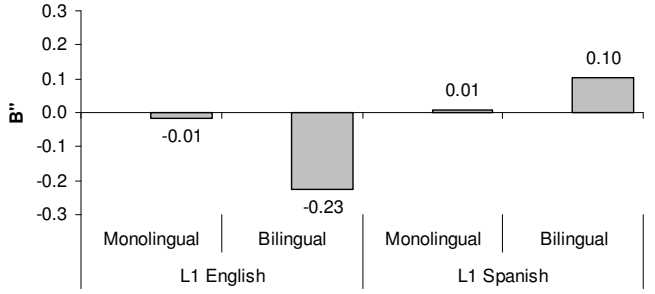


Figure 2. Mean bias (B") for all four participant groups in Experiment 1. Negative values indicate a bias to respond "English"; positive values indicate a bias to respond "Spanish"

Discussion

Regardless of language background or prior linguistic experience, all four groups of participants were able to complete the language identification task at levels that were significantly above chance. This result suggests that the visual speech signal alone provides sufficient information for an observer to correctly identify the language being spoken. That both monolingual and bilingual observers completed this task successfully replicates the earlier results of Soto-Faraco et al. (2007), who found that knowledge of only one of the test languages was sufficient to allow visual-only discrimination of Spanish and Catalan. The present results demonstrate that monolingual and bilingual participants not only can

discriminate between two different languages in visual-only displays of speech, but that they are able to accurately identify languages in a 2AFC task.

Unlike the results of Soto-Faraco et al. (2007), who found that bilingual observers were more successful in completing the discrimination task, we found no significant differences in sensitivity (A') between any of the four participant groups. Monolingual participants were just as sensitive as bilingual participants at identifying which language was spoken in the video clips. This result suggests that participants may have performed the 2AFC task by considering whether the stimulus was presented in their L1, or not in their L1, as opposed to making an English vs. Spanish judgment.

Measures of response bias (B") revealed that all four participant groups exhibited some preference to respond with their native language. The bias was particularly strong in the group of English-dominant bilinguals. The monolingual participants showed less response bias than the bilingual participants, although this difference failed to reach statistical significance. Familiarity and naturalness may underlie the patterns of bias observed in Experiment 1. Monolingual English and Spanish speakers who possess knowledge of only one of the two test languages may have responded based on whether they recognized a familiar word or temporal pattern in their L1, reflecting the naturalness of the stimulus. When no familiar words or patterns were present in the video, or when the stimulus looked unnatural, these participants may have indicated that the language was their non-native language. In the case of the bilinguals, all of the video clips had the potential to contain familiar words, segments, or syllable structures, and thus they all appeared to be natural. The bilingual participants, upon finding some degree of familiarity or naturalness in the signal, may have processed the visual signal as belonging to the L1 because of L1 dominance.

The English-dominant bilinguals, who exhibited a significant bias to respond "English", differed from the other three participant groups; the Spanish-dominant bilinguals failed to show a statistically significant native language bias, suggesting that they may have completed the task in an English mode, and adopted an English perceptual set. All paperwork and instructions were presented to the Spanish-dominant bilinguals in their non-native language (English), whereas the English monolinguals and English-dominant bilinguals received paperwork and task instructions in their native language. Using their non- native language as the primary mode of presentation may have attenuated the native language bias.

The results of Experiment 1 provide new insights into the robustness of the visual properties of speech. Several of the findings first reported in Soto-Faraco et al. were confirmed in the present study. They found that monolingual and bilingual observers could discriminate between Spanish and Catalan in visual-only displays of speech. Our results demonstrate that observers differing in language background and prior linguistic experience are able to identify languages based solely on the visual information. While Soto-Faraco et al. found that bilingual observers were better at completing a discrimination task, we found no significant differences in A' between monolingual and bilingual observers in our identification task. However, the effects of native language and prior linguistic experience were reflected in the differences in response bias (B") in the present study.

Although we replicated the basic findings reported by Soto-Faraco et al. (2007), neither their study, nor Experiment 1 explained *how* participants carried out the visual-only language identification task. What cues do observers use to identify the language spoken in visual-only speech? The remaining experiments described below examine the contribution of stimulus length, rhythmic properties, and lexical information to visual-only language identification. Unlike Experiment 1 which analyzed differences in language identification between monolingual and bilingual speakers of English and

Spanish, only monolingual English speakers took part in the remaining three experiments. Monolingual English speakers were chosen for two reasons. First, the results of Experiment 1, as well as those of Soto-Faraco et al., suggest that knowledge of one language is sufficient for visual-only language identification and discrimination tasks. Second, the monolingual English speakers in Experiment 1 did not perform differently than the bilingual participants, and showed less response bias.

Experiment 2: Rhythmic Cues to Language Identification

The results of Experiment 1 demonstrated that observers can identify language from visual-only displays of speech. Experiment 2 was designed to assess the contribution of stimulus length and prosodic differences to visual-only language identification. The high level of accuracy obtained in Experiment 1 may have been due in part to the nature of the stimulus set, which consisted of sentence-length utterances. Participants viewed sentences of varying lengths, ranging from 2 to 12 words in both languages. Soto-Faraco et al. (2007) found that language discrimination was better in longer phrases than in shorter phrases. We predict that the same would be true for a visual-only language identification task. Longer utterances provide larger samples of speech and more opportunity for the observer to extract information necessary for accurate language identification. For this reason, both sentences and isolated words were used in Experiment 2 to test whether longer utterances would facilitate language identification. We were also interested in determining whether the limited information from words would provide sufficient information to permit reliable language identification.

In addition to manipulating stimulus length, we also manipulated the direction of the video clips. Temporally-reversed ("backwards") versions of both the words and sentences were included in the stimulus set to assess whether participants can make accurate judgments about the language once lexical information has been eliminated. One possible way observers might extract language-specific information through visual speech is through rhythmic or prosodic information. Previous studies on visual-only speech perception have reported that observers are able to extract speaking-rate and stress differences from visual-only displays of speech (Green, 1987; Berstein, Eberhardt, Demorest, 1986). Thus, it is possible that observers in our experiments would be able to attend to rhythmic differences in the visual displays. As discussed earlier, Spanish is a syllable-timed language and English is a stress-timed language. Thus in Spanish, the vocalic gestures are more evenly-spaced in terms of duration while in English they are more variable. Temporal reversal of words and sentences preserves these global prosodic differences, but eliminates fine articulatory dynamics. That is, temporal reversal of the sentences and words creates stimuli which maintain overall temporal and rhythmic properties associated with Spanish and English, while at the same time eliminate the more fine-grained gestural-articulatory information necessary for lexical access. If participants use differences in the global rhythmic properties to identify language, we would expect that they should also be able to identify languages in the temporally-reversed stimuli, although they should be more accurate in the forwards condition where both lexical and rhythmic information are preserved. In contrast, if participants are unable to use prosodic cues, performance on the backwards stimuli should be extremely poor.

Experiment 2 examined both length and direction of visual-only stimuli. Manipulating the stimuli in this way allows us to investigate the potential contribution of rhythmic cues to visual-only language identification and to determine if single word utterances contain sufficient information for language identification. The experiment was divided into two parts. In Experiment 2A, participants were not informed that half of the video clips would be temporally-reversed. In Experiment 2B, the stimuli were blocked by direction, and participants were explicitly told that they would be viewing both forwards and backwards video clips.

Methods: Experiment 2A

Stimulus Materials. A total of 320 video clips were utilized in Experiment 2: 20 English and 20 Spanish sentences, and 20 English and 20 Spanish words, each spoken by two talkers, and presented in two directions (forwards and backwards). The 80 forwards sentences utilized in this experiment were the same as those used in Experiment 1 described above. The 80 word stimuli were recorded in the same way as the sentences described in Experiment 1. As with the sentences, each word was produced by the same male and female talker. The word stimuli included days of the week, animals, and the numerals one through ten (see Appendix 2). All video clips were temporally-reversed on an Apple Macintosh computer using Final Cut Pro, resulting in an additional 80 backwards sentences and 80 backwards words.

Participants. Thirty-four students enrolled in an introductory Psychology class at Indiana University participated in Experiment 2A. None of the participants who took part in Experiment 2A had completed Experiment 1. All were monolingual speakers of English who reported little or no knowledge of Spanish, and no history of a speech or hearing disorder at the time of testing. Participants received partial course credit for their participation.

Procedure. The general procedure for Experiment 2A was similar to the procedures used in Experiment 1. Participants were presented with two blocks of 160 stimuli. One block consisted of forwards and backwards words; the other block consisted of forwards and backwards sentences. The presentation of the blocks was counterbalanced across participants. After seeing each video clip, participants were asked to decide if the person in the video was speaking English or Spanish. A button box was used to record the participants' responses. The participants were not informed that half of the video clips in each block were time-reversed. No feedback was provided.

Results: Experiment 2A

As in Experiment 1, non-parametric measures of Sensitivity (A') and Bias (B") were calculated for each participant. Mean values of A' and B" are presented in Figures 3 and 4. A one-sample t-test of A' scores for the four conditions revealed that participants were sensitive to differences between the languages at levels statistically above chance (forwards sentences $t(33) = 16.84$, $p < .001$; backwards sentences $t(33) = 7.69$, $p < .001$; forwards words $t(33) = 7.96$, $p < .001$; backwards words $t(33) = 4.025$, $p < .001$). This finding indicates that participants were able to reliably identify the language from visual-only stimuli in all conditions. Moreover, the languages could be accurately identified when presented in the backwards condition. A repeated-measures ANOVA of A' scores with Stimulus Direction (forwards vs. backwards) and Length (words vs. sentences) as within-subjects variables revealed a significant main effect of Stimulus Direction ($F(1,33) = 4.42$, $p = 04$) and Length ($F(1,33) = 28.04$, $p < .001$). Participants identified the language as English or Spanish better when the stimuli were presented forwards (A'= 0.73) than backwards (A'= 0.64). The length of the stimuli also affected sensitivity. Participants were more accurate when presented with sentences (A'= 0.71) than with isolated words (A'= 0.66). The Direction by Length interaction approached significance ($F(1,33) = 3.81$, $p = .059$). Post-hoc analyses of this interaction revealed that participants were better able to identify the language being spoken in forwards sentences than in forwards words ($t(33) = -2.95$, $p = .006$). In the forwards conditions, observers' sensitivity was increased with increased length of the (words A'= 0.70, sentences A'= 0.78). In the backwards condition, however, longer utterances did not increase performance (words A'= 0.62, sentences A'= 0.66; $t(33) = -.942$, $p = .35$).

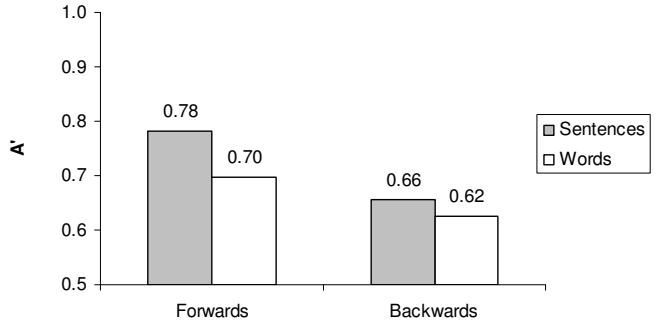


Figure 3. Mean sensitivity (A') in all four stimulus conditions for Experiment 2A.

Bias (B") scores for each of the participants were also calculated. A repeated-measures ANOVA of B" scores revealed a significant main effect of direction ($F(1,33)=22.03$, $p<.001$), indicating that participants were more biased to respond "English" for the forwards stimuli (B"= -0.09), and "Spanish" to the backwards stimuli (B"= 0.05). The main effect of Length was not significant. The Direction by Length interaction also reached significance (F (1,33) = 8.44, p = .006). Examination of this interaction revealed that participants displayed a greater bias to respond "English" when presented with forwards sentences than with forwards words (words B"= -0.03, sentences B"= -0.15; $t(33)$ = 2.29, p = .028). In the backwards condition, although the overall trend was a greater bias towards Spanish, the B" scores were not significantly different for words and sentences (words B"= 0.04, sentences B"= 0.07; $t(33)$ = -1.21, p = 0.23).

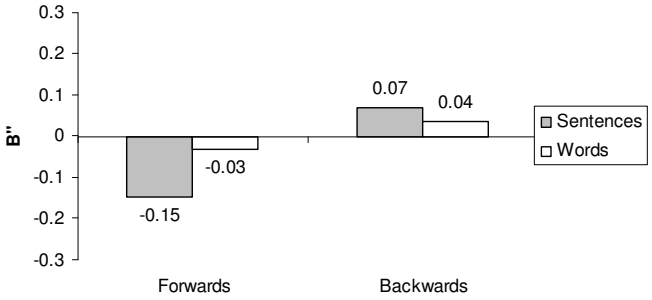


Figure 4. Mean bias (B") in all four stimulus conditions for experiment 2A. Negative values indicate a bias to respond "English"; positive values indicate a bias to respond "Spanish".

Methods: Experiment 2B

A modified version of Experiment 2A was conducted to examine the effects of direction when participants were explicitly told that some of the stimuli had been temporally-reversed. Because temporal reversal of the stimuli eliminated fine articulatory details and lexical cues, we hypothesized that awareness of the direction of the stimuli would force participants to rely on the prosodic information present in the video clips.

Stimulus Materials. The stimulus materials used in Experiment 2B were the same as those used in Experiment 2A.

Participants. A total of 33 introductory Psychology students took part in this experiment. A total of 13 participants were eliminated: three participants were eliminated because they had studied Spanish; one because of native Spanish speaking parents; one had undergone speech therapy; four due to computer malfunction; an additional four participants were eliminated so that the number of participants in each block order condition was equivalent. The remaining 20 participants were monolingual speakers of English who reported little or no knowledge of Spanish, and no history of a speech or hearing disorder. Participants received partial course credit for their participation. None of the participants had taken part in the previous experiments.

Procedure. Four blocks of visual-only stimuli (forwards words, forwards sentences, backwards words, and backwards sentences) were presented to participants. Prior to the presentation of each stimulus block, participants were told whether the stimuli would be presented forwards or backwards, and whether they would be viewing single words or whole sentences. Participants were divided into four groups based on the order of block presentation: 1) forwards words, forwards sentences, backwards words, backwards sentences, 2) forwards sentences, forwards words, backwards sentences, backwards words, 3) backwards words, backwards sentences, forwards words, forwards sentences, and 4) backwards sentences, backwards words, forwards sentences, forwards words. After viewing each video clip, participants were asked to decide if the person in the video was speaking English or Spanish. As in Experiment 2A, each block consisted of an equal number of English and Spanish tokens spoken by both the male and female talkers. No feedback was provided.

Results: Experiment 2B

The same statistical analyses carried out on the data from Experiment 2A were performed on the data collected in Experiment 2B. A summary of the A' scores is shown in Figure 5. A one-sample t-test of sensitivity (A') scores revealed that, as in Experiment 2A, participants could identify language at levels above chance (forwards sentences $t(19) = 7.75$, $p < .001$, backwards sentences $t(19) = 6.52$, $p < .001$; forwards words $t(19) = 5.73$, $p < .001$; backwards words $t(19) = 2.11$, $p = .04$). A repeated-measures ANOVA with Stimulus Direction (forwards vs. backwards) and Length (word vs. sentence) as within-subjects variables revealed a significant main effect of Direction ($F(1,19) = 10.95$, $p = .004$) and Length ($F(1,19) = 7.23$, $p = .01$). As observed in Experiment 2A, participants were more sensitive to language differences when the stimuli were presented forwards (A'= 0.68) than backwards (A'= 0.60), and were also more accurate with sentences (A'= 0.67) than words (A'= 0.61). The Direction by Length interaction was also significant ($F(1, 19) = 12.62$, $p = .002$). Post-hoc paired samples t-tests on this interaction revealed that accuracy was affected by length in the backwards condition (words A'= 0.56, sentences A'= 0.64; $t(19) = -3.65$, $p = .002$). In contrast to the results of Experiment 2A, no difference in length was found in the forwards direction (words A'= 0.67, sentences A'= 0.69; $t(19) = -1.33$, $p = .19$).

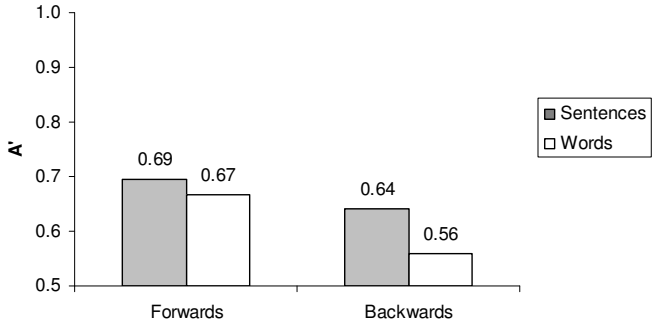


Figure 5. Mean sensitivity (A') in all four stimulus conditions for Experiment 2B.

Measures of response bias (B") were also calculated. A summary is presented in Figure 6. A repeated-measures ANOVA with Stimulus Direction (forwards vs. backwards) and Length (word vs. sentence) as within-subjects variables revealed a significant main effect of Direction ($F(1,19) = 7.66$; $p = .012$). Participants were more likely to respond "English" in the forwards condition than in the backwards condition. The general pattern of response bias is similar to that observed in Experiment 2A, but the magnitude of bias was attenuated. The main effect of stimulus Length and the Direction by Length interaction were not significant.

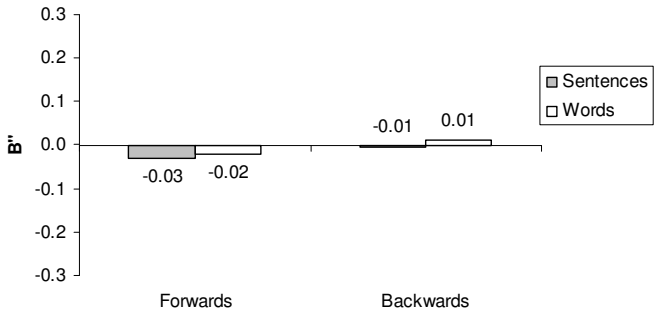


Figure 6. Mean bias (B") in all stimulus conditions for experiment 2B. Negative values indicate a bias to respond "English"; positive values indicate a bias to respond "Spanish".

Discussion: Experiments 2A and 2B

Experiments 2A and 2B were designed to examine the contribution of rhythmic information to visual-only language identification. As previously mentioned, global rhythmic differences between English and Spanish are retained in temporally-reversed stimuli. The results of these two experiments demonstrate that observers can identify differences in rhythmic structure from visual-only stimuli, and that they use this information in a language identification task. Participants' ability to reliably identify the language from the backwards stimuli suggests that even when access to lexical information is eliminated, a sufficient amount of prosodic information is still available to facilitate identification. Moreover, the present results demonstrate that monolingual speakers of English are able to make reliable judgments about language identity based on the visual information alone in the backwards stimuli. The results of this experiment suggest that observers perceive and utilize prosodic information associated with English and Spanish. We conclude that the rhythmic properties of a language are one cue that participants use to determine language identity from visual-only displays of speech.

Sensitivity to the language differences in the signal was greater when the stimuli were presented in the forwards condition as compared to the backwards condition. Greater sensitivity in the forwards condition was attained because forwards stimuli contain all possible cues to language identification; that is, forwards stimuli contain both rhythmic and lexical information, whereas only rhythmic cues are retained in backwards stimuli. The finding that sensitivity to language differences in the forwards condition was greater also suggests that participants use other sources of information in addition to rhythmic information to make their decisions about language identity. If rhythm and timing were the only properties observers attended to, then performance on the forwards and backwards stimuli would have been equivalent. In addition, stimulus length was also found to influence performance; sentence-length stimuli provided more information to language identity than isolated words. The sentence-length utterances contained more information than the words, and also provided participants with more time to make their decisions.

In addition to differences in sensitivity, response bias was also affected by the stimulus condition. Participants showed a greater tendency to respond "English" when they were presented with forwards stimuli and "Spanish" when presented with backwards stimuli. The differences in response bias suggest that in the backwards condition, when a word or a sentence appeared to be less natural and less familiar, or did not contain any recognizable information, participants were more likely to respond that the stimulus was Spanish.

Participants in Experiment 2A were not told that stimuli would be presented to them in two directions. When presented with stimuli, observers may have been making their decisions based on whether the stimulus display appeared natural or familiar. In the backwards condition, stimuli appeared less natural and less familiar, influencing the observers to identify these stimuli more often as Spanish; the forwards stimuli, because they were more familiar and natural, were more likely to be judged as English. In Experiment 2B, participants were explicitly told whether the stimuli were temporally-reversed. Thus, this group of participants was aware that they could no longer rely strictly on naturalness or familiarity to make their decisions, because half of the stimuli would appear unnatural; they also were aware that they would not be able to access lexical information in half of the stimuli. Participants' knowledge of stimulus direction altered their strategy in this task, and resulted in smaller response bias.

We also found that response bias towards English or Spanish was slightly greater with sentences than with words, although this difference was not statistically significant in all conditions. In the forwards condition, response bias to English was greater with the sentences than with words; observers were slightly more biased to respond "Spanish" when presented with a backwards sentence than with a backwards word. Participants may have exhibited stronger biases when presented with longer utterances because the additional length provided more cues to naturalness. Longer utterances also provided more information about gestures and articulation, which afforded participants more opportunity to decide if the stimulus looked natural or familiar. Sentence-length utterances offered more articulatory and timing information than word-length utterances.

The rhythmic properties of a language, which were maintained in the temporally-reversed versions of the stimuli, provided sufficient cues to language identity. Thus, it is not necessary for lexical information to be present for reliable language identification to occur. In the forwards condition, however, when both the rhythmic and lexical properties of the language were present, overall performance was enhanced. Greater sensitivity to the linguistic differences in the forwards stimuli suggests that a combination of rhythmic cues and lexical information is more beneficial than having only one available set of cues.

Experiment 3: Lexicality Judgments

Greater accuracy in the forwards condition in Experiments 2A and 2B suggests that participants attended to other properties of the stimulus, in addition to rhythm, when completing the language identification task. We hypothesize that observers extract both rhythmic cues and lexical information when making their decisions. Research on lip-reading has shown that both lexical and segmental information can be extracted from isolated words in the visual-only modality (Lachs et al., 2002; Kaiser et al. 2003). The purpose of Experiment 3 was to examine participants' ability to extract lexical information from visual-only isolated words, using a lexical decision task.

If participants accessed and used lexical information to carry out the language identification tasks in our earlier experiments, they should be more likely to report that forwards English stimuli are "words"

than Spanish stimuli. We also expected participants to be more likely to indicate that backwards video clips were "nonwords" than forwards video clips.

Methods

Stimulus Materials. The stimulus materials used in Experiment 3 consisted of the same forwards and backwards words utilized in Experiments 2A and 2B.

Participants. The participants in Experiment 3 were 32 introductory Psychology students at Indiana University. All participants met the same specifications described for Experiments 2A and 2B above. Partial course credit was awarded to all those who participated in this experiment. None of the participants had taken part in any of the previous experiments.

Procedure. Participants were presented with a single block of 160 trials mixed by talker, language, and stimulus direction. In contrast to the previous three language-identification tasks, participants were instructed to decide if the talker was saying a "word" or a "nonword." Participants were not informed that the words were spoken in English and Spanish, nor were they told that half of the video clips had been temporally-reversed. No feedback was provided.

Results

The number of "word" and "nonword" responses in each of the four conditions was calculated and these response frequencies were then analyzed using a Chi-square test of independence to determine if the distribution of responses was different across conditions. Collapsing over the direction of the stimuli, the distribution of "word" and "nonword" responses was significantly different for the English and Spanish stimuli (χ^2 (1, N = 5106) = 32.425, $p < .001$). This indicates that participants were more likely to categorize an English stimulus as a word than a Spanish stimulus (60% for English, and 52% for Spanish). The overall differences in frequency distribution reported for the total number of English and Spanish videos were also present when the stimuli were subdivided further. Chi-square analyses comparing forwards English and forwards Spanish words was significant (χ^2 (1,N = 2552) = 39.507, $p < .001$), indicating that there were more "word" responses to the forwards English stimuli (75%) than to the forwards Spanish stimuli (63%). Finally, the backwards Spanish stimuli were labeled as "words" less often than the backwards English stimuli (47% for English, and 42% for Spanish; χ^2(1, N = 2554) = 4.673, $p < .05$).

Collapsing over language, the distribution of "word" and "nonword" responses for the forwards and backwards stimuli was also significant (χ^2(1, N = 5106) = 319.36, $p < .001$). The participants categorized the forwards stimuli as "words" more often than the backwards stimuli (69 % for forwards video clips, and 44% for backwards videos). The overall pattern of responses found for direction was also observed within each language. Forwards English videos were labeled as words on 75% of the trials, whereas backwards English videos were labeled as words in only 46% of the trials. The chi-square analyses of this distribution was significant (χ^2(1, N = 2555) = 215.45, $p < .001$). The distributions of the forwards and backwards Spanish stimuli was also significantly different (χ^2(1, N = 2551) = 114.14, $p < .001$). Forwards Spanish stimuli were judged to be words more often than backwards Spanish stimuli (63% for forwards Spanish, and 41% for backwards Spanish.

In short, when observers were asked to make word/nonword judgments on isolated visual displays of English and Spanish words, they displayed a highly consistent pattern that differed statistically from chance expectation.

Discussion

The chi-square analyses indicated that observers' responses were not randomly distributed across the different stimulus conditions. "Word" and "nonword" responses varied systematically depending on the experimental conditions. Moreover, observers were more likely to judge an English stimulus as a word than a Spanish stimulus. The same preference for the "word" response was also observed with the forwards stimuli, regardless of the language of the stimulus. The forwards versions of both the English and Spanish stimuli were labeled as words more often than the backwards versions of the same stimuli.

The main goal of this experiment was to examine the extent to which participants access the lexicon when engaging in a visual-only language identification task. Although there is some evidence that lexical information may be accessed due to the higher frequency of "word" responses with the forwards English stimuli as opposed to the forwards Spanish stimuli, it is not possible to describe the extent to which lexical information contributes to visual-only language identification. Only monolingual speakers of English took part in this experiment, and it was thus assumed that these observers did not posses a Spanish lexicon. The fact that participants labeled approximately half of the Spanish stimuli as words suggests that they may have been making "word"/ "nonword" decisions based on whether the stimuli looked as if they could be possible words in English and not as a result of explicitly recognizing a stimulus as a specific lexical item. The forwards English stimuli were judged to be "words" the most frequently, followed by the Spanish words. In the backwards condition, the Spanish stimuli identified as "nonwords" more often than the English stimuli.

The pattern of responses observed in this experiment suggest that as in Experiment 2, the participants were attending to more global properties of the stimuli that are related to naturalness and familiarity, as opposed to making their decisions based on whether they recognized a specific word in their language. In the forwards English condition, the greatest number of cues to identity, both lexical and temporal information, is maintained in a coherent manner, and these stimuli should appear to the most natural-looking of all four stimulus types. The forwards Spanish stimuli are potentially recognizable as language, consisting of a combination of sounds and gestures that are also possible in English, but appear less recognizable than the English words. The backwards English and Spanish stimuli may maintain some of the rhythmic properties associated with each language, but lack the specific details necessary to identify a particular word.

Although the ability to recognize lexical information may contribute to more accurate language identification, the results of this experiment suggest that lexical properties of visual speech may not be as robust as the more global rhythmic and timing information. We conclude that observers may have been basing their decisions on whether the stimulus appeared as if it *could* be a word in English, or whether it looked highly unnatural and was therefore unlikely to be a possible word in English.

Experiment 4: Direction

The previous three experiments investigated participants' ability to identify language in visual-only stimuli and examined the extent to which they utilized prosodic and lexical information when making their decisions. In Experiment 3, participants may have made their word/nonword judgments based on the naturalness of the stimuli. That is, the forwards stimuli are considered natural, since they are actual language productions, whereas the backwards stimuli are unnatural. The goal of Experiment 4 was to investigate the question of articulatory naturalness by examining whether participants can reliably

identify the direction (forwards or backwards) of a silent video clip. We were also interested in determining if performance on this task would be affected by the language of the stimulus.

Methods

Stimulus Materials. The stimulus materials used in Experiment 4 consisted of the same set of video clips used in Experiments 2A and 2B: forwards and backwards visual-only video clips of English and Spanish words and sentences spoken by a male and female talker.

Participants. Twenty-five additional participants took part in this experiment. Three participants were eliminated due to computer malfunction, and two others for not following directions. Of the remaining 20 participants, 14 were introductory Psychology students who received course credit for taking part in this experiment. The other six participants were paid $10 for participating. None of the participants had completed any of the previous experiments described in this paper.

Procedure. Each participant was presented with one block of 160 words and one block of 160 sentences that were mixed by talker and language, but separated by stimulus length. All participants were presented with the words block first, followed by the sentences block. After viewing each video clip, participants were instructed to decide if the video they had just seen was forwards or backwards. The participants were not told that half of the video clips were in English and that half were in Spanish. No feedback was provided.

Data Analysis. As in Experiments 1 and 2, sensitivity (A') and bias (B") were the primary means of measuring performance on this task. In contrast to the previous experiments, however, participants were not asked to make language judgments, but instead were asked to identify direction. For this reason, in Experiment 4, a response of "forwards" to a forwards stimulus was considered a "hit". A false alarm occurred when a participant incorrectly identified a backwards stimulus as being forwards. Negative B" scores would thus indicate a tendency to respond "forwards," whereas positive scores would be indicative of a bias to respond "backwards."

Results

To examine observers' ability to identify the direction of each video clip, sensitivity (A') scores in the four stimulus conditions were calculated. A summary of these scores is presented in Figure 7. A one-sample t-test of A' scores for each condition was significant, indicating that participants were able to reliably discriminate between the forwards and backwards video clips (English sentences $t(20) = 6.23$, $p < .001$; English words $t(20) = 7.77$, $p < .001$; Spanish sentences $t(20) = 5.26$, $p < .001$; Spanish words $t(20) = 8.30$, $p < .001$). Thus, participants were able to reliably determine if the video clip they had just seen had been presented to them forwards or backwards. A repeated-measures ANOVA with Stimulus Language (English vs. Spanish) and Length (word vs. sentence) as within-subjects variables was conducted, and revealed a significant main effect of Length ($F(1,20) = 5.36$, $p = 0.03$). Observers were better able to identify a video clip as forwards or backwards when presented with an isolated word (A"= 0.73) than when presented with a sentence (A'= 0.68). Thus, in contrast to our earlier findings in Experiment 2, participants' ability to judge the direction of a stimulus was not enhanced when the video was longer in duration. The main effect of Stimulus Language and the Language by Length interaction were not significant. That the main effect of language was not significant indicates that participants were able to determine the direction of the video clip regardless of the language of presentation.

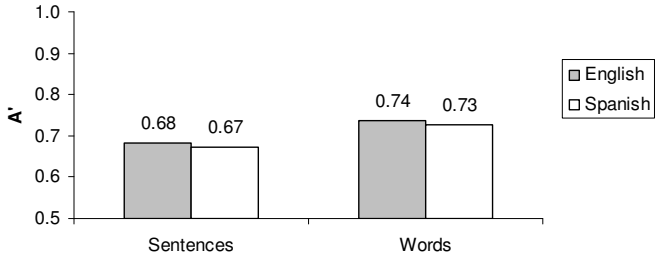


Figure 7. Mean sensitivity in four stimulus conditions in Experiment 4.

Mean bias (B") scores for each condition are presented in Figure 8. As shown here, all B" scores were negative, indicating a bias towards the "forwards" response alternative in all conditions. A repeated-measures ANOVA with Stimulus Language (English vs. Spanish) and Length (words vs. sentences) as within-subjects variables revealed a significant main effect of Stimulus Language ($F(1,20) = 8.36$, $p = .009$). This result indicates that participants had a greater tendency to respond "forwards" when presented with an English video (B"= -0.16) than with a Spanish video (B"= -0.08). The main effect of Length, and the Language by Length interaction did not reach significance.

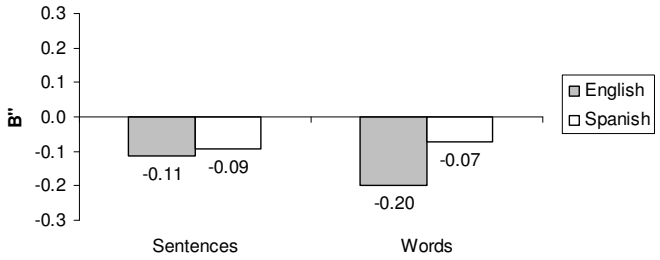


Figure 8. Response bias (B") in all four stimulus conditions in Experiment 4. Negative values indicate a bias to respond "forwards," whereas positive values indicate a tendency to respond "backwards."

Discussion

Overall, the results of Experiment 4 show that observers can successfully identify the direction (forwards or backwards) of a visual-only video clip. This finding suggests that participants are able to reliably identify differences in naturalness in the visual-only modality. Analysis of sensitivity (A') revealed no significant effects of presentation language. Participants were able to reliably identify the direction of a video clip regardless of the language in which it was spoken. We conclude that natural productions of speech (i.e. forwards utterances in both languages) were more natural-looking to observers because they specify gestural properties that are identifiable as being possible in language.

An examination of response bias (B"), however, did reveal an effect of language; participants were biased to respond "forwards" across all conditions, but this bias was strongest when the language of

presentation was English. Differences in naturalness and familiarity may be able to explain this trend. Because the participants were all monolingual English speakers, English sentences and English words would appear to be the most natural utterances to observers, and would also be the most familiar. Experiment 2 showed that the temporally-reversed versions of the stimuli maintained some of the global prosodic characteristics associated with English, and that this information could be reliably perceived from silent video clips. Thus, all English stimuli, including those that were presented backwards, contained some familiar properties of the native language which may have influenced observers to respond "forwards."

On the other hand, the Spanish stimuli would have appeared less natural and less familiar to participants. In the forwards condition, Spanish words and sentences may have looked as though they contained possible consonant and vowel gestures, but were less familiar in terms of their global prosodic characteristics. In the backwards condition, the Spanish words and sentences contained little, if any, familiar information to which the observers' could attend. Thus, participants were less likely to respond "forwards" when presented with a Spanish stimulus because of their lack of experience and familiarity with the rhythmic properties of the language. With the exception of the backwards Spanish stimuli, the video clips contained information that was in some way familiar to participants, influencing them to respond "forwards" more often than "backwards." The number of cues and the degree of familiarity and naturalness was greater for the English stimuli, reflecting why observers were more biased to identify these stimuli as being presented forwards than with the Spanish stimuli.

The second finding of this experiment is that participants exhibited greater ability to identify direction when the stimuli were word-length utterances than when they were sentences. This finding contrasts with the results obtained in experiments 2A and 2B, in which sensitivity to language differences was greater with longer utterances. It is possible that in this experiment, longer stimuli provided more opportunity for participants to believe that they had seen a familiar structure, resulting in greater confusion and lower accuracy with longer utterances.

General Discussion and Conclusions

In this paper we investigated how observers identify language from visual-only displays of speech. Our main goals were to replicate and extend the earlier findings of Soto-Faraco et al. (2007) using a different methodology. Overall, the results of Experiment 1 confirmed their earlier findings that language identification is possible from information in visual-only displays of speech. Although we found no differences in measures of sensitivity between monolingual and bilingual speakers of Spanish and English, the effect of prior linguistic experience was observed in measures of response bias; bilingual English speakers differed from all other participant groups, showing a bias for their native language.

A second goal of our investigation was to determine *how* observers identify languages when provided with visual-only information, by examining their reliance on prosodic information, lexicality, and naturalness. In Experiment 2, we directly examined the contribution of prosody to language identification by temporally reversing the video clips. We found that even when the visual stimuli were presented backwards, participants were still able to reliably identify stimuli as English or Spanish, although performance was significantly better in the forwards condition. We also found that observers were able to identify languages from short, isolated words, as well as sentences, but that sensitivity to language differences was greater in longer utterances. To examine if observers were accessing and using lexical information in the previous experiments, a lexical decision task was conducted in Experiment 3. The differential pattern of response frequencies suggested that observers may have been accessing some lexical information, but we concluded that "word/nonword" decisions were more likely influenced by the

perceived naturalness of the stimuli. Observers' attention to naturalness of the stimuli was investigated further in Experiment 4, in which participants were asked to decide if a video had been presented to them forwards or backwards. Participants were able to reliably identify isolated words and sentences as forwards or backwards, indicating that they were able to detect whether a stimulus looked like a natural language production. Although observers demonstrated a bias to respond "forwards" to all stimuli, the bias to respond "forwards" was stronger when the observers were presented with English stimuli than with Spanish stimuli. Based on the findings described above, we conclude that observers' prior linguistic experience influenced the way they performed the visual-only identification tasks, and that they were able to identify languages from visual-only stimuli using prosody and naturalness.

In their recent study, Soto-Faraco et al. (2007) found that linguistic experience affected observers' ability to discriminate languages when provided only with visual information. Bilingual speakers of Spanish and Catalan exhibited the highest discrimination scores, followed by the Spanish monolinguals. Monolingual speakers of English and Italian were unable to complete the discrimination task successfully, leading the authors to conclude that knowledge of at least one of the languages presented in the visual-only displays was necessary for reliable discrimination. Although observers in Experiment 1 did not exhibit sensitivity (A') differences, the effects of linguistic experience were revealed in differences in response bias. Analysis of response bias (B") revealed significant differences between the English-dominant bilinguals and the other three groups of observers. English dominant bilinguals exhibited a strong response bias toward their native language, whereas response bias for the other three groups did not differ significantly from each other. Although all four groups of observers showed some tendency to respond more with their native language, bias was stronger with the bilinguals than the monolinguals.

The bias exhibited by the English-dominant bilinguals can be attributed both to linguistic experience and methodological factors All stimulus materials contained sentences that were potentially recognizable to this group of bilinguals. Upon viewing a video clip, the English-dominant bilinguals were more likely to indicate that the stimulus was English based on their L1 dominance. The information presented in the video clips may have been processed through their L1, influencing participants to indicate that the stimulus was English more often than it was Spanish. The Spanish-dominant bilinguals also showed a bias to respond more with their native language, although this tendency did not reach significance. All paperwork and instructions were presented in English. Thus, the Spanish-dominant bilinguals did not show the same native-language effects because they were perceptually set in an English mode.

Effects of prior linguistic experience were also observed in the B" scores obtained in Experiments 2 and 4. In Experiment 2, monolingual English observers displayed a bias toward responding "English" when presented with forwards stimuli, and "Spanish" when presented with backwards stimuli. Because the monolingual English participants had more experience with English, the forwards stimuli were more familiar and natural to the observers. This familiarity influenced them to respond "English" more often when they viewed a forwards video clip. In the backwards condition, the stimuli appeared less familiar, resulting in a greater tendency to respond "Spanish." Thus, when observers were able to recognize a stimulus as a natural articulatory pattern, they showed a greater likelihood of indicating that the video clip was English.

The B" scores obtained in Experiment 4 suggested similar effects of prior linguistic experience. In this experiment, observers were presented with forwards and backwards English and Spanish words and sentences, and were asked to decide if the video clip had been presented "forwards" or "backwards." The general tendency observed here was to judge all stimuli as "forwards," but the bias to respond

"forwards," was greater for the English videos than for the Spanish videos, again revealing effects of prior linguistic experience. When presented with the English stimuli – regardless of direction – observers attended to both prosodic characteristics and naturalness. Because all of the English video clips maintained the basic temporal patterns of the observers' L1, participants had a tendency to indicate that the stimuli were forwards because in some respects, they all appeared to be possible and natural. As monolingual English-speaking participants have more experience with English utterances, English stimuli looked more natural, which may account for why more English video clips were categorized as "forwards" than Spanish video clips.

The effects of prior linguistic experience were also observed in the differential pattern of response frequencies obtained Experiment 3. The greater frequency of "word" responses to English stimuli than Spanish stimuli is clearly a consequence of participants' being monolingual English speakers. The increased number of "word" responses to all English stimuli – regardless of direction – is likely due to the prosodic cues preserved in both directions. Because the English stimuli contained some degree of naturalness or familiarity, they were categorized more often as "words" than the Spanish stimuli, which exhibited a different prosodic pattern.

We determined that prosodic information and naturalness of the stimuli were two sources of information that observers used when identifying the language spoken in a visual-only video clip. As previously mentioned, the contribution of prosodic information to visual-only language identification was examined in Experiment 2. Temporally-reversed video clips of words and sentences in English and Spanish were presented to observers, who were then asked to decide the language of the video clip. We found that observers were able to reliably identify the language even from backwards stimuli, suggesting that gross differences in prosody are sufficient to support language identification. Lexical information does not need to be present in order for observers to identify languages; prosodic cues alone provide sufficient information for language identification in this task. That sensitivity to language differences was greater in the forwards condition, however, indicates that the presence of additional information available in the forwards stimuli improves identification accuracy.

The objective of Experiment 3 was to determine the extent to which observers were able to access lexical information when provided with visual-only video clips of isolated English and Spanish words. Response frequencies in all stimulus conditions revealed a systematic pattern; "word" responses were more frequent for English stimuli versus Spanish stimuli, and also for forwards videos versus backwards videos. We hypothesized that monolingual English participants would judge English words as "words," and all other stimuli as "nonwords" based on observers' lack of experience with Spanish. However, many of the Spanish video clips were also judged as "words," suggesting that participants were not accessing specific lexical items, but instead may have been making their decisions based on the naturalness of the articulatory gestures and visual trajectories. In the forwards condition, both the English and Spanish video clips appeared natural because they contained temporal and gestural patterns that naturally occur in language. The backwards video clips, however, only maintain gross rhythmic information, and only the temporal patterns of English would have seemed familiar to this group of observers. Thus, backwards Spanish stimuli were judged as "nonwords" more often than backwards English stimuli because they lacked cues to naturalness and familiarity.

The results of Experiment 4 provided additional support for our hypothesis that differences in naturalness were detectable from visual-only displays of speech. Participants were able to reliably identify a stimulus as "forwards" or "backwards" regardless of the language of presentation. This result suggests that visual displays encode a number of highly salient properties (i.e. prosodic, articulatory, and perhaps lexical) that make them appear natural to observers.

Taken together, the results of the four experiments reported here demonstrate that visual displays of speech contain highly detailed information about the speech signal, and that observers' prior linguistic experience affects the way in which these sources of information are processed. We found that prosodic and lexical information, as well as cues to naturalness, are present in the visual signal. Observers are able to attend to and reliably use these sources of information in order to identify English and Spanish in silent video clips. Future investigations of visual-only language identification and discrimination will provide additional insights into how observers complete these tasks, and assess the extent to which lexical, segmental, and suprasegmental (prosodic) information is accessed during visual-only perception.

References

Abercrombie, D. (1965). *Studies in phonetics and linguistics*. London: Oxford University Press.

Bernstein, L.E., Eberhardt, S., & Demorest, M. (1986). Judgments of intonation and contrastive stress during lipreading. *Journal of the Acoustical Society of America*, *80*, S78.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, *32*, 85-99.

Davis, H., & Silverman, S. R (Eds.). (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart, & Winston.

Grabe, E., & Low, E.L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Papers in Laboratory Phonology 7* (pp. 515-546). Berlin: Mouton de Gruyter.

Green, K. (1987). The perception of speaking rate using visual information from a talker's face. *Perception & Psychophysics*, *42*(6), 587-593.

Grier, J. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Phsychological Bulletin*, *75*(6), 424-429.

Hardison, D. (2003). Acquisition of second language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24,* 495-522.

Hazan, V., Senemma, A., & Faulkner, A. (2002). Audiovisual perception in L2 learners. In H. L. Hansen and B. Pellom (Eds.), *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, September 16-20 (pp. 1685-1688).

Huarte, A., Molina, M., Manrique, M., Olleta, I., & García-Tapia, R. (1996). *Protocolo para la valoracions de la audicion y el lenguaje, en lengua española, en un programa de implantes cochleares*. Editorial Garsi, Grupo Masson: Madrid.

Kaiser, A.R., Kirk, K., Lachs, L., & Pisoni, D.B. (2003). Talker and lexical effects on audiovisual word recognition by audilts with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46,* 390-404.

Lachs, L. (1999). Use of partial stimulus information in spoken word recognition without auditory stimulation. In *Research on Spoken Language Processing No. 23* (pp. 81-118). Bloomington, IN: Speech Research Laboratory, Indiana University Bloomington.

Lachs, L., Weiss, J., & Pisoni, D.B. (2002). Use of partial stimulus information by cochlear implant patients and normal hearing listeners in identifying spoken words: Some preliminary analyses. *The Volta Review, 102*(4), 303-320.

Massaro, D. (1987). Speech perception by ear and eye. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). Hillsdale: Lawrence Erlbaum Associates.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Pike, K. (1946). *The intonation of American English*. (2nd Edition). Ann Arbor: University of Michigan.

Soto-Faraco, S., Navarra, J., Weikum, W.M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J.F. (in press, 2007). Discriminating languages by speech reading. *Perception and Psychophysics*, *69*, 218-237.

Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). Hillsdale: Lawrence Erlbaum Associates.

Weikum, W., Vouloumanos, A., Navarro, J., Soto-Faraco, S., Sebastian-Galles, N., & Werker, J.F. (2007). Visual language discrimination in infancy. *Science*, *316*(5828), 1159.

Werker, J., Frost, P., & McGurk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology*, *46*, 551-568.

Appendix A: List of sentences used in Experiments 1, 2A, 2B, and 4

English CID sentences list #9-10 (Davis & Silverman, 1970).

1. Where can I find a place to park?
2. I like those big red apples we always get in the fall.
3. You'll get fat eating candy.
4. The show's over.
5. Why don't they paint their walls some other color?
6. What's new?
7. What are you hiding under your coat?
8. How come I should always be the one to go first?
9. I'll take sugar and cream in my coffee.
10. Wait just a minute!
11. Breakfast is ready.
12. I don't know what's wrong with the car, but it won't start.
13. It sure takes a sharp knife to cut this meat.
14. I haven't read a newspaper since we bought a television set.
15. Weeds are spoiling the yard.
16. Call me a little later!
17. Do you have change for a five-dollar bill?
18. How are you?
19. I'd like some ice cream with my pie.
20. I don't think I'll have any dessert.

Spanish sentences, adaptation of CID list #9-10 (Huarte et al., 1996)

1. El desayuno está preparado en la mesa.
2. Qué le pasará al coche, que no funciona.
3. ¿Crees que el cuchillo cortará bien la carne?
4. No he leído un periódico desde que compré la televisión.
5. Las malas hierbas están estropeando el jardín de mi casa.
6. Llámame si puedes un poco más tarde, por favor.
7. ¿Tienes cambios de mil pesetas en la cartera?
8. ¿Qué tal estás?
9. Me gustaría tomar un poco de helado de chocolate con la tarta.
10. Creo que no tomaré ningún postre.
11. ¿Donde puedo encontrar un sitio para aparcar?
12. Me gustan las manzanas grandes y rojas que hay en los árboles.
13. Si comes muchos dulces, vas a engordar
14. La película ha terminado tarde.
15. ¿Por qué no pintas las paredes de otro color?
16. ¿Cuál es la noticia mas importante hoy?
17. ¿Qué escondes debajo del abrigo azul?
18. Espera un minuto en la puerta del cine.
19. Pondré azúcar y leche en mi café.
20. ¿Cómo puedo ser siempre el primero en llegar?

Appendix B: List of words used in Experiments 2A, 2B, 3, and 4

List of common English words

1. Monday
2. Wednesday
3. Friday
4. Saturday
5. Sunday
6. One
7. Three
8. Four
9. Five
10. Seven
11. Eight
12. Nine
13. Ten
14. Bird
15. Fish
16. Chicken
17. Duck
18. Dog
19. Donkey
20. Giraffe

List of common Spanish words

1. Lunes
2. Miércoles
3. Viernes
4. Sábado
5. Domingo
6. Uno
7. Tres
8. Cuatro
9. Cinco
10. Siete
11. Ocho
12. Nueve
13. Diez
14. Pájaro
15. Pez
16. Gallina
17. Pato
18. Perro
19. Burro
20. Jirafa

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Executive Function, Working Memory, Perceptual-Motor Skills, and Speech Perception in Normal-Hearing Children: Some Preliminary Findings[1]

Jennifer Karpicke, Christopher M. Conway, and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH grant DC00111. We wish to thank Dr. Larry E. Humes and Luis Hernandez for their help and support on this project.

Executive Function, Working Memory, Perceptual-Motor Skills, and Speech Perception in Normal-Hearing Children: Some Preliminary Findings

Abstract. Speech perception involves a number of complex cognitive processes. Previous work has suggested that executive function, working memory, and perceptual-motor skills play a role in children's language development. In order to further investigate this relationship, we evaluated the correlations between perception of degraded speech, which represents an approximation of the auditory signal received by a cochlear implant user, with several tasks involving executive function, working memory, and sensory-motor function. Our data revealed that age and performance on two tasks, one representing executive function and one representing perceptual-motor skills, were significantly correlated with children's perception of highly degraded speech. Moreover, correlations between each of these tasks and the perception of degraded speech remained strong and significant even when the effects of age were partialled out. These results suggest that processes attributed to executive function, such as attention, planning, motor control, hand-eye coordination, and problem solving, underlie spoken language processing and its development. The present findings with normal-hearing, typically-developing children provide an initial benchmark for more detailed investigation of individual differences in performance and audiologic outcome among profoundly deaf children who use cochlear implants.

Introduction

Profoundly deaf children who use cochlear implants (CIs) show a large degree of variability in terms of post-implantation audiologic outcomes. Children with very similar hearing losses and etiologies of deafness may obtain drastically different benefit from their CI (Pisoni et al., 2000). Variables such as age at onset of deafness, length of auditory deprivation, and age at implantation have been found to be associated with a wide range of outcome measures in implanted children (Fryauf-Bertschy et al., 1992; Osberger et al., 1991). Other language-related factors, such as mode of communication (Pisoni et al., 1997) and parents' knowledge of vocabulary (Stallings et al., 2000), also correlate with children's language outcomes after receiving a CI. Recently, several studies have shown that other neurocognitive measures, such as motor skills (Horn et al., 2006) and working memory span (Cleary et al., 2002b), also correlate with spoken language processing in children with implants. To better understand language outcome after cochlear implantation, it is important to further investigate these additional factors as possible predictors of language benefit. The current research investigates the relationship between spoken language processing, executive function, working memory, and perceptual-motor skills in normal-hearing children, with the motivation of applying our findings to the field of cochlear implantation. We hypothesize that performance on the executive function, working memory, and perceptual-motor tasks will be correlated with children's performance on a degraded speech perception task. Before describing the current project in more detail, we first review the findings associated with the role of executive function, working-memory, and perceptual-motor skills in speech perception and spoken language development.

Executive Function and Spoken Language Processing

Executive functioning is a term used to describe certain behaviors which are attributed to the functions of the frontal lobe (Stuss, 1992), such as attention, problem solving, planning, inhibiting

reflexive behaviors, monitoring behaviors, and goal-directed behavior (O'Reilly & Munakata, 2000). Russell (1948) investigated the role of the frontal lobe in development and found that the frontal regions were of great importance in the childhood years in terms of conditioning behavior patterns for the rest of the brain. Because of this, it is not surprising that executive dysfunction is linked to a variety of developmental disorders including attention deficit hyperactivity disorder (ADHD), Tourette's syndrome, dyslexia, and autism (Chelune et al. 1996; Ozonoff & Jensen, 1999; Pennington & Ozonoff, 1996; Helland & Asbjørnsen, 2000).

One of the most well-known executive function tasks is the Stroop Color Naming task (Stroop, 1935), in which the subject must inhibit a reflexive response to read printed color words, while naming the color of ink in which the words are printed. For example, the word 'blue' would be printed in red ink, and the subject must say 'red' while inhibiting the reflex to read the word 'blue'. Deficits in the ability to perform the Stroop task have been found in children with dyslexia (Helland & Asbjørnsen, 2000).

The Wisconsin Card Sorting Task (WCST) is another common measure of executive functioning. In this procedure, the subject is given a stack of test cards and must sort them based on the shape, color, or number of stimuli on the cards. The experimenter tells the subject if he is right or wrong so that the subject can learn the rules for matching the cards. After a certain number of correct matches, the rules change without notice. Performance on this test measures the subject's ability to flexibly shift responses and inhibit the reflex to follow the previous set of rules. Liss et al. (2001) investigated autistic children's performance on the WCST and found that children with autism were less likely to inhibit perseverative responses when the rules changed than were children with developmental language disorders. Similar results have been found when comparing perseveration in children with autism to typically-developing controls (Bennetto, Pennington, & Rogers, 1996; Prior & Hoffman, 1990; Ozonoff & McEvoy, 1994; Rumsey, 1985). In addition to these studies in children, individual differences in executive function have been found in adults. Performance on frontal lobe tasks is correlated with a variety of other information processing tasks even in typical populations (Miyake et al., 2000).

From these findings we can infer that executive functions may also play a role in the development of other cortical processes. In particular, children who show delays in the development in executive functions may also show delays in spoken language processing. Very little research has been conducted on the topic of executive function and language. Luria (1961) proposed a close interaction between language and executive function. Singer and Bashir (1999) described a case study of a 16-year-old boy with language-learning disorder, which involved problems with speech production, word finding, and language formulation. They found that their subject struggled with several domains of executive function including attention, inhibition, maintenance, adaptation and self-regulation. Similarly, patients with frontal lobe damage also show deficits in both written and verbal fluency (Kimberg et al., 1996). Because of the lack of research in the field of frontal lobe functions and language development, one of our goals was to investigate these factors in typically-developing children, ultimately applying our findings to children who use cochlear implants.

Working Memory and Spoken Language Processing

Working memory (for review, see: Conway, et al., 2005) has been identified as being involved in complex cognitive behaviors including reasoning and problem solving (Engle, 2002). The two most common tasks used to assess verbal working memory are digit span and nonword repetition. The digit span task is a common component of intelligence testing, which requires the subject to remember and repeat a sequence of digits either forwards (forward span) or backwards (backward span). Forward digit span (FDS) taps into the subject's ability to phonologically encode and verbally rehearse sequential

materials. Backward digit span (BDS), on the other hand, involves not only encoding and rehearsal, but also mentally manipulating the series of digits, which involves executive function and cognitive control.

Nonwords are novel, phonologically possible words that have no meaning or semantic representation in long-term memory. In a nonword repetition task, the subject is asked to repeat back spoken nonwords one at a time. Nonword repetition is a complex task which requires phonological encoding, memory, articulatory planning, and speech production. The relationship between speech perception, digit span, and nonword repetition tasks has been investigated in adults as well as in typically-developing children (for reviews, see Baddeley, 2003; Baddeley, Gathercole, & Papagno, 1998; Gathercole, 1999; 2006; Gupta & MacWhinney, 1997). Several studies have found that digit span and nonword repetition are correlated with children's vocabulary development (Adams & Gathercole, 1996; Edwards, Beckman, & Munson, 2004; Gathercole & Baddeley, 1989; Gathercole et al., 1999; Gathercole, Willis, Emslie, & Baddeley, 1992; Marjerus et al, 2006; Michas & Henry, 1994).

Digit span and nonword repetition have also been investigated in profoundly deaf children who use CIs. More importantly, several measures of working memory have been linked to language outcome measures (Pisoni & Geers, 2000, Burkholder & Pisoni, 2003; Burkholder & Pisoni, 2006), as well as other measures of working memory and spoken language processing (Cleary et al., 2002a; Dillon et al., 2004) in these children. However, the relationship between working memory and speech and language abilities still only accounts for about twenty percent of the variance in language outcomes of children with CIs. Therefore, it is necessary to investigate other cognitive factors that may affect language outcome in deaf children by assessing the correlations between spoken language processing, working memory, and other related tasks that draw on processes associated with frontal lobe function.

Perceptual-Motor Skills and Spoken Language Processing

In typically developing children, motor and language milestones tend to occur in synchrony (Lenneberg, 1967; Siegel, 1992), leading researchers to wonder if delays or deficits in one domain may also show up in the other. In fact, motor control and coordination have been found to be empirically linked with language abilities in both children and adults. For example, sequential fingertip tapping skill is correlated with phonological decoding (i.e., reading) abilities in normal adults (Carello, 2002). This finding implies that the processes which underlie difficulties in reading may be related to motor and coordination development. Children with specific language impairment (SLI) have been found to perform more poorly than age-matched controls on tasks involving motor control and visual discrimination (Powell & Bishop, 1992). Similarly, twin studies in which one or both twins have SLI, revealed a genetic link between language, motor, and working memory impairment (Bishop, 2000).

Following these earlier studies, investigators have recently begun to consider the role of motor development in language outcomes in children with CIs. Several longitudinal studies were completed, in which motor assessments made before the child received an implant were compared to the child's audiologic outcome measures post-implantation. These studies have found that children who present with higher motor scores on the Vineland Adaptive Behavior Scales (Sparrow et al., 2005) do better on assessments of language, vocabulary, and spoken word recognition than children with lower motor scores (Horn et al., 2005). Specifically, Horn et al. (2006) found that fine, but not gross, motor abilities were highly correlated with the expressive and receptive language abilities of children with cochlear implants. Similarly, the ability to correctly reproduce geometrical designs has been found to be predictive of implant success in children (Horn et al., 2004, in press; Fagan et al., in press). These studies suggest that relations between children's motor performance and their speech perception abilities warrant more detailed investigation.

Degraded Speech Perception as a Measure of Spoken Language Processing

The use of CI simulated speech as a measure of spoken language processing has become a common experimental tool over the last decade. This method allows researchers to use normal-hearing, typically-developing subjects and provide them with an auditory simulation of a cochlear implant in order to study perceptual learning and adaptation to spectrally-degraded speech (e.g., .Rosen et al., 1999; Fu & Galvin, 2003). Effects of context on understanding degraded speech stimuli have also been reported, showing that context plays an important role in sentence perception (Conway et al., 2007; Kalikow, et al., 1977; Miller & Selfridge, 1950; Rubenstein, 1973). Little research, however, has been conducted concerning the perception of degraded speech by children, especially for degradations simulating the effects of cochlear implants.

Eisenberg et al. (2002) studied the perception of degraded lexically-easy and lexically-hard sentences by normal-hearing children and profoundly deaf children who used cochlear implants. Lexically-easy sentences contained keywords that were high in word frequency (i.e., more common in the language) and low in neighborhood density (i.e. have few similar-sounding neighbors and therefore are less confusable). Lexically-hard sentences contained keywords that were low in word frequency (i.e., less common in the language) and high in neighborhood density (i.e. have more similar sounding neighbors and are therefore more confusable). Research has shown that word frequency and neighborhood density have an effect on speech perception under degraded conditions (Meyer & Pisoni, 1999; Luce & Pisoni, 1998). Eisenberg et al. (2002) found that normal-hearing children performed better on the perception of lexically-easy words and sentences under cochlear implant simulation than on lexically-hard words and sentences, demonstrating that word frequency and neighborhood density influence spoken language processing under degraded listening conditions. In addition, the children were more likely to correctly perceive degraded sentences than degraded isolated words, suggesting a benefit from the presence of contextual cues when listening to degraded speech.

Eisenberg et al. (2002) also presented these same lexically-balanced words and sentences in the clear to children who use CIs. Similar trends were observed. The children who used CIs performed better on lexically-easy sentences and words than on lexically-hard sentences and words. These findings replicated and extended earlier work by Kirk et al. (1995), who found that word frequency, neighborhood density, and context play a role in CI users' performance on speech perception tasks.

Except for the recent study by Eisenberg et al. (2002), there has been no research on perception of CI-simulated speech in children. It is important to continue the research in this field in order to investigate the degree to which children show individual differences in degraded speech perception, and to determine what other behavioral measures might be used to predict differences in spoken word recognition. Research on this problem has a direct clinical application to the field of cochlear implantation because the electrical signal received by CI users is highly degraded. In this study, we want to know if the same mechanisms that predict normal-hearing children's speech perception performance for CI-simulated speech also predict deaf children's audiologic and speech perception outcomes following implantation.

The present study was carried out to assess predictors of spoken language processing performance in children by investigating the relationship between spoken language processing and other cognitive processes. Specifically, we measured the speech perception of normal-hearing children listening to degraded (CI-simulated) sentences, and compared their performance on this task with measures of executive function, working memory, and perceptual-motor skills.

Method

Each of the participants in this study completed ten tasks in one session that lasted between sixty and ninety minutes, with breaks provided as needed. All testing was completed in a sound-proof booth. This paper summarizes performance on five of these tasks. For tasks involving published materials, the tests were administered as described in the testing manuals. All testing was completed at Indiana University by the first author. The specific materials and set-up for each task are described below.

Participants

Fifteen normal-hearing children were tested in the Speech Research Laboratory at Indiana University-Bloomington. The children were all monolingual English speakers who ranged in age from five years, five months (5;5) to eight years, eleven months (8;11) of age (mean 7;5). Children were recruited from the LEARN Home Schooling Network in Bloomington, IN as well as from the Department of Psychological and Brain Sciences KID Information Database.

Materials and Procedure

A brief pure-tone audiometric screening was administered to each child at 250, 500, 1000, 2000, and 4000 Hz using a portable audiometer (Maico Hearing Instruments, Model MA27A). Responses were required at 20 dB HL for each frequency. Each ear was tested separately, and all children passed the screening at all frequencies and in both ears.

Degraded Sentence Perception. Speech perception abilities were measured using lexically-controlled sentences which were degraded using a cochlear implant simulator. The sentences consisted of twenty lexically-easy (i.e., high word frequency, low neighborhood density) and twenty lexically-hard (i.e., low word frequency, high neighborhood density) sentences. Each sentence contained three keywords. Audio recordings of the sentences in the clear were obtained from Laurie Eisenberg, and are the same as in Eisenberg et al. (2002).

We degraded the sentences to four spectral channels using a sine wave vocoder cochlear implant simulator (www.tigerspeech.com), and presented them to each child through a loudspeaker (Advent AV570) at 65 dB SPL. The children were instructed to listen closely to each sentence and then repeat back what they heard, even if they were only able to perceive one word of the sentence. Two practice sentences were presented before testing. Children were given feedback after they made their responses to the practice sentences, but received no feedback during testing. All 40 of the test sentences (20 ‘easy’ and 20 ‘hard’) were presented to the children. The order of presentation of the test sentences was randomized for each child, and each sentence was presented only once. The child’s responses were recorded onto digital audio tape (DAT), and were later scored off-line based on number of keywords correctly repeated for each sentence.

WISC-III Forward and Backward Digit Span. The forward and backward digit span portions of the WISC-III intelligence scale (Wechsler, 1991) were administered to the children to obtain a measure of verbal immediate memory span. The testing materials were prerecorded by a young adult female talker, and were presented to the child through a loudspeaker (Advent AV570) at 65 dB SPL. Number sequences for both the forward and backward span tasks ranged from two to ten digits in length, with two strings of digits presented at each sequence length. For the forward span test, the child was asked to repeat the digits in the same order in which they were presented. The backward span task

required the children to repeat the digits in the opposite order. The children's responses were recorded on digital audio tape (DAT). Responses were scored on-line, using the tape as a scoring crosscheck. Testing was terminated when the child missed two sequences of the same length.

Memory for Dot Patterns. The memory for dot locations subtest of the Children's Memory Scale (CMS; Cohen, 1997) was used as a measure of both immediate and delayed recall of a spatial pattern. The children were shown a picture of six blue dots inside a large white background. The dot pattern was presented to the child for five seconds before being taken out of sight. The child was then asked to replicate the dot pattern by placing six blue chips onto a 3x4 grid. The child was allowed to place the chips on the grid in any order with no time restriction. The final pattern produced by the child was recorded and no feedback was given on the child's performance. The chips were then cleared from the child's grid, and the same dot pattern was shown to the child again for five seconds and then taken out of view. The child was then asked to replicate the dot pattern. This process was repeated a third time, resulting in a total of three learning trials in which the same dot pattern was used. Next, a trial of red dots was presented and the child was asked to replicate it. The red dot trial was not scored, but rather served as a distracter. The child was then asked to recall from memory the initial blue dot pattern that had been presented three times (immediate recall trial). At the conclusion of the experiment (after a delay of approximately 30 minutes), the child was again asked to replicate the blue dot pattern from memory (delayed recall trial). The child's replications were scored based on total number of chips placed correctly on the grid. Therefore, a child could receive a total raw score of up to twenty-four points for the three learning trials plus the immediate recall trial, and up to six points for the delayed recall portion of the task. These raw scores were then converted into scaled scores, taking into account the age of the child.

NEPSY – A Developmental Neuropsychological Assessment. The NEPSY (Korkman et al., 1998) is a clinical neuropsychological test battery designed to assess children's development in the following five domains: Attention/Executive Functions, Language, Visuo-spatial Processing, Sensorimotor, Memory and Language (for review, see Ahmad & Warriner, 2001). Results from two subtests of the NEPSY, described below, will be reported in this paper.

The Design Copying subtest of the NEPSY is part of the visuo-spatial processing domain, which includes a battery of tests aimed to assess the child's non-verbal visuo-spatial skills such as body movement and hand-eye coordination. The children were given eighteen geometric designs and asked to copy each design. All eighteen designs from the subtest were copied by each child. The child was not allowed to erase any mistakes made, and was not allowed to turn the paper while drawing. Each design was scored on a four-point scale taking into consideration things such as angle, completeness, and proportion. The scoring criteria differed for each design. The child's pencil grip and the presence or absence of hand tremors were also noted by the experimenter.

The Knock and Tap subtest of the NEPSY was administered to assess the child's attention and executive functioning, specifically, their ability to coordinate motor responses, inhibit reflexive responses, and shift responses when a rule change was introduced (i.e. inhibit perseveration). First, the experimenter demonstrated that when she knocked on the table the child was to tap on the table with his preferred hand (the child's non-preferred hand rested on the table at all times). She also demonstrated that if she tapped on the table, the child was to knock on the table. Four practice trials were carried out with this set of rules as many times as necessary until the child understood the rules. All children required only one practice session. A total of fifteen test trials were then completed under this set of rules. Then, a new set of rules was introduced to the child. Now, if the experimenter hit the side of her fist on the table, the child was to knock on the table, and if she knocked the child was to hit the side of

his/her fist on the table. However, if the experimenter tapped on the table, the child was to do nothing. Six practice trials were administered with the new set of rules as many times as necessary until the child understood the rules. All children required only one practice session. Fifteen test trials were then administered under the new rule set. The child's response was recorded for each trial. The number of correct responses out of a total possible raw score of thirty points was then converted into a percentile ranking based on the child's age.

Results and Discussion

Mean performance on the degraded sentence perception task is shown in Table 1, which reports the average number of target words correctly perceived by the children. Examination of this table reveals that children performed numerically better on the lexically-easy sentences (41% correct) than on the lexically-hard sentences (36% correct). However, this difference was not statistically significant: Easy vs. Hard, $t(14)=2.02$, $p=.06$.

Table 1: Degraded Speech Perception Scores

Task	Mean	SD
Easy Sentences	24.7	10.07
Hard Sentences	22.3	10.4
Total	47	19.9

Table 1. Means and standard deviations for number of keywords correct on the lexically-easy and lexically-hard sentences (total possible keywords correct = 60 for each sentence type), and for the total number of keywords correct across sentence type (total possible keywords correct = 120).

The children's scores on the other experimental tasks are summarized in Table 2. For the forward digit span (FDS) task, the children produced an average of 7.1 correct sequences (2 of length 2, 2 of length 3, 2 of length 4, and 1 of length 5). The children produced fewer correct sequences on backward digit span (BDS; mean of 4.1 correct sequences: 2 of length 2, and 2 of length 3). The difference between children's performance on FDS and BDS was statistically significant: FDS vs. BDS, $t(14)=7.25$, $p<.001$. The total digit span performance (TDS = FDS+BDS) had a mean of 11.4 correct sequences.

The CMS Learning scaled scores (total of the children's replications on the first 3 exposures) had a mean of 10.1 (50^{th} percentile). The CMS Total scaled scores (CMS Learning score + Immediate Recall score) had a mean of 11.0 (63^{rd} percentile). The CMS Delay scaled scores (Delayed Recall score) had a mean of 11.3 (63^{rd} percentile).

The NEPSY Design Copying Subtest scaled scores had a mean of 12.8. There are no percentile conversions available for this individual score, because it is part of a larger battery of tests which comprise the visuo-spatial processing domain of the NEPSY. Therefore, only standardized data for the sum of scaled scores in the entire domain were available. Since the Design Copying was the only visuo-spatial processing domain task of the NEPSY that we completed, we are unable to report percentile rankings for this scaled score. The NEPSY Knock and Tap raw scores had a mean of 28.5 and a standard deviation of 1.06. This translates to the 26^{th} to 75^{th} percentile, which is considered to be the expected level of performance for typically developing children.

Table 2: Means and Standard Deviations of Experimental Subtests

Task	Mean	SD
FDS	7.1	2.1
BDS	4.1	1.2
TDS	11.4	3.1
CMS-Learn	10.1	3.65
CMS-Total	11.0	3.85
CMS-Delay	11.3	2.38
Design Copy	12.8	2.73
Knock&Tap	8.5	1.06

Table 2. Means and standard deviations for various tasks including: forward digit span (FDS), backward digit span (BDS), total digit span (TDS = FDS+BDS), CMS-Learning scaled scores (CMS-Learn = replications of first 3 exposures), CMS-Total scaled scores (CMS-Total = CMS-Learn + Immediate Recall), CMS-Delay scaled scores (Delayed Recall), Design Copying scaled scores, and Knock and Tap raw scores (total possible = 30).

To investigate the relations between working memory and degraded speech perception, we ran a series of correlations comparing performance on the speech perception task with performance on the digit span and CMS tasks. The results of this analysis are summarized in Table 3. The only significant correlation that emerged from this analysis was the correlation between the total digit span score (TDS) and the scores on lexically-hard (Hard Sent., r=.53, p<.05) and total sentence perception (Total Sent., r=.52, p<.05). FDS was also correlated with scores on lexically-hard sentences (r=.50, p=.06) and total sentence perception (r=.46, p=.09) although both correlations were only marginally significant.

However, when we performed a partial-correlation analysis controlling for age, both correlations became non-significant (Table 4), indicating that younger children have more problems with both digit span and speech perception than older children. The fact that digit span correlates with speech perception only when age is not controlled for suggests that any association between performance on these two tasks is largely accounted for by a single common source of variance having to do with chronological age.

Table 3: Correlational Analysis Between Working Memory and Speech Perception Tasks

Task	Easy Sent.	Hard Sent.	Total Sent.
FDS	r = .387 p = .154	r = .498 p = .059	r = .456 p = .088
BDS	r = .427 p = .112	r = .315 p = .253	r = .380 p = .126
TDS	r = .508 p = .053	r = .526 p = .044*	r = .523 p = .041*
CMS-Learn	r = .109 p = .698	r = .110 p = .696	r = .113 p = .689
CMS-Total	r = .221 p = .429	r = .207 p = .460	r = .219 p = .432
CMS-Delay	r = -.079 p = .778	r = -.053 p = .852	r = .-068 p = .810

* = sig. to .05

Table 3. Correlational analysis of digit span and CMS: dot location memory tasks with degraded speech perception scores. Age is included as a variable.

Table 4: Partial-Correlation Analysis Between Digit Span and Speech Perception (Controlling for Age)

Task	Easy Sent.	Hard Sent.	Total Sent.
FDS	r = .040 p = .893	r = .254 p = .380	r = .162 p = .580
BDS	r = .140 p = .633	r = -.007 p = .980	r = .066 p = .823
TDS	r = .155 p = .597	r = .224 p = .441	r = .203 p = .487

Table 4. Partial-correlation analysis of digit span and speech perception scores, controlling for age.

To assess the relationship between speech perception, executive function, and perceptual-motor skills, we performed a correlational analysis comparing the two NEPSY subtests with scores from the speech perception task. The results of this analysis are summarized in Table 5. Both the Design Copying and the Knock and Tap scores were significantly correlated with the speech perception measures (r>.59, p<.02).

Table 5: Correlations Between EF and P-M skills and Speech Perception

Task	Easy Sent.	Hard Sent.	Total Sent.
Design Copy	r = .663 p = .007**	r = .653 p = .008**	r = .676 p = .006**
Knock & Tap	r = .593 p = .020*	r = .386 p = .156	r = .501 p = .057

* = sig. to .05, ** = sig. to .01

Table 5. Correlational analysis of executive function and perceptual-motor tasks (NEPSY: Design Copying scaled scores and NEPSY: Knock & Tap percentile rankings) with speech perception. Age included as a variable.

Strong positive and significant correlations (r's >.6, p's<.01) between scores on the Design Copying task and all of the degraded sentence perception measures were observed. Children who were better able to perceive speech under degraded conditions also performed better at copying geometric designs. A scatterplot of the individual scores is shown in Figure 1. These results suggest that executive function and perceptual-motor skills which are involved in copying geometric designs are associated with speech perception, word recognition, and spoken language processing.

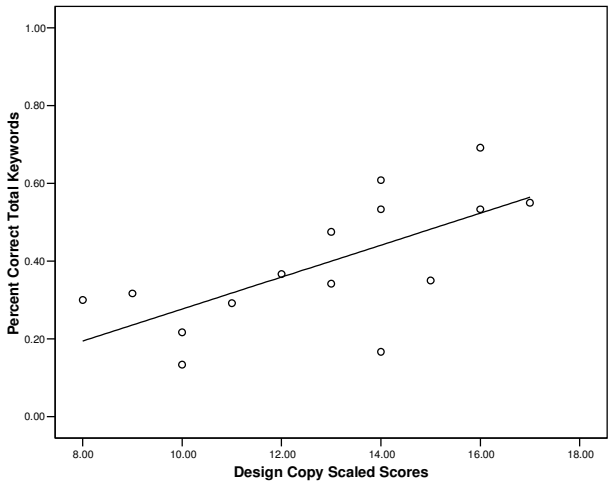


Figure 1. Scatterplot of performance on the Design Copying task and percent correct total keywords (easy + hard sentences) on the sentence perception task.

Scores on the Knock and Tap test were also positively correlated (r=.59, p<.05) with speech perception, but only for the perception of lexically-easy sentences. Children who were better at perceiving lexically-easy sentences under degraded conditions also performed better on the Knock and Tap task. A scatterplot of the scores on these two tasks is shown in Figure 2. This indicates that the perception of high-frequency, low-density words is linked to the executive function and perceptual-motor skills involved in completing the Knock and Tap task.

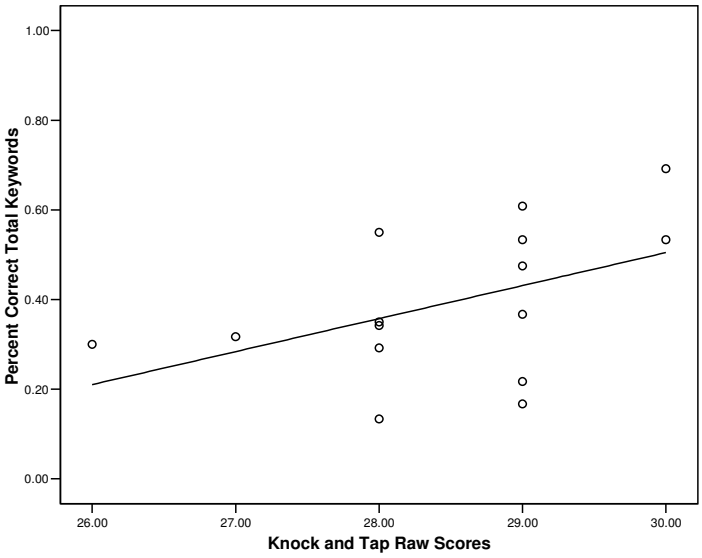


Figure 2. Scatterplot of performance on the Knock and Tap task and percent correct total keywords (easy + hard sentences) on the sentence perception task.

In order to evaluate the effect of age on these correlations, we performed a partial-correlation analysis controlling for age. The results of this analysis are summarized in Table 6. When age is partialled out of the correlation, the Design Copying test scores are still significantly correlated (r's>.59, p's<.05) with all three speech perception measures. This indicates that the link between the abilities used on Design Copying and those used in degraded speech perception is not a result of age. In other words, the association between these two tasks cannot be explained simply due to younger children performing more poorly than older children.

Table 6:
Partial-Correlation Analysis Between EF and P-M skills and Speech Perception (Controlling for Age)

Task	Easy Sent.	Hard Sent.	Total Sent.
Design Copy	r = .631 p = .015*	r = .597 p = .024*	r = .648 p = .012*
Knock & Tap	r = .677 p = .008**	r = .349 p = .221	r = .533 p = .050*

* = sig. to .05
** = sig. to .01

Table 6. Partial-correlation analysis of executive function and perceptual motor tasks (NEPSY: Design Copying scaled scores and NEPSY: Knock & Tap percentile rankings) with speech perception, controlling for age.

The correlation between the Knock and Tap test and the perception of lexically-easy sentences actually became stronger when age was controlled for (r=.68, p<.01). The correlation between Knock and Tap scores and the total keywords perceived was also significant (r=.53, p=.05). This pattern suggests that the association between the Knock and Tap task and the degraded speech perception task also are not due to age-related factors.

We also found that both the Knock and Tap and Design Copying scores were strongly correlated with one another (r=.64, p=.01) even when age was partialled out of the correlation (r=.62, p<.05). This result suggests that both tasks share a common source of variance that is independent of age, and some overlap exists between the resources need to complete these tasks, even though they were placed in separate domains in the NEPSY test battery.

Overall, our results indicate varying degrees of correlation between working memory, executive function, perceptual-motor skills, and speech perception abilities in normal-hearing children. While the correlations between speech perception and measures of working memory (Digit Span and CMS: Memory for Dot Locations) reflect age related factors at this stage in our analysis, the executive function and perceptual-motor tasks were found to correlate with speech perception regardless of age. Therefore, it is likely that some aspects of motor and frontal lobe functioning may play a role in spoken language processing and perception.

General Discussion

We originally predicted that measures of executive function, working memory, and perceptual-motor skills would be correlated with children's performance on a degraded speech perception test, based on earlier empirical evidence that spoken language processing is associated with attention, memory, and motor functions. The present findings provide insight into the relationship between speech perception, executive function, and perceptual-motor skills in typically-developing children. These findings are particularly interesting because, not only are executive function and perceptual-motor tasks correlated

with a measure of spoken language processing, they are also correlated with each other. This indicates that the development of attention, coordination, and speech perception and production develop at similar rates in children, and that these abilities may reflect common organizational processes.

It is important to consider the following observations regarding our results. First, the correlation we initially found between digit span and degraded speech perception was expected, based on earlier studies showing that spoken language processing and verbal working memory measures are linked in typically-developing children and in children who use CIs. However, this correlation became weaker when age was controlled for in the analysis. Whereas previous studies have shown a link between working memory and vocabulary development (e.g., Baddeley, 2003; Baddeley et al., 1998), our data suggest that verbal working memory may not be related to the perception of spoken language under degraded listening conditions, at least for the current set of stimuli.

Second, we found no significant correlations between memory for dot locations and children's speech perception abilities. Previous research has shown that cochlear implant children perform more poorly on measures of spatial memory span than normal-hearing children (Cleary, et al., 2001). However, in a study using the same CMS dot locations test with children who use cochlear implants, their overall performance was comparable to the published norms for the task, but was still slightly lower than scores obtained from a normal-hearing control group (Cleary & Pisoni, 2007). It is possible that visuo-spatial memory is not closely linked to speech perception when compared to phonological working memory. Further examination of these tasks in typically-developing children is warranted.

Third, the Design Copying and Knock and Tap tasks were found to strongly correlate with the degraded speech perception task. Previous studies have shown that non-verbal cognitive development is highly predictive of language development in children as young as 2 to 4 years of age (Oliver, Dale, & Plomin, 2004), and that children who show deficits in language development also show deficits in non-verbal domains (Viding et al., 2003). Korkman and colleagues (2001) found that the subtests of the NEPSY are highly correlated with age, especially in younger children (ages 5 to 8). Their findings not only indicate that the NEPSY is a developmentally sensitive test, but also magnify the importance of our findings, because the correlations we found between these tasks were not a function of age. The development of executive function, speech perception, and perceptual-motor abilities apparently varies from child to child, with children's performance on one task being highly predictive of their performance on the others.

The finding that frontal lobe functions are related to language development is not surprising when framed in the theory of embodied cognition (for review, see Wilson, 2002). This approach suggests that cognitive and sensory processes do not function independently of one another, but rather are controlled by a complex, integrative system which encompasses brain, body, and world (Clark, 1997). Developmental research has shown that milestones in both language and motor development follow a similar timetable, and that motor development successfully predicts later language development in children (Lenneberg, 1967; Siegel, 1982). In line with this view, it has been found that there are distinct developmental periods for frontal lobe functions during which children's attentional and self-regulational abilities are developed and organized (Case, 1992). In fact, some researchers believe that the frontal lobe is directly responsible for guiding the actions of other perceptual, cognitive, and physical systems, such as language (Stuss 1992; Thatcher, 1992).

Fourth, we found a strong relationship between the Design Copying and the Knock and Tap tasks. Although the NEPSY test battery places Design Copying and Knock and Tap in separate domains (visuo-spatial processing, and attention/executive function, respectively), children may actually be using

similar neural and cognitive resources to complete these tasks. For example, replicating a drawing of a geometric design involves, at a minimum: visuo-spatial processing, attention, planning, and precise fine-motor coordination. Likewise, performing the Knock and Tap task involves, at a minimum: attention, inhibition, motor coordination, and restraining perseveration. It is apparent that some overlap occurs in the cognitive functions required to complete these two tasks, all of which have been attributed to the functioning of the frontal lobe (i.e. executive function).

Finally, we found that Design Copying was correlated with both lexically-easy and lexically-hard sentence perception (and overall performance on the speech perception task), while Knock and Tap was correlated with lexically-easy sentence perception but not with lexically-hard sentence perception. This pattern is particularly interesting because no significant difference was found between children's performance on the two types of sentences. It has been reported that there is a difference between the perception of lexically-easy and lexically-hard words under adverse listening conditions. Because lexically-hard words are used less frequently and are more easily confusable than lexically-easy words, lexically-easy words are generally perceived better than lexically-hard words under degraded conditions. Therefore, the perception of lexically-hard words is a more cognitively challenging task, which requires the listener to encode and discriminate fine phonetic distinctions in the speech signal (especially when the signal is degraded) in order to perceive the words correctly. The reason why Design Copying correlates with lexically-hard sentence performance while Knock and Tap correlates with lexically-easy sentence performance is unclear at this point. The two tasks correlate with each other, indicating that they may overlap in some executive function domain. It is possible that there are different dimensions to the various executive functions such as attention. For example, Knock and Tap and lexically-easy sentence perception may involve a form of general attention, while Design Copying and lexically-hard sentence perception may involve attention for fine details.

To summarize, we found no relationship between degraded speech perception and verbal or spatial working memory tasks in typically-developing children. However, we did find strong positive relations between speech perception and both a test of visuo-spatial processing (Design Copying) and a test of attention/executive function (Knock and Tap). The Design Copying and Knock and Tap tasks correlated with one another, indicating that while they are placed in different NEPSY domains, there may be some overlap in the cognitive functions (such as attention, planning, and motor coordination) required to complete these tasks. Upon further investigation of this relationship, we found that these two tasks were correlated with different speech perception measures in terms of the lexical content of the sentences. This suggests that executive function is not a homogenous psychological construct and may reflect different subskills and processing domains.

Finally, the sample size of the present study (N=15) is small and these results require confirmation with a larger sample of children. In addition, caution should be exercised when generalizing the results obtained for the degraded speech used in this study to other forms of spoken language perception tasks. In this study, we were interested in a fairly severe form of degradation that closely mimicked CI speech. The results for this form of degradation, however, may not generalize to other forms of degradation examined previously such as background noise, reverberation, or filtering. One could argue that, the more severe the stimulus degradation, the greater the role to be played by higher level cognitive processing used for deciphering the distorted input. The use of CI simulated speech as a performance measure has been criticized because subjects are acutely exposed to this type of degradation. The argument has been made that measures obtained from normal-hearing subjects under CI simulation may not be directly comparable to the measures obtained from CI users who are chronically exposed to acoustic degradation and therefore experience an effect of learning with continued exposure. In addition, the children assessed in this study already had typically developing spoken language abilities.

Their ability to make use of context or to make sense of degraded input may differ from that in children without typically developing language systems, such as many profoundly impaired children who receive cochlear implants.

Conclusions

Spoken language processing is a complex task that involves the processing and encoding of fine acoustic details. Motor and memory abilities have been found to be linked to children's ability to perceive language under highly degraded conditions. We found that executive function and perceptual-motor tasks strongly correlated with typically-developing children's ability to perceive degraded speech signals. These findings indicate that the development of certain aspects of executive function, such as attention, planning, motor control and coordination, visuo-spatial processing, and inhibition are closely linked with the development of spoken language processing in children. All of these executive functions are attributed to the frontal regions of the brain, indicating an important role of frontal lobe development and coordination in language development. Aside from their general theoretical impact in terms of the role of cognitive control in language processing, the present findings have implications for the study of individual differences in deaf children who have received CIs. Research on this unique clinical population in our lab is focused on discovering factors that may help predict profoundly deaf children's outcome and benefit achieved after receiving an implant. Understanding the contribution of such factors will allow for advancements in clinical protocols, ultimately improving the techniques used for aural habilitation and rehabilitation of children and adults who receive cochlear implants as a treatment for profound deafness.

References

Adams, A.-M., Gathercole, S.E. (1996). Phonological working memory and spoken language development in young children. *The Quarterly Journal of Experimental Psychology, 49A*, 216-233.

Ahmad, S.A., Warriner, E.M. (2001). Review of the NEPSY: A developmental neuropsychological assessment. *The Clinical Neuropsychologist, 15(2)*, 240-249.

Baddeley, A.D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189-208.

Baddeley, A., Gathercole, S., Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158-173.

Bennetto, L., Pennington, B.F., Rogers, S.J. (1996). Intact and impaired memory functions in autism. *Child Development, 67,* 1816-1835.

Bishop, D.V. (2000). Motor immaturity and specific speech and language impairment: Evidence for a common genetic basis. *American Journal of Medical Genetics, 114(1)*, 56-63.

Burkholder, R.A., Pisoni, D.B. (2003). Speech timing and working memory in profoundly deaf children after cochlear implantation. *Journal of Experimental Child Psychology, 85*, 63-88.

Burkholder, R.A., Pisoni, D.B. (2006). Working memory capacity, verbal rehearsal speed, and scanning in deaf children with cochlear implants. *In Advances in the spoken language development of deaf and hard-of-hearing children*. Oxford, UK: Oxford University Press, 328 357.

Carello, C., LeVasseur, V.M., Schmidt R.C. (2002). Movement sequencing and phonological fluency in (putatively) nonimpaired readers. *Psychological Science, 13(4)*, 375-379.

Case, R. (1992). The role of the frontal lobes in the regulation of cognitive development. *Brain and Cognition*, 20, 51-73.

Chelune, G.J., Ferguson, W., Koon, R., Dickey, T.O. (1986). Frontal lobe disinhibition in attention deficit disorder. *Child Psychiatry and Human Development, 16*, 221-234.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge: MIT Press.

Cleary, M., Dillon, C., Pisoni, D.B. (2002a). Imitation of nonwords by deaf children after cochlear implantation: Preliminary findings. *Annals of Otology, Rhinology, and Laryngology, Suppl. 189*, 91-96.

Cleary, M., Pisoni, D.B., Kirk, K.I. (2002b). Working memory spans as predictors of spoken word recognition and receptive vocabulary in children with cochlear implants. *The Volta Review, 102(4)*, 259-280.

Cleary, M., Pisoni, D.B. (2007). Visual and visual-spatial memory measures in children with cochlear implants. Unpublished manuscript.

Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review, 12(5)*, 769-786.

Conway, C.M., Karpicke, J., Pisoni, D.B. (2007). Contribution of implicit sequence learning to spoken language processing: Some preliminary findings with hearing adults. *Journalof Deaf Studies and Deaf Education, 12,* 317-334.

Cohen, M.J. (1997). *Children's Memory Scale*. San Antonio: The Psychological Corporation.

Dillon, C.M., Cleary, M., Pisoni, D.B., Carter, A.K. (2004). Imitation of nonwords by hearing impaired children with cochlear implants: Segmental analyses. *Clinical Linguistics and Phonetics, 18(1)*, 39-55.

Edwards, J., Beckman, M.E., Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research, 47*, 421-436.

Eisenberg, L.S., Martinez, A.S., Holowecky, S.R., Pogorelsky, S. (2002). Recognition of lexically controlled words and sentences by children with normal hearing and children with cochlear implants. *Ear and Hearing, 23(5)*, 450-462.

Engle, R.W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19-23.

Fagan, M.K., Pisoni, D.B., Horn, D.L., Dillon, C.M. (in press). Neuropsychological processes associated with vocabulary, reading, and working memory in deaf children with cochlear implants. *Journal of Deaf Studies and Deaf Education.*

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D.M.R., Gantz, B.J., Woodworth, G.G. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *Journal of Speech, Language, and Hearing Research, 40*, 183-199.

Fu, Q.J., Galvin, J.J. (2003). The effects of short-term training for spectrally mismatched noise-band speech. The *Journal of the Acoustical Society of America, 113(2)*, 1065-1072.

Gathercole, S.E., Baddeley, A.D. (1993). *Working memory and language*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Gathercole, S.E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27*, 513-543.

Gathercole, S.E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences, 3*, 410-419.

Gathercole, S.E., Baddeley, A.D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28*, 200-213.

Gathercole, S.E., Service, E., Hitch, G.J., Adams, A.-M., Martin, A.J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology, 13*, 65-77.

Gathercole, S.E., Willis, C., Emslie, H., Baddeley, A.D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*, 887-898.

Gupta, P., Lipinski, J., Abbs, B., Lin, P.-H. (2005). Serial position effects in nonword repetition. *Journal of Memory and Language, 53*, 141-162.

Gupta, P., MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59*, 267-333.

Helland, T., Asbjørnsen, A. (2000). Executive functions in dyslexia. *Child Neuropsychology, 6(1)*, 37-48.

Horn, D.L., Davis, R.A.O., Pisoni, D.B., Miyamoto, R.T. (2004). Visuomotor integration ability of pre-lingually deaf children predicts audiological outcome with a cochlear implant: a first report. *International Congress Series, 1273*, 356-359.

Horn, D.L., Fagan, M.K., Dillon, C.M., Pisoni, D.B., Miyamoto, R.T. (in press). Visuo-motor integration skills of prelingually deaf children: Implications for pediatric cochlear implantation. *Laryngoscope*.

Horn, D.L., Pisoni, D.B., Sanders, M., Miyamoto, R.T. (2005). Behavioral assessment of prelinguallydeaf children before cochlear implantation. *Laryngoscope, 115*, 1603-1611.

Horn, D.L., Pisoni, D.B., Miyamoto, R.T. (2006). Divergence of fine and gross motor skills in prelingually deaf children: implications for cochlear implantation. *Laryngoscope, 116*, 1500-1506.

Kalikow, D.N., Stevens, K.N., Elliott, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61*, 1337-1351.

Kimberg, D.Y., D'Esposito, M., Farah, M.J. (1996). Frontal lobes: Cognitive neuropsychological aspects, in Feinberg, T.E., Farah, M.J. (eds.): *Behavioral Neurology and Neuropsychology*. New York: McGraw-Hill.

Kirk, K.I., Pisoni, D.B., Osberger, M.J. (1995). Lexical effects of spoken word recognition by pediatric cochlear implant users. *Ear and Hearing, 16*, 470-481.

Korkman, M., Kirk, U., Kemp, S. (1998). *NEPSY: A Developmental Neuropsychological Assessment*. China: PsychCorp.

Korkman, M., Kirk, U., Kemp, S. (2001). Effects of age on neurocognitive measures of children ages 5 to 12: A cross-sectional study on 800 children from the united states. *Developmental Neuropsychology, 20(1)*, 331-354.

Lenneberg, E.H. (1967). *Biological foundations of language*. New York: Wiley and Sons.

Liss, M., Fein, D., Allen, D., Dunn, M., Feinstein, C., Morris, R., Waterhouse, L., Rapin, I. (2001). Executive functioning in high-functioning children with autism. Journal *of Child Psychology and Psychiatry, 42 (2)*, 261-270.

Luce, P.A., Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19(1)*, 1-36.

Luria A. R. (1961) *The Role of Speech in Regulation of Normal and Abnormal Behaviour.* Oxford: Pergamon Press.

Marjerus, S., Poncelet, M., Greffe, C., & Van der Linden, M. (2006). Relations between vocabulary development and verbal short-term memory: The relative importance of short-term memory for serial order and item information. *Journal of Experimental Child Psychology, 93*, 95-119.

Meyer, T.A., Pisoni, D.B. (1999). Some Computational Analyses of the PBK Test: Effects of Frequency and Lexical Density on Spoken Word Recognition. *Ear and Hearing, 20(4)*, 363-371.

Michas, I.C. & Henry, L.A. (1994). The link between phonological memory and vocabulary acquisition. *British Journal of Developmental Psychology, 12*, 147-164.

Miller, G.A. & Selfridge, J.A. (1950). Verbal context and the recall of meaningful material. *American Journal of Psychology, 63*, 176-185.

Miyake, A. Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100.

O'Reilly, R.C, Munakata Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain.* Cambridge, MA: MIT Press.

Oliver, B., Dale, P.S., Plomin, R. (2004). Verbal and non-verbal predictors of early language problems: An analysis of twins in early childhood back to infancy. *Journal of Child Language, 31*, 609-631.

Osberger, M.J., Todd, S.L., Berry, S.W., Robbins, A.M., Miyamoto, R.T. (1991). Effect of age ofonset of deafness on children's speech perception abilities with a cochlear implant. *Annals of Otology, Rhinology, and Laryngology, 100 (11)*, 883-888.

Orsini, A., Grossi, D., Capitani, E., Laiacona, M., Papagno, C., Vallar, G. (1987). Verbal and spatial immediate memory span: Normative data from 1355 adults and 1112 children. *Italian Journal of Neurological Science, 8(6)*, 539-548.

Ozonoff, S., Jensen, J. (1999). Specific executive function profiles in three neurodevelopmental disorders. *Journal of Autism and Developmental Disorder, 29*, 171-177.

Ozonoff, S., McEvoy, R. E. (1994). A longitudinal study of executive function and theory of mind development in autism. *Development and Psychopathology, 6,* 415-431.

Pennington, B.F., Ozonoff, S., (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*, 51-87.

Pisoni, D.B., Svirsky, M.A., Kirk, K.I., Miyamoto, R.T. (1997). Looking at the ''stars'': A first report on the intercorrelations among measures of speech perception, intelligibility, and language development in pediatric cochlear implant users. *Research on Spoken Language Processing Progress Report No. 21(1996–1997)* Bloomington, Indiana: Indiana University: 51-91.

Pisoni, D.B., Cleary, M., Geers, A., Tobey, E. (2000). Individual differences in effectiveness of cochlear implants in children who are prelingually deaf: New process measures of performance. *Volta Review, 101*, 111-164.

Pisoni, D.B., Geers, A.E. (2000). Working memory in deaf children with cochlear implants: Correlations between digit span and measures of spoken language processing. *Annals of Otology, Rhinology, and Laryngology, Suppl.* 185, 92-3.

Powell, R.P., Bishop, D.V. (1992). Clumsiness and perceptual problems in children with specific language impairment. *Developmental Medicine and Child Neurology, 34(9)*, 755-765.

Prior, M. R., Hoffmann, W. (1990). Brief report: Neuropsychological testing of autistic children through an exploration with frontal lobe tests. *Journal of Autism and Developmental Disorders, 20,* 581-590.

Rosen, S., Faulkner, A., Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America, 106(6)*, 3629-3636

Rubenstein, H. (1973). Language and probability. In G.A. Miller (Ed.), *Communication, language, and meaning: Psychological perspectives* (pp. 185-195). New York: Basic Books, Inc.

Rumsey, J. M. (1985). Conceptual problem-solving in highly verbal, nonretarded autistic men. *Journal of Autism and Developmental Disorders, 15,* 23-36.

Russell, W.R. (1948). Function of frontal lobes. *Lancet, 254*, 356-360.

Siegel L., Saigal, S., Rosenbaum, P., Morton, R.A., Young, A., Berenbaum, S., Stoskopf, B.(1982). Predictors of development in preterm and full-term infants: a model for detecting the at risk child. *Journal of Pediatric Psychology, 7*, 135-148.

Singer, B.D., Bashir, A.S. (1999). What are executive function and self-regulation and what do they have to do with language-learning disorders? *Language, Speech, and Hearing Services in Schools, 30*, 265-273.

Sparrow, S.S., Cicchetti, D.V., Balla, D.A. (2005). *Vineland-II: Vineland Adaptive Behavior Scales, Second Edition.* Circle Pines, MN: AGS Publishing.

Stallings, L.M., Kirk, K.I., Chin, S.B., Gao, S. (2000). Parent Word Familiarity and the Language Development of pediatric Cochlear Implant Users. *The Volta Review, 102(4)*, 237-285.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, J8,* 643-662.

Stuss, D.T. (1992). Biological and psychological development of executive functions. *Brain and Cognition, 20*, 8-23.

Thatcher, R.W. (1992). Cyclical cortical reorganization during early childhood. *Brain and Cognition, 20*, 24-50.

Viding, E., Price, T.S., Spinath, F.M., Bishop, D.V.M., Dale, P.S., Plomin, R. (2003). Genetic and environmental mediation of the relationship between language and nonverbal impairment in four year-old twins. *Journal of Speech, Language, and Hearing Research, 46*, 1271-1282.

Wechsler, D. (1991). *Wechsler intelligence scale for children – third edition*. San Antonio: The Psychological Corporation.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review, 9 (4)*, 625-636.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Audiovisual Perception of Spoken Words in Speech and Nonspeech Modes: Measures of Architecture and Capacity[1]

Nicholas A. Altieri and James T. Townsend[2]

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This study was supported NIH Grants DC-00111 and DC-00012 to Indiana University. I would like to acknowledge Jeremy Loebach, Mario Fific, and David B. Pisoni for insightful comments.
[2] Indiana University, Bloomington. jtownsen@indiana.edu

Audiovisual Perception of Spoken Words in Speech and Nonspeech Modes: Measures of Architecture and Capacity

Abstract. Contemporary models of audiovisual speech perception attempt to explain accuracy data based on curve fitting and optimization techniques (see Braida, 1991; Massaro, 1987). Research on audiovisual speech perception lacks a formal mathematical foundation because current models do not make predictions about reaction time or adequately describe how the audio and visual channels are processed in the black box. The double factorial paradigm (DFP) developed by Townsend and Nozawa (1995) uses systems factorial technology to provide a framework for investigating how different information channels are processed. In Experiment 1, participants were required to make one response if auditory information, visual information, or both forms of information were present and a negative response on target absent trials. Data from the audiovisual detection task with the word "base" as the stimulus showed that processing architecture was either coactive or parallel self-terminating. A second experiment again using the double factorial paradigm methodology (Experiment 2) required participants to distinguish between two spoken words: "base" and "face." The data showed that processing was mostly coactive, but possibly parallel self-terminating in some cases. Processing capacity was limited in both experiments, indicating a lack of redundancy gain. Overall, these results suggest that the audio and visual channels are combined into a single processor, although inhibition or competition may exist between channels.

Introduction

The cognitive or information processing approach to psychology seeks to understand in a mathematically rigorous fashion how information is processed in the "black box." Given a certain number of distinct inputs to the system, the output or subject's response is measured, but the psychologist would ultimately want to understand the cognitive mechanisms that produced the output. Speech perception for example, is a multimodal perceptual faculty that relies on auditory, visual, and even haptic information as inputs to the system—where word or segment recognition is the output (Fowler & Dekle, 1991; Sumby & Pollack, 1954). Sumby and Pollack demonstrated the importance of the contribution of visual information in speech perception by showing that the proportion of audiovisual gain remains identical across all signal to noise ratios. It is also well established that when listeners are presented with incongruent audiovisual stimuli, the resulting percept is different than either the audio or visual stimuli, as is the case in the McGurk effect. The auditory stimulus was the utterance /ba/, which was dubbed over a visually articulated /ga/, and in the majority of cases, subjects reported experiencing the "perceptual fusion" /da/ (see McGurk & McDonald, 1976).

Researchers have long investigated the output of the black box and established that fact speech perception is a multi-modal phenomenon. However, broad classes of models related to the way audio and visual stimuli are processed in black box such as serial, parallel, or coactive processing have not been falsified or investigated. The mechanisms that listeners use to extract and combine information from different modalities in real time are not understood. An investigation of the processing architecture (i.e., parallel or serial) in the "black box" would provide a fundamental foundation for scientific investigations of audiovisual perception.

Most research on audiovisual speech perception assumes that listeners somehow combine information from the individual modalities, without explaining how integration occurs in the "black box"

or a neurologically based model. The Fuzzy Logic Model of Perception FLMP is one class of models, which assumes *a priori* that audiovisual integration occurs in an optimal fashion, where the relationship between audio and visual information are multiplied and divided by the sum of the alternatives (Massaro, 2004). FLMP uses a formulation of Bayes' theorem to determine the probability that a certain syllable, word, or phoneme was processed given the available audio and visual parameters.[3] A second model referred to as the pre-labeling integration model (PRE) is founded upon multidimensional signal detection theory, and assumes that the unimodal information scores will be used optimally, and that the predicted AV scores should be greater than or equal to observed unimodal identification scores (Braida, 1991).

While FLMP and PRE account for confusion data when tested in audiovisual perception experiments (Grant, 2002; Grant, Tufts, & Greenberg, 2007; Massaro, 2004), they do not attempt to explain how cognitive systems process information from the audio and visual channels. The question is how are the audio and visual channels utilized and combined in real time to form a unified percept? A second and related point is that models of audiovisual perception do not make fine-grained predictions about reaction time data, which generally precludes mathematical modeling of dynamic processes.

Figure 1 shows two prominent conceptual accounts or neural representations of how integration might occur in an information processing system, along with a serial processing model where processing cannot begin on the second channel until it finishes on the first channel. The parallel model has independent channels where separate decisions are made on each channel. In this framework, the audio and visual speech streams are processed separately and simultaneously just prior to the decision stage. A separate decision is made on each channel or modality and a subsequent decision is made using an AND or an OR gate. Consider for example a case where a listener is given a task where they have to respond "yes" if presented with /ba/ in either the audio or visual modality. When /ba/ is presented, each channel accumulates information and if the auditory channel exceeds threshold, the listener responds "yes" regardless of whether the visual channel is finished accumulating information. In a coactive model, the information from each channel is combined into a common information processor that counts information from each source. Once the counter in this common processor exceeds threshold, a decision is made. Lastly, in the serial model, processing on the audio and visual components of /da/ or /da/ cannot occur simultaneously. If the auditory component is processed first, for example, then processing in the visual domain cannot begin until the audio channel is completely finished. If the system is self-terminating, then a decision can be made when the audio channel finishes, whereas if the stopping rule is exhaustive, both channels must finish.

FLMP and PRE do not make explicit predictions about serial, parallel, or coactive processing architecture. Massaro (2004) claims that the algorithms used in FLMP can implement either the parallel or coactive models depicted in Figure 1. A major undertaking in this project is to garner behavioral evidence to distinguish between the models depicted in Figure 1. Two general candidates for audiovisual speech recognition include coactive processing and parallel non-convergent processing, although serial processing will also be considered.

[3] In a two alternative forced choice task where the listener has to distinguish between /ba/ and /da/, the probability of a given value is a function of the audio and visual parameters: $p(/da/ \mid A \& V) = aivj/[aivj + (1-ai)(1-vj)]$.

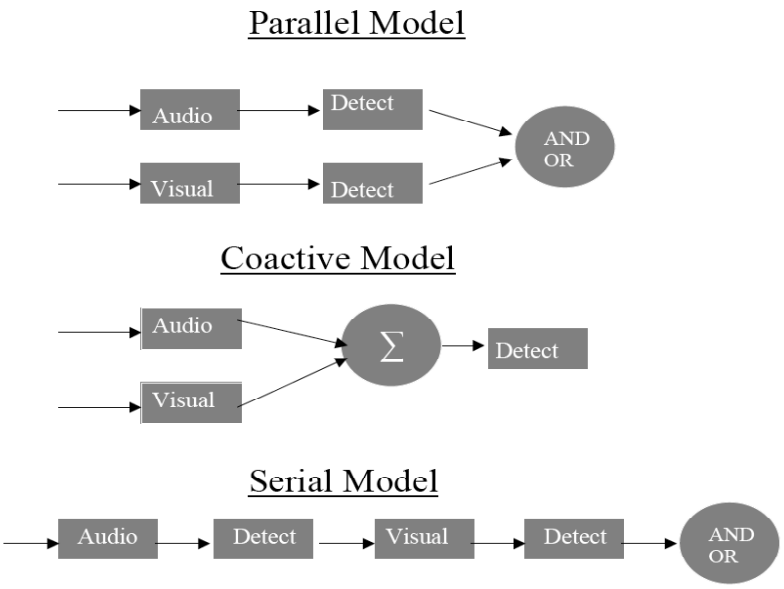


Figure 1. Audiovisual Processing accounts. On top is a schematic representation of a parallel model with an OR as well as an AND gate. The coactive model below assumes that each channel is pooled into a common processor where evidence is accumulated prior to making a decision. The serial model at the bottom assumes that processing occurs one stage at a time. Processing cannot begin on the second channel on stage two unless processing on the first channel is completed.

Audiovisual Speech Perception

While formal mathematical models have not been applied to distinguish coactive versus parallel processing, there has been discussion in the audiovisual speech perception literature pertaining to different processing architectures for the audio and visual channels. For instance, speech perception theorists from different schools of thought like motor theory (Liberman & Mattingly, 1985) direct realism (Fowler & Rosenblum, 1991), and other general processing theories (see Bernstein, 2005) differ in how they conceptualize audiovisual information processing. Motor theory and direct realism for instance, assume that the primitives of speech perception are articulatory gestures.[4] Rosenblum (2005) argues that the evidence of the importance of multimodal speech perception supports gesture based theories, and draws the conclusion that multimodal speech is the *primary* function of perception. He argues that information in the speech signal is present in every modality, and the perceptual processes involved in recognizing speech are "unconcerned" with regard to modality. Gesture based theories do not make explicit mathematical predictions with regard to the mappings between the auditory and visual channels. However, one way to illustrate this framework in the context of audiovisual perception is to conceptualize the information from the audio and visual modalities becoming "integrated" and combined into a single channel "early" in the decision process prior to word recognition (where the decision process considers only the sum of the information and not the information in the individual modalities), as depicted in the "coactive" model in Figure 1.

[3] In the case of motor theory, the motor gestures that produced the sounds are recovered by the listener using analysis by synthesis. For direct realism, information about gestures is carried by the speech signal and is perceived directly. For simplicity, motor theory and direct realism will be treated identically with regard to audiovisual perception in this paper.

Behavioral studies have provided some support for the view that speech perception is "unconcerned" with source modality, or that audiovisual integration occurs early, i.e., prior to word or segment recognition. Green and Miller (1985) demonstrated that visually perceived rate of articulation influences auditory segment perception. They used a McGurk paradigm to show that visual information about place of articulation can influence properties like voice onset time. Subjects were shown audiovisual clips of a talker saying a syllable that varied auditorially and visually on a continuum from /bi/ to /pi/. The corresponding visual information was played either fast or slow. They showed that slowly articulated syllables increased the percentage of time that subjects perceived /bi/ relative to /pi/. Because visual information influences the perception of features that are the components of word recognition, these findings indicate "early" integration of audiovisual channels, in which audio and visual information is combined into a single channel prior to word and segment recognition. They argued that the results were indicative of a decision process that has access to both auditory and visual information and combines the two sources of information prior to recognition.

Neuroimaging evidence from audiovisual speech perception tasks has suggested similar conclusions about the presence of coactive processing. Calvert and Campbell (2003) showed that silent lipreading tasks activate the primary auditory cortex. Subjects were presented with either sequences of still key frames or moving images of the same duration of a talker saying nonsense syllables. Subjects were instructed to look for a visible target syllable like "voo" in a sequence of other nonsense syllables. In contrast to resting conditions in which letters were superimposed on a resting face, sequences of still key frame images produced activation in the posterior cortical areas associated with the perception of biological motion. Activation was also observed in canonical speech processing areas including Broca's area, the superior temporal sulcus (STS). However, moving images produced greater activation in these regions compared to still frames. They concluded that visual speech accesses areas traditionally believed to be auditory processing regions for language, which is possibly due to "dynamic audiovisual integration mechanisms" in the STS (Calvert & Campbell, 2003).

Super-additive activation in the STS has also been observed in congruent audiovisual speech perception tasks (Calvert et al., 1997), while incongruent audiovisual speech has yielded sub-additive activation in the STS (Calvert, Campbell, & Brammer, 2000). Super-additive activation occurs when the amount of activation recorded in a brain area in the bimodal condition is greater than the sum of the activation levels from each unimodal condition. The observation of super-additive levels of activation in the STS indicates the possibility that there are neurons and brain regions that only respond, or mostly respond to audiovisual input. The existence of neurons that respond selectively to audiovisual input provides at least some evidence that the brain might be implementing an information processing system analogous to the coactive model depicted in Figure 1 where the audio and visual components of the signal are combined into one channel prior to segment or word recognition.

Nonetheless, the conclusion that multi-sensory neurons are responsible for processing audiovisual speech is not uniformly accepted. The BOLD response is a measure of the blood oxygen level in a brain region and therefore represents an indirect measure of neural activity. fMRI designs also suffer from poor temporal resolution. Observations of super-additive levels of activation in the STS could be due to "commingled" unisensory neurons (Bernstein, Auer, & Moore, 2004; Meredith, 2002). That is, areas that are believed to respond only to audiovisual speech in reality contain large numbers of unisensory neurons. Furthermore, the STS responds not only to speech, but also to complex nonspeech gestures (Puce, Allison, Bentin, Gore, & McCarthy, 1998). When presented with pairs of moving eyes or moving mouths, bilateral activation was observed in the posterior STS, while the control stimuli consisting of moving checkered patterns did not activate the STS or surrounding areas. These data appear to indicate that the auditory and visual streams are not converging to a common processor, and therefore there is insufficient evidence for a coactive processing model.

Bernstein (2005) argued instead that while speech is part of a highly specialized cortical system, not all motor and perceptual areas of the cortex seem to be devoted to speech perception, as gestural theories would assume. According to Bernstein, auditory and visual speech stimuli might be processed separately and simultaneously and "converge" only after phonetic perception and word recognition. Bernstein reasons that multimodal perception of the speech signal involves separate and simultaneous analysis of the audio and visual inputs. According to this account, the information from the audiovisual speech streams is processed in parallel, where extensive unisensory processing occurs before the binding of auditory and visual speech representations. This view is analogous to the parallel model discussed in Figure 1, which differs architecturally from the coactive model where one common processor integrates audio and visual information prior to phonetic perception.

Double Factorial Paradigm: Assessing Architecture and Capacity

Given the coactive and parallel models of integration in the context of Rosenblum (2005) and Bernstein's (2005) respective analyses on audiovisual speech perception, it is pertinent to return to the purpose of this project by finding a way to distinguish between these two models. The double factorial paradigm (DFP) developed by Townsend and Nozawa (1995) is an experimental methodology that can be used to obtain behavioral evidence to distinguish parallel from coactive processing. The description of the coactive and parallel models in the speech perception literature, while of theoretical importance, requires a more specific mathematical formulation along with behavioral data if they are to be adequately distinguished due to conflicting and imprecise accounts discussed in previous paragraphs.

The methodology for assessing mental architecture involves a factorial methodology that captures potential interactions between factors. One statistic that has been used to analyze interactions is the mean interaction contrast, or MIC = RTll – RTlh – (RThl – RThh) (see Sternberg, 1969). In this formula, RT designates reaction time, and each subscript represents the level of one factor like presence or absence of a feature or brightness: h = high, which indicates fast reaction times and l = low, which indicates slower reaction times. The hh condition for example might represent audio and visual stimuli of a high level of clarity, which a listener would be able to identify more quickly than if the audio of visual portions (or both) were degraded or less salient. One shortcoming of the MIC is that it is a coarse measure representing only one point at each level (i.e., the mean or median of the distribution). Townsend and Nozawa (1995) developed a more sensitive measure that analyzes the curve of the entire distribution of reaction times referred to as the *survivor interaction contrast* (SIC). The SIC is defined as $SIC(t) = Sll(t) - Slh(t) - (Shl(t) - Shh(t))$. Notice that the SIC uses the same sequence of terms as the MIC, only this time survivor functions are used rather than mean reaction times. Let $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function of the density function $f(t)$ of reaction times. The survivor function $SIC(t)$, is a distribution function indicating the probability that a process is still going on. If audiovisual stimuli is presented, then $SIC(t)$ would indicate the probability that the word, phoneme, or stimulus has not been recognized and identified by the subject by time t.

The SIC function makes several predictions about processing architecture. For the type of parallel processing described by the non-convergent model which assumes that each channel has its own decision stage, the SIC function can be positive or negative depending on the stopping rule. A parallel model with separate decisions and an exhaustive stopping rule predicts a negative SIC curve. "Exhaustive processing" refers to a stopping rule in a parallel system where each channel must finish processing before a decision is made. The reason for underadditivity in parallel exhaustive models is because each element must be completed before the system terminates. In other words, the processing of the system is determined by the slowest element. On the lh or hl trials, the longest time tends to be closer to the longest time on the ll

trials. Thus, the difference between Sll (t) – Slh (t) is generally smaller than the difference between Shl (t) – Shh(t).

The case is exactly the opposite for parallel minimum time self-terminating models (or horse race models), which terminate when the fastest element finishes. The SIC function for these models is positive since the difference between Sll(t) – Slh(t) is generally greater than the difference between Shl(t) – Shh(t). The reason is because the lh trials have an element that takes less time to process.

Coactivation might be considered a class of parallel models where the information from each channel is pooled into a single channel governed by a Poisson summation process. The survivor interaction function for Poisson summation models is negative at the beginning for low t, and becomes positive at later times t. The mean interaction contrast is positive. While the shape of the SIC function may not conform to intuition, it does make sense mathematically. The rate of coactive models is the sum of the rates of each channel—the sum of the audio and visual channels. For certain time t, the contrast will either be positive or negative. The SIC function is a function of the rate parameter and the curvature corresponds to the sign of the second derivative, which as stated above is negative for low t, and becomes positive as t increases (Townsend & Nozawa, 1995).

Finally, serial processing predicts an MIC of 0 regardless of whether the stopping rule is exhaustive or self-terminating. When processing in serial with a self-terminating stopping rule, the SIC(t) function is flat and equal to 0 at each point. Interestingly in the exhaustive case, the SIC(t) resembles an S-shaped curve with a negative region for early processing times and a positive region for later processing times (Townsend & Nozawa, 1995). The negative and positive regions of the curve are equal to each other in serial exhaustive model, and if we integrate over the curve, the total area is equal to zero.

Capacity and Audio-Visual Gain in Speech Perception

A second feature of the DFP is its ability to assess the *capacity* of the system. Capacity is a measure that determines how the number of channels present affects the processing speed at a given time t. In other words, is there a cost, benefit, or no change in processing when both audio and visual channels are present (redundant target) relative to conditions when only the audio or visual channel is operating (single target)? If the processing rate is unaffected by increasing the number of channels, the system operates at unlimited capacity, if it slows down, then it operates at limited capacity, and if there is a benefit in processing rate, then it operates at super capacity.

Measuring processing capacity requires looking at the ratio of the integrated Hazard functions. The form of the hazard function is given below.

$$h(t) = f(t)/[S(t)] \quad (1)$$

Where f(t) is the probability density function, and S(t) is the survivor function which yields the probability that a process has not yet finished. The hazard function h(t) indicates the probability that a process will terminate at the next moment (t + 1) in time given that it has not yet terminated at time t.

To calculate the capacity coefficient C(t) at each point in time, we calculate the integrated hazard function for the conditions where the subject is presented with the redundant target and divide it by the sum of the integrated hazard functions of the single target conditions (Townsend & Nozawa, 1995). The subscripts A and V indicate the audio and visual channels.

$$C(t) = HAV(t)/[HA(t) + HV(t)] \quad (2)$$

The integrated hazard function H(t) is equivalent to log[1 – F(t)] or log[S(t)], and in the field of physics it is used as a measure of the total energy consumed. The system operates at super capacity at a certain point in time t if C(t) is greater than 1 at that point, unlimited capacity if it equals 1, and limited capacity if it is less than 1 (Wenger & Townsend, 2000).

As previously stated, it is known that congruent audiovisual information about spoken words facilitates accuracy levels in perception (Sumby & Pollack, 1954). However, the notion of processing capacity as defined above has generally been left unaddressed in the audiovisual speech perception literature, although research has been conducted investigating redundant target effects for nonspeech auditory and visual stimuli (see Berryhill, Kveraga, Webb, & Hughes, 2007; Miller, Kuhlwein, & Ulrich, 2004; Schroter, Ulrich, & Miller, 2007 for a discussion). Berryhill et al. (2007) presented subjects with congruent audiovisual stimuli (with a visual lead (SOA) of 0, 75ms, 150ms, and 225ms). The stimuli consisted of symbolic tokens of the numerals 1 and 2 presented in the visual modality, and auditory tokens of a talker saying 1 or 2, where the task of the participants was to determine whether '1' was presented or '2' was presented. Each trial was an audio only trial, visual only trial, or audiovisual trial (redundant target). They observed limited capacity, or lack of redundancy gain, when presentation of the audio and visual components was synchronized. When the lead (SOA) of the visual stimuli increased, capacity became less limited, and at SOAs of 150 and 225ms, a redundancy gain was observed.

In this study, the double factorial paradigm was applied in two separate experiments to test architecture and capacity in a control study where subjects were not required to attend to speech (i.e., nonspeech mode: Experiments 1A and 1B). A second Experiment (2) was conducted where subjects were required to distinguish between two spoken words. Both experiments used RT data to test audiovisual processing architecture and capacity using the formal framework of the double factorial paradigm. These experiments were designed to look inside the black box and begin to analyze whether processing of audiovisual components is parallel, coactive, or even serial in tasks where subjects were required to identify the presence of a talker or distinguish between spoken words of English. Experiment 1 was an audiovisual detection task using video clips of a single talker as stimuli. This experiment was a control study where subjects were exposed to a talker speaking a word of English. They were required to focus on the surface properties of the stimuli to judge whether a stimulus was present or absent, and were required to detect stimuli rather than engage in spoken word recognition. We compared the results (i.e., architecture and capacity) from Experiment 1 with the results from Experiment 2. The purpose was to assess whether the results from the speech perception experiment were particular to high-level cognition such as spoken word recognition, or whether they reflect general audiovisual processing mechanisms involved in simply identifying "complex" stimuli like the moving face of a talker.

Experiment 1A

Participants

Seven subjects (four females and three males) with normal or corrected vision were paid ten dollars per session for their participation. Data analysis was not conducted for one subject who only completed one session.

Materials

The experiment was carried out in the Speech Research Laboratory (SRL) at Indiana University in Bloomington. The stimulus materials included an audiovisual movie clip of a female talker from the Hoosier Multi-talker Database saying the word "Base". A total of eight different stimuli were created

from this video clip: two audio files at two levels of saliency, two video files at two levels of saliency, and four audiovisual clips at each factorial combination of high-high, high-low, low-high, and low-low levels of saliency. The audio, visual, and audiovisual files were edited using Final Cut Pro HD version 4.5. The audio files were sampled at a rate of 48 kHz at a size of 16 bits. The high saliency audio files were presented at 57 dB and the low saliency audio files presented at a volume of 45 dB. The brightness level on the video files was manipulated to create two different levels of saliency. On the low saliency video files, the brightness was reduced 90 steps using the brightness video filter. This had the effect dimming and reducing the contrast of the video, making it mode difficult to perceive the talker's articulators. Both audio and video files lasted for a total duration of approximately 1,600 milliseconds.

Design and Procedure

Subjects were seated 14 to 16 inches in front of a Macintosh computer equipped with Beyer Dynamic-100 headphones. Each trial began with a fixation cross appearing in the center of the computer screen followed by either a target absent trial, or one of the eight possible stimuli: target present and target absent. One fourth of the trials were target absent trials in which no stimulus appeared after the plus sign on the center of the screen. The stimulus trials included either audio only, visual only, or audiovisual stimuli. Experiment 1 was an OR design where subjects were instructed to respond, as quickly and accurately as they possible by pressing the button labeled "Base" if they heard either the word "base" (audio only), saw the talker utter the word "base", or were exposed to a redundant target where both the audio and visual components of the word "base" were present. They were instructed to respond by pressing the button labeled "Nothing" if no stimulus appeared on the screen. There was a 750-millisecond delay between trials.

There were a total of 800 target-absent trials, 800 audio only trials, 800 audiovisual trials, and 800 visual only trials for a total of 3,200 trials per subject (1/4 Nothing, 1/4 A only, 1/4 V only, 1/4 AV). This included 200 trials in each redundant target condition (hh, hl, lh, ll). Participants were run for 40 blocks at 80 trials each with a break scheduled between each block. Participants also received sixteen practice trials at the onset of each experimental session that were not included in the subsequent data analysis. The experiment lasted approximately 45 minutes and was conducted over a course of 4 days.

Results and Discussion

Percentage of errors averaged across all participants was less than 2 %. Evidence of a speed-accuracy trade off was not observed. Therefore, only reaction time results will be presented.

The primary focus in Experiment 1A was on the set of SIC curves for each participant, which are distribution free (Townsend & Nozawa, 1995). Data from each participant was analyzed separately rather than averaged together since the results would have obscured individual differences and possibly led to different conclusions (see Townsend & Fific, 2004). ANOVAs and the mean interaction contrast (MIC) were analyzed in this experiment because they can help confirm or disconfirm interactions between the factorial conditions, which is an important tool for disconfirming serial processing. Serial processing would display a MIC of 0 (no interaction) and a flat SIC. The integral of the SIC curve is equal to the mean interaction contrast. Results of the SIC and mean interaction will be discussed together. Finally, the capacity coefficient, C(t), which is a measure of the system's capacity at time t, was also of interest and will be addressed in subsequent analyses.

SIC curves for four participants who demonstrated selective influence appear in Figure 2. The MIC appears in Table 1 along with the ANOVA results for the four factorial conditions. A bin size of 10 milliseconds was used to calculate each survivor function in each experiment. Recall that each participant

completed 200 trials in each factorial condition, but errors and outliers (+ or – 3.0 SD from the mean) were eliminated from the analysis.

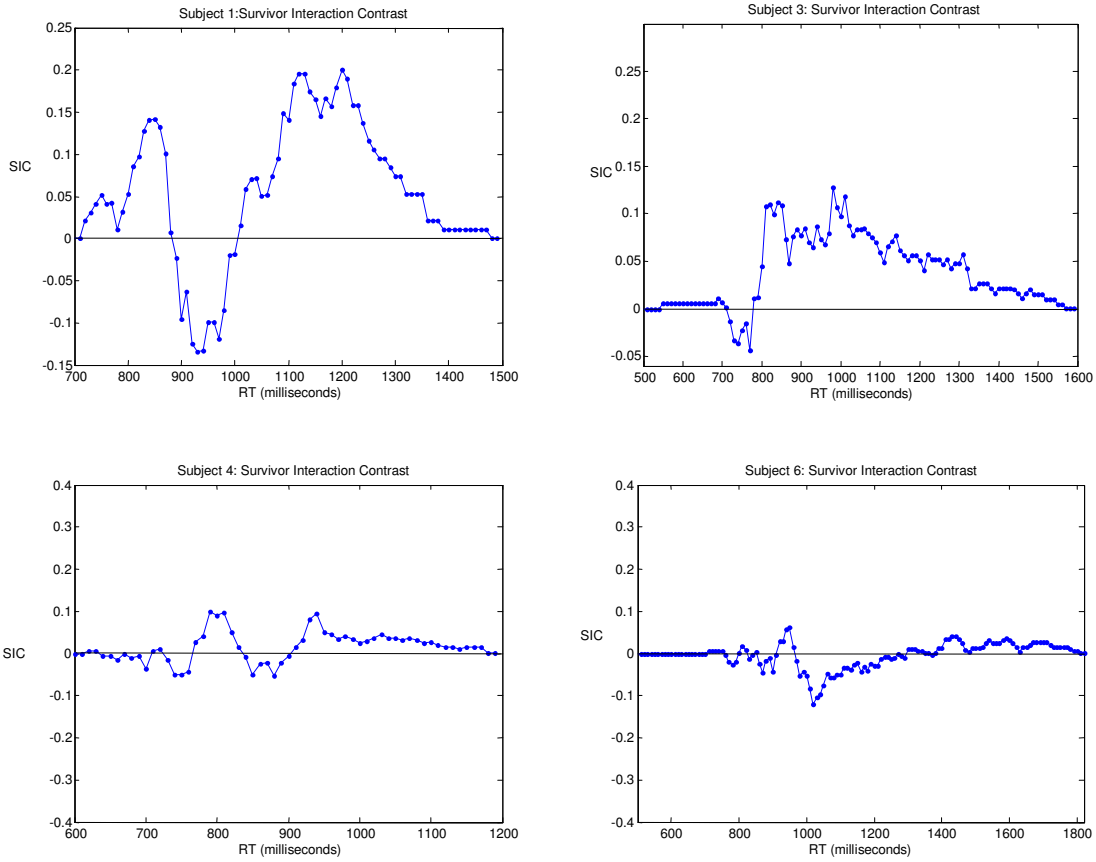


Figure 2. SIC curves for four subjects 1, 3, 4, and 6. These were the subjects who showed selective influence.

Subject 1's ANOVA results shown in table 2 disconfirm serial processing. The SIC curve, while mostly positive, dips below zero yielding a small range of negativity around 900 ms. Since the SIC curve is not entirely positive, it fails to confirm parallel self-terminating processing behavior in this subject. One possible explanation, given the inconsistent curve and the positive interaction is that Subject 1 used dual processing strategies during the task, switching from parallel to serial.

Subject 3's results reveal a positive SIC curve with negativity for early processing times and a positive MIC. This indicates coactive or possibly parallel self-terminating processing that finishes when either the audio or visual channel has reached a decision. The significant results provided by the ANOVA in Table 2 support this conclusion, along with the fact that the capacity coefficient discussed in the following section indicates severely limited capacity.

Subject 4's results suggest either serial self-terminating or indeterminate behavior due to weak selective influence between the redundant target conditions. Subject 4's SIC curve was neither positive nor negative and fit the line $SIC(t) = 0$ with a root mean squared error of .019 and a sum of squared errors

of 1.57. Since the MIC was close to zero and the ANOVA did not even approach significance, we can tentatively accept the result that the behavior for this participant was serial self-terminating.

Subject 6's SIC curve was a flat line like subject 4's curve. Likewise, these results taken together with the MIC are indicative of serial self-terminating, or again indeterminate behavior due to weak selective influence. The fit to the flat line SIC(t) = 0 had a root mean squared error of .017 and a sum of squared errors of 1.40. The corresponding ANOVA did not show a trend toward significance.

The SIC curves and ANOVA results from subject 3 suggested parallel or coactive processing. The SIC curves and ANOVA results from subjects 4 and 6 on the other hand, indicated serial processing. Subject 4's SIC curve was flat and the MIC was close to zero. It is possible in some instances that subjects process audiovisual material in a serial manner and self-terminate when a decision is made. Subject 6's SIC curve, similar to Subject 4's, was generally flat at SIC(t) = 0. The MIC was close to zero without a trend toward an interaction between channels. These ANOVA results added to the evidence that Subject's 4 and 6 processed the audiovisual stimuli in a parallel self-terminating manner.

Subject	df1	df2	*F*	*p*	MIC
1	1	181	11.019	.001	44.28
3	1	180	6.23	.013	41.90
4	1	171	.365	.546	7.28
6	1	181	.014	.905	3.45

Table 1. General Linear Model showing the level of interaction between the audio and visual channels. The Mean Interaction Contrast (MIC) is also displayed**.** This table shows the F value for the mean interaction, the p value (sig. = .05), and the mean interaction contrast.

The second part of this analysis involves examining the system's capacity. Specifically, we were investigating whether having both channels operating increases efficiency, decreases efficiency. Results of the measured capacity coefficient C(t) are compared with the bound for super capacity discussed previously in the introduction in addition to Grice's inequality (see Townsend & Nozawa, 1995).

The performance of each subject in the redundant target condition was compared with the predictions of an unlimited capacity parallel processing model (i.e., C(t) = 1). Figure 3 shows plots of the capacity coefficient for each of the six participants in Experiment 1A. The solid line at C(t) = 1 is the bound for unlimited-super capacity. Data points above the line are indicative of super capacity, data points below the line are indicative of limited capacity, and data points hovering around the line indicate unlimited capacity. The boundary indicated by C(t) = 1/2 represents the Grice bound for limited to extremely capacity. Grice's inequality is defined below:

$$C(t) < MAX[HV(t), HA(t)] / [HV(t) + HA(t)] \quad (3)$$

The value in the numerator is the highest unimodal hazard function or the slower of the two processes. When the distributions of completion times for each channel are identical, Grice's inequality = 1/2.

The definition of "fixed capacity" is the average of the two single target integrated hazard functions (if we assume equal distribution parameters), which means that when two channels are operating, fixed capacity is C(t) = 1/2. Most of the data points fall below Grice's bound for extremely limited capacity and generally hover around C(t) = 1/2. Experiment 1 data support a limited to extremely

limited or fixed capacity model since C(t) < 1 (where C(t) ~ 1/2) for all six subjects across all time bins, even for small values of t.

The data from Experiment 1A indicated variable processing strategies for subjects. One possible reason for variability in processing strategies might have been the long exposure times of the stimuli combined with the simple experimental design. Therefore, the processing architecture data obtained in Experiment 1A is inconclusive. However, the capacity coefficient remained consistent across subjects, which supports the hypothesis that processing capacity is extremely limited in audiovisual detection tasks.

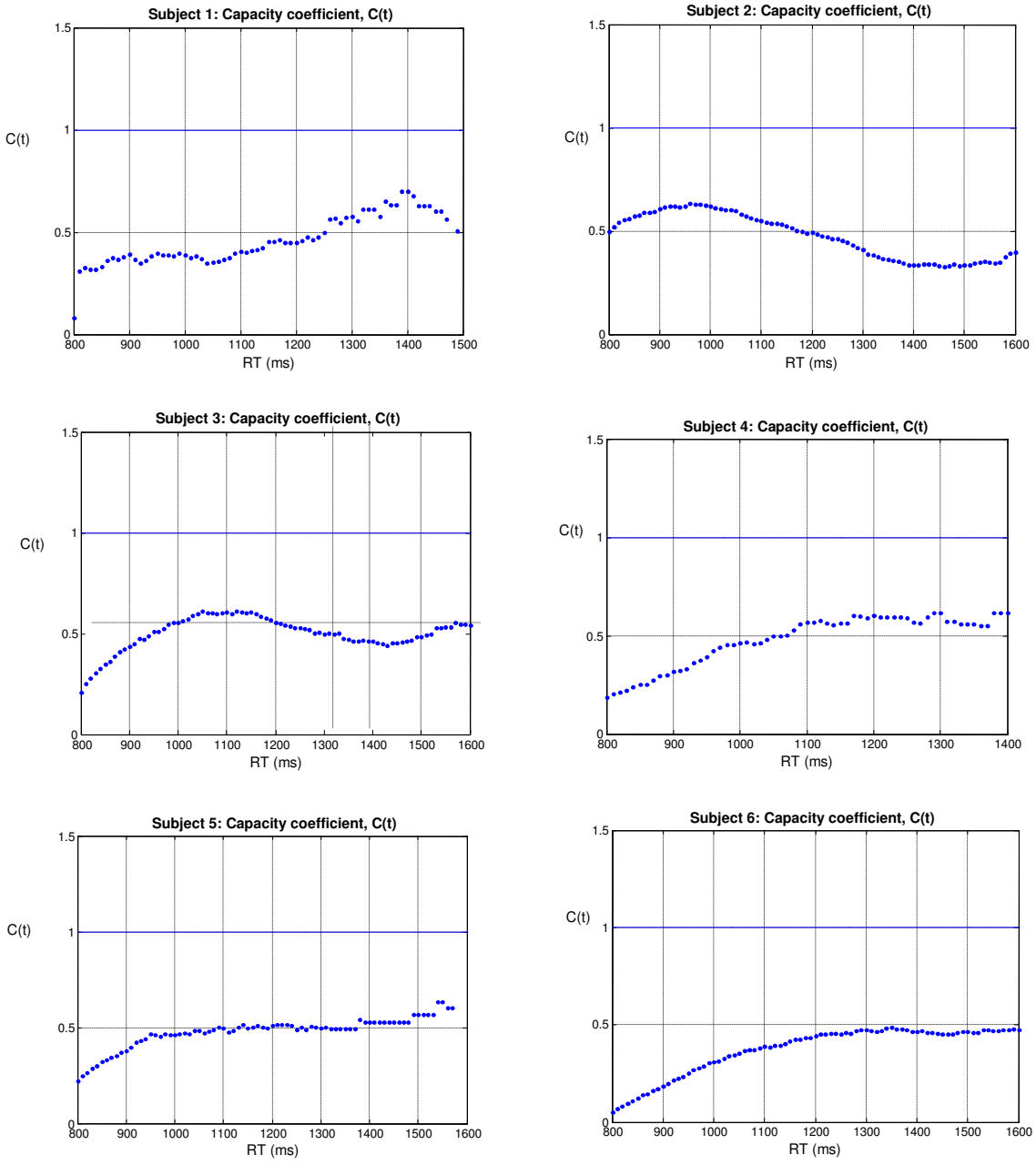


Figure 3. The Capacity coefficient for each of the six participants in Experiment 1A. Processing capacity was extremely limited for each subject.

Experiment 1B

Experiment 1B was a modification of Experiment 1A. The audio and visual stimuli used in Experiment 1A (the female talker saying the word "Base") were shortened where only the first five frames of the video and corresponding audio files were used. This manipulation had the effect of shortening the duration of the audiovisual stimuli from 1,624 ms to approximately 150~160 ms. The purpose of this manipulation was to improve selective influence by helping to reduce eye movements and variability in each of the redundant target reaction time distributions. The audiovisual files were cropped beginning at the onset of the word in "Base". The SIC curves in Experiment 1A were highly variable. Of the four subjects showing selective influence of experimental manipulation, two yielded SIC functions that were basically flat, indicating parallel processing.

Since the stimuli lasted over 1,000 milliseconds in Experiment 1A, it was possible for subjects to move their eyes and therefore potentially shift processing strategies. The purpose of Experiment 1B was to eliminate variable processing strategies by manipulating the duration of the stimulus materials.

Participants

Five participants (two males and three females) with normal or corrected vision were paid ten dollars per session for their participation.

Materials

The materials were identical to those used in Experiment 1. The audiovisual files were shortened using Final Cut Pro HD version 4.5.

Design and Procedure

The design and procedure was identical to task used in Experiment 1A.

Results and Discussion

Percentage of errors averaged across all participants was less than 5%. As in the case of Experiment 1A, evidence of a speed-accuracy trade off was not observed. Therefore, only reaction time results are discussed.

Participants in Experiment 1B showed less between subject variability in the SIC curves. Participants in Experiment 1A on the other hand, either failed to show selective influence, or yielded SIC curves that were indicative of parallel self-terminating processing or coactive processing, or even serial-self terminating processing.

Subject	df1	df2	F	p	MIC
1	1	191	199.6	< .001	21.940
2	1	192	4.592	< .05	17.049
3	1	195	7.520	< .05	20.618
4	1	195	1.768	.185	14.162
5	1	195	7.670	< .01	10.00

Table 2. This table indicates the level of audio and visual channel interaction for each subject. The mean interaction contrast is also indicated.

Each of the five subjects demonstrated selective influence. SIC curves are shown for each of the five subjects in Figure 4 below. Each participant completed 200 trials in each factorial condition—the same amount of trials that were completed in Experiment 1A. Errors and outliers (+ or – 3.0 SD from the mean) were eliminated from the analysis. Table 2 displays the F values and MIC for each of the five subjects.

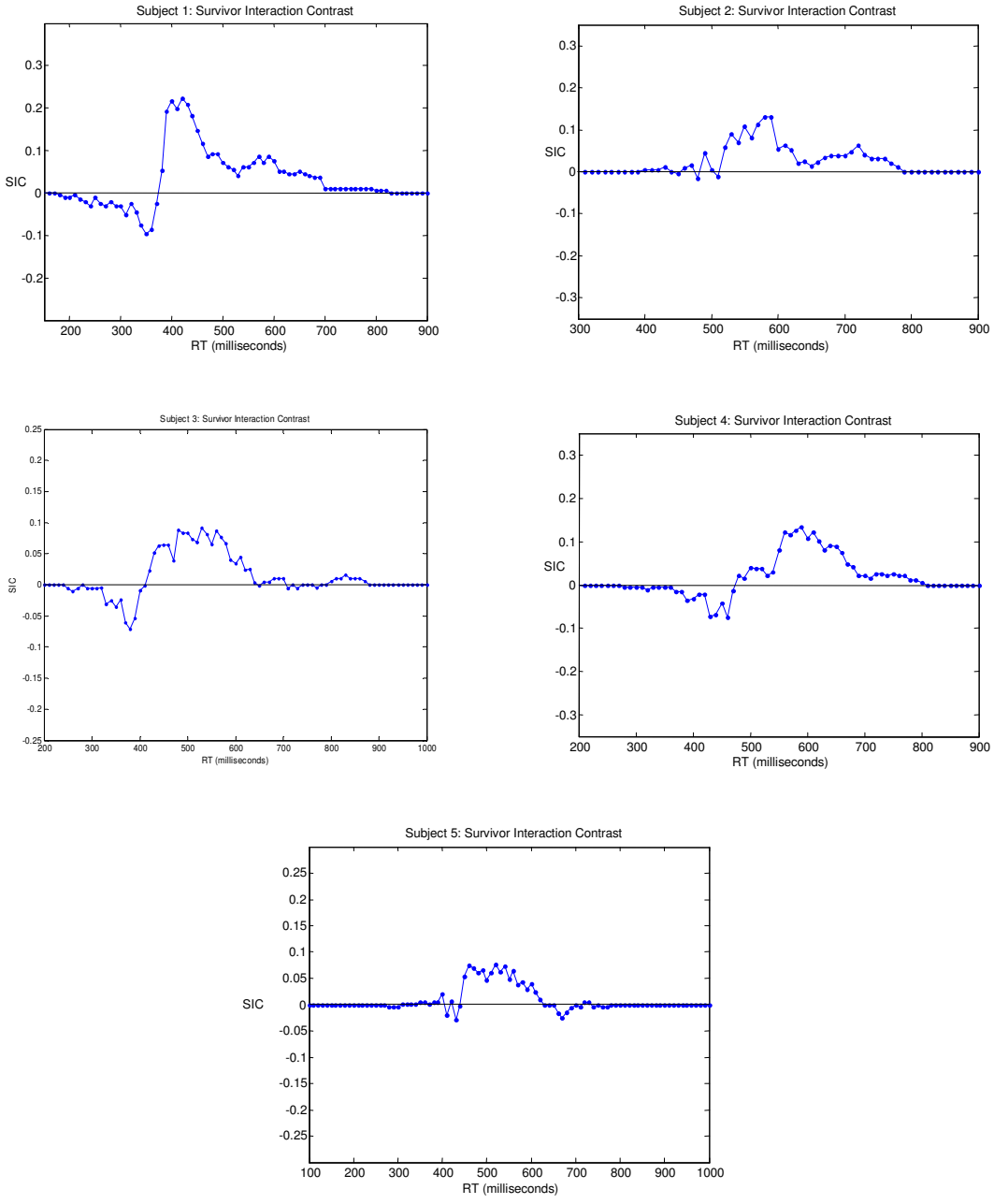


Figure 4. SIC curves for all five subjects in Experiment 1B. Each subject showed selective influence.

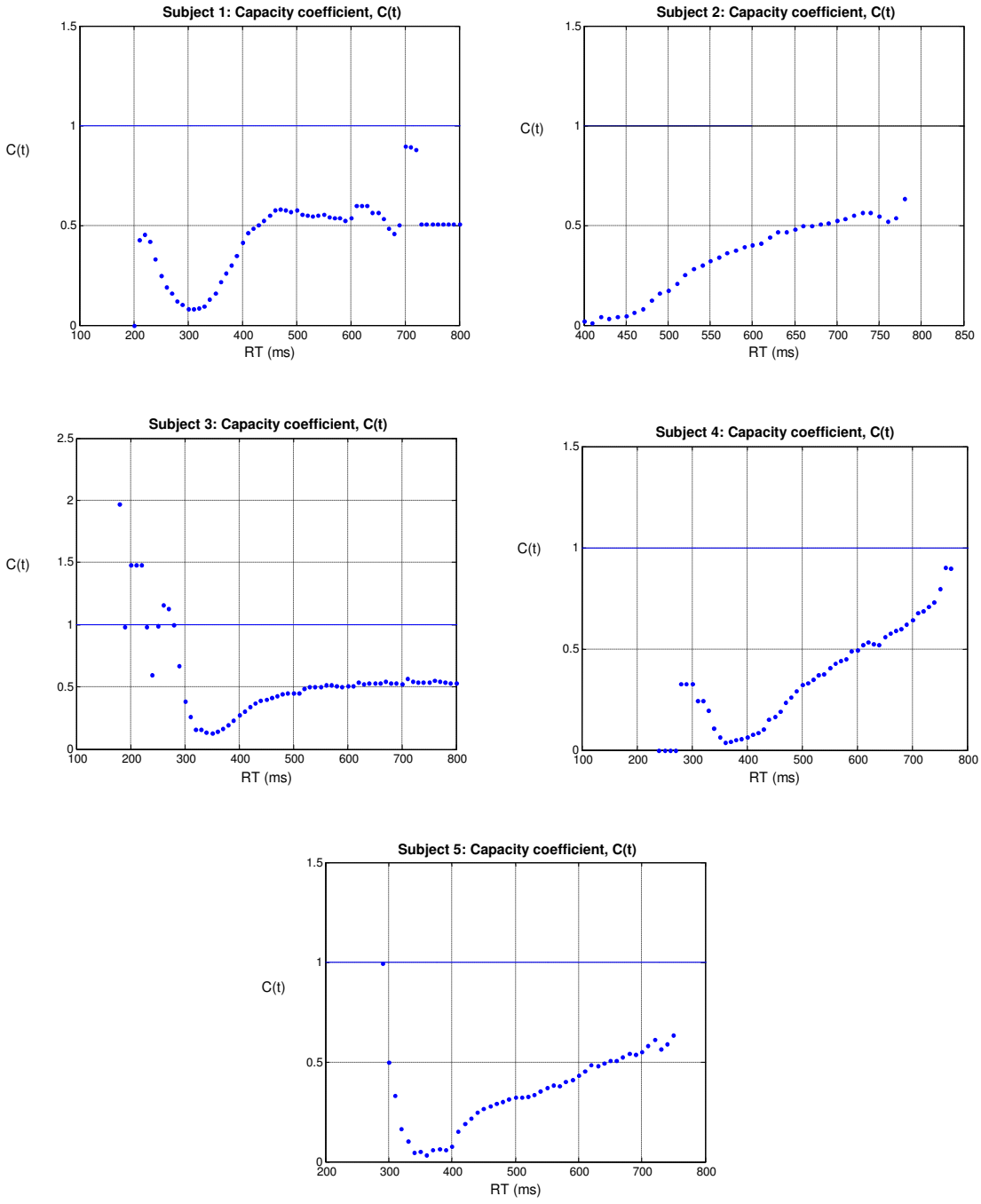


Figure 5. The Capacity coefficient for each of the five participants in Experiment 1B. Processing capacity was extremely limited for each subject.

The SIC for each subject was over-additive (> 0), strongly suggesting parallel or coactive processing strategies. The SIC and MIC for subjects 2 and 5 was almost entirely over-additive, indicating parallel self-terminating processing. The SIC for subjects 1, 3, and 4 show negativity for early stages of processing. The MIC was positive for subjects 1 and 4, although the ANOVA on the interaction was statistically significant for subject 1 but not 4. The positive MIC supports the hypothesis that processing

was coactive for these subjects, but the case is weaker for subject 4 whose F value was not statistically significant. Negativity at early processing stages is indicative of coactive processing, while positive SIC functions as previously discussed indicate parallel self-terminating processing.

The capacity functions for each subject shown in Figure 5 differ slightly from those obtained in Experiment 1A. The capacity coefficient C(t) for each subject was below 1 indicating fixed or limited capacity. Data from each participant shows that the Grice inequality was violated across many points in time. The capacity data differ slightly from the data in Experiment 1A because Miller's inequality was violated in Subject 3's and Subject 5's data. In short, while largely consistent with the data obtained in Experiment 1A, capacity, at least for some subjects, was not as limited at early processing times.

Experiment 2

Experiment 2 was designed to test architecture and capacity in a speech recognition task where participants have to distinguish between two words. Experiments 1A and 1B were control tasks where participants engaged in the detection, but not recognition, of audiovisual stimuli.

Participants

Five female subjects with normal or corrected vision were paid $10/session for their participation. Data from one subject was removed since that individual did not complete all experimental sessions.

Materials

The stimulus materials included two audiovisual movie clips of a female talker from the Hoosier Multi-talker Database saying the words "Base" and "Face". The two words in this set are intended to be confusable, with only the onset phoneme (/b/ versus /f/) differing between them. A total of eight different stimuli were created from each video clip: two audio files at two levels of saliency, two video files at two levels of saliency, and four audiovisual clips at each factorial combination of high-high, high-low, low-high, and low-low levels of saliency. The audio, visual, and audiovisual files were created using Final Cut Pro HD version 4.5. The audio files were sampled at a rate of 48 kHz using 16 bit encoding. Pink noise was generated using Adobe Audition and mixed into each audio file to create two different signal-to-noise ratios, and hence two different levels of saliency. The two signal-to-noise ratios for both stimuli were 40 dB for the high condition and 0 dB for the low condition.

The brightness level on the video files was manipulated in the same way as in Experiment 1A and 1B. The audio and video files lasted for a total duration of 1,616 milliseconds for "Base" and 1,683 milliseconds for the word "Face". The beginning of each audio and video file was edited in Final Cut Pro in order to create identical onset times for the spoken stimuli.

Design and Procedure

Subjects were seated in front of a Macintosh computer equipped with *Beyer Dynamic-100* headphones. Each trial began with a plus sign appearing in the center of the computer screen followed by the word "base" or "face." Trials are either audio alone, visual alone, or AV. Subjects were instructed to respond, as quickly and accurately as possible by pressing the button labeled "Base" if they either heard the word "base", saw a video of the talker saying "base", or both. Subjects were instructed to press the button labeled "Face" if they heard the word "face", saw a video of the talker saying "face", or both. There was a 1,000 millisecond delay between trials.

Each subject was presented with 3,360 total trials with 1,120 audio only trials (560 "base" + 560 "face), 1,120 visual only trials, (560 "base" + 560 "face"), and 1,120 audiovisual trials (560 "base" + 560 "face). Additionally, there were 280 trials in each redundant target condition (hh, hl, lh, ll). Table 3 below shows a diagram of the experimental design. Participants were run for 28 blocks at 120 trials each with a break scheduled between each block and each experimental session lasted approximately one hour. The experiment required four to five days to complete. Participants also received sixteen practice trials at the onset of each experimental session that were not included in the subsequent data analysis.

Audio	Visual	Correct Response
A_{Base}	V_{Base}	Base
A_{Base}	∅	Base
∅	V_{Base}	Base
A_{Face}	V_{Face}	Face
A_{Face}	∅	Face
∅	V_{Face}	Face

Table 3. This table shows each stimulus-response category (Base and Face) along side each factorial condition.

Results and Discussion

Percentage of errors averaged across all participants was less than 10 %. The error rate was likely higher in Experiment 2 due to the increased complexity of the task requiring subjects to distinguish between two similar spoken words of English. Each subject was close to or above 90 % accuracy across conditions. Evidence of a speed-accuracy trade off was not observed in the redundant target condition.

As in Experiments 1A and 1B, the initial analysis consisted of an investigation of the SIC curves and corresponding ANOVAs. Each subject showed selective influence. SIC curves for 4 subjects in Experiment 2 appear in Figure 6. ANOVA results and the MIC are shown in Table 4. The different nature of the tasks in Experiments 1 and 2 was the reason that subjects failed to show selective influence (or showed weaker selective influence) in the former experiment but not the latter. Although the duration of the stimuli remained the same between Experiments 1A and 2, participants were required to simply detect the presence of a moving image or sound in the first experiment, whereas in Experiment 2, the task was more likely to be cortically driven requiring them to distinguish between two words. More evidence was able to accumulate in each channel in Experiment 1 because the stimuli remained on for a longer time (compared with shorter durations in Experiment 1B), resulting in a smaller difference in completion times between the high and low conditions.

Recall that each participant in Experiment 2 completed 28 blocks consisting of 120 trials. Overall, the data demonstrate consistent processing between subjects. Each subject's SIC curve is over additive (greater than 0) at most points. Furthermore, each subject's MIC is positive and the corresponding one-way ANOVA indicates either a strong trend, or a significant positive interaction between the audio and visual channels. The positive SIC curve with the MIC and ANOVA results indicate parallel processing while observing a minimum time or self-terminating stopping rule.

Subject	df1	df2	F	p	MIC
1	1	263	3.34	.12	38.0
2	1	261	2.99	~ .05	21.4
3	1	261	3.87	< .05	36.7
4	1	201	.71	< .50	22.1

Table 4. This table shows the F value for the mean interaction, the p value (sig. = .05), and the mean interaction contrast for Experiment 2.

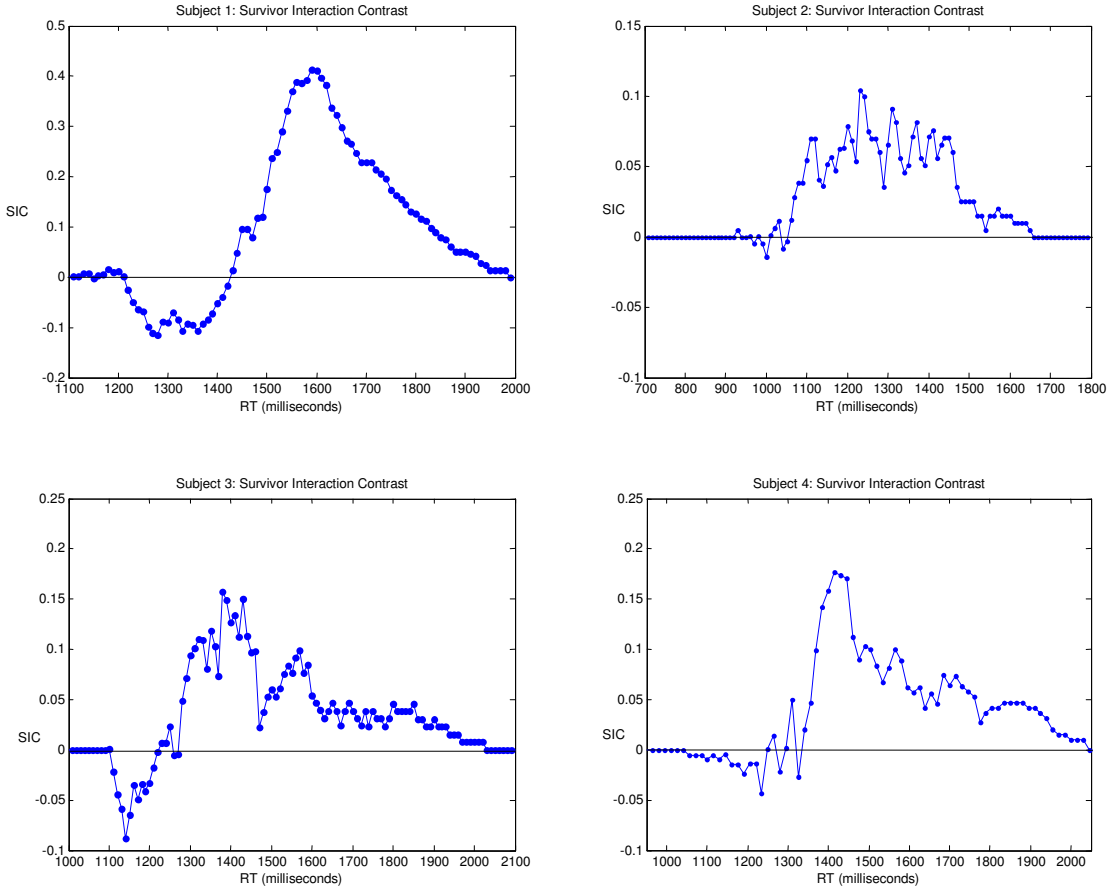


Figure 6. SIC curves for each the four subjects in Experiment 2. Each subject showed selective influence.

Subject 1's SIC curve and ANOVA suggests coactive processing in both conditions. The curve is mostly over additive with a degree of negativity for reaction times around 1300 milliseconds. Secondly, the mean interaction is positive (MIC = 38), although the p value from the ANOVA indicates only a trend toward a significant positive interaction ($p \approx .10$).

Subject 2's SIC curve was entirely over additive, strongly suggesting parallel self-terminating processing. Secondly, this participant's MIC was positive and the ANOVA indicates a marginally significant interaction between the audio and visual channels.

Subject 3's results indicted coactive processing, both in the over additive SIC curve showing negativity for early processing times, and the positive MIC. Results from the ANOVA also indicated a significant interaction with $p < .05$.

The negativity for early processing times in Subject 4's data also suggests coactive processing. The MIC was positive for this subject adding further evidence for coactive architecture rather than serial exhaustive. We did notice that the *F* value was low and the *p* value did not approach significance suggesting that the power was too low for this particular subject to draw strong conclusions. Nonetheless, these data are consistent and strikingly suggest coactive or parallel processing since the SIC curves and interactions were overwhelmingly positive, with a range of negativity for early processing times for three out of four subjects.

Figure 7 displays the capacity C(t) plots for all three subjects. The solid flat line at C(t) = 1 represents the bound for super capacity. The plots were consistent in showing that capacity at all points in time was extremely limited. The Grice Lower bound was violated for each subject at nearly every point in time, while the bound for super capacity was not surpassed at any point in time. These data, as in Experiments 1A and 1B, are indicative of an extremely limited capacity or fixed capacity coactive model. In order for a coactive model to predict extremely limited capacity, strong inhibition between auditory and visual channels is necessary. Independent coactive models always produce violations of the unlimited—supercapacity bound and do not violate Grice's inequality for extremely limited capacity (see Townsend & Nozawa, 1995).

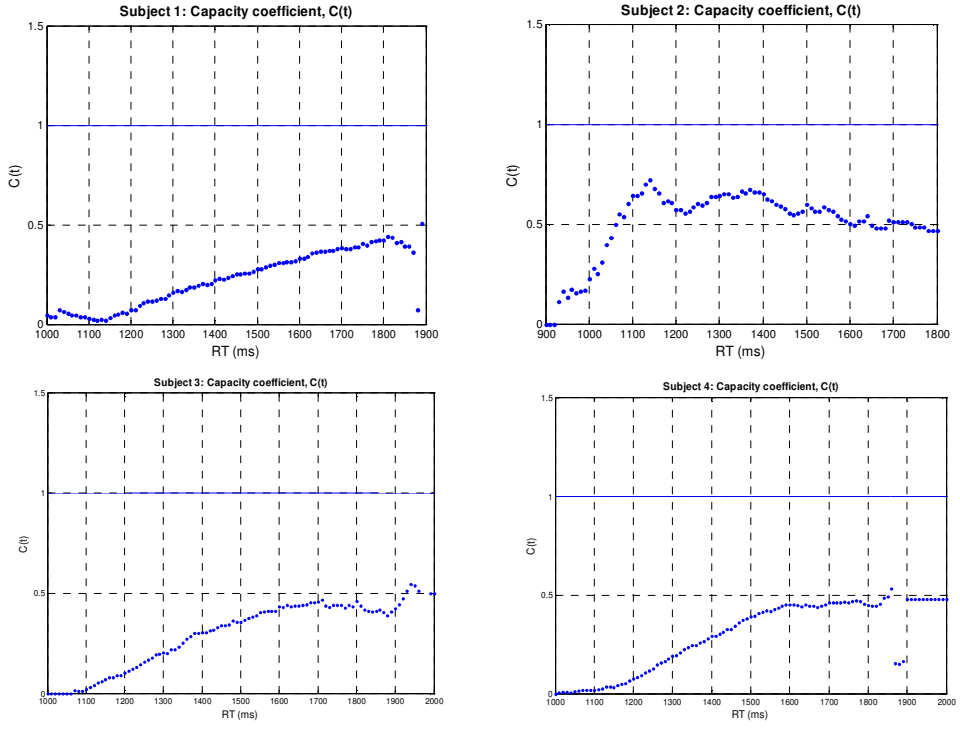


Figure 7. The Capacity coefficient for each of the four participants in Experiment 2. Processing capacity was extremely limited for each subject, as was the case in Experiments 1A and 1B.

General Discussion and Conclusion

Experiments (1A & 1B) and 2 were designed to test different models regarding how listeners process audiovisual stimuli in real time. Both versions of Experiment 1 were designed to test how participants processed audiovisual speech stimuli in a detection task, while Experiment 2 examined how participants process audiovisual speech stimuli in a word discrimination task where they were required to distinguish between real spoken words. Recall that previous work on audiovisual speech perception was generally unconcerned with dynamic models of audiovisual perception, and primarily sought to account for accuracy data (Braida, 1991; Massaro, 2004). In particular, previous research did not attempt to account for how information from the auditory and visual channels was utilized by the "black box" prior to or during the decision process.

Models of dynamic audiovisual speech perception are relevant to current work in the field. Theorists of direct perception and motor theory (Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985), and contrasting theories (Bernstein, 2005) make different claims about how audiovisual information is used during perception and word recognition. Mathematical tools founded upon factorial methodology which make specific claims about reaction time distributions, are an appropriate tools to begin investigating these claims. Factorial methodology was employed to assess the processing architecture and capacity in a detection task and word discrimination task. Our primary focus was analyzing the SIC and determining what form of processing emerged. In addition to investigating the shape of the SIC curves, we looked at the capacity coefficient to determine whether processing time increased, decreased or remained the same when two channels were present relative to the cases when only one channel was present. Experiments 1 and 2 began to reveal how the audio and visual channels are integrated. Data show that the main candidates for processing architecture are parallel with a self-terminating decision rule, or possibly coactive with extreme capacity limitations.

Data from the detection task in Experiment 1A revealed inconsistent results. SIC curves were either inconclusive as to the nature of processing taking place due to the fact that selective influence between conditions was either weak or not present. Processing appeared to be parallel self-terminating, while one subject showed coactivation and the rest demonstrated either serial or indeterminate processing. Experiment 1B, a modified version of Experiment 1A with shorter stimulus durations produced clearer results. Processing for each subject was most likely parallel self-terminating for 2 subjects, while 3 participants showed architecture consistent with coactive processing. Capacity between these two experiments was consistent, where the capacity coefficient C(t) was overwhelmingly negative for each of the subjects.

Data from the word discrimination task in Experiment 2 showed that processing was either coactive or parallel self-terminating. Capacity coefficients obtained in Experiment 2 revealed extremely limited capacity, which was consistent with the capacity measured in Experiment 1A and 1B. Extremely limited capacity is observed in serial models, and parallel models with negative inhibition, but is not typical of coactive models (Townsend & Nozawa, 1995; Townsend & Wenger, 2004). Hence it is important to begin understanding why coactive architecture indicated by the negativity in the SIC(t) function was observed in conjunction with extremely limited capacity in Experiments 1B and 2. The fact that capacity was extremely limited might indicate strong inhibition or competition between the audio and visual channels. Inhibitory links between channels might begin to explain why extremely limited capacity was observed in conjunction with coactive processing. However, simulations have demonstrated that coactive processing models are usually super capacity even with negative inhibition between channels (Townsend & Nozawa, 1995; Townsend & Wenger, 2004).

It is worth mentioning that extremely limited capacity was observed even though previous studies have consistently observed "audiovisual enhancement" in accuracy scores when audiovisual conditions were compared to audio only conditions (Sumby & Pollack, 1954). Audio and visual processing channels might simultaneously engage in inhibition (slowing the system down) while enhancing the quality of the information at the decision stage.

It is also important to continue investigating the nature of the limitations in processing capacity obtained in these experiments. If limitations in audiovisual processing capacity result from between channel inhibition, it would be worthwhile to understand how this inhibition might be manipulated or offset. Recent research involving discrimination of the numerals "1" and "2" in the visual modality with congruent speech stimuli indicates that manipulating the SOA (the lead of the visual stimuli in milliseconds) might decrease capacity limitations. At SOAs of 150 milliseconds or more, "redundant target effects" (i.e., supercapacity) were observed, which might indicate coactive processing (Berryhill et al., 2007).

Another worthwhile future direction will be to explore capacity and processing architecture using incongruent audiovisual stimuli as in the McGurk effect. The use of incongruent audiovisual stimuli will allow investigators to explore how audiovisual inhibition as indicated by the capacity coefficient C(t) might be enhanced or otherwise altered, and explore whether processing architecture remains consistent with cases where the audio visual are congruent in both AND as well as OR experimental designs.

References

Bernstein, L.E., (2005). Phonetic perception by the speech perceiving brain. In D.B. Pisoni & R.E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 79-98). Malden, MA: Blackwell Publishing.

Bernstein, L.E., Auer, E.T., & Moore, J.K. (2004). Audiovisual speech binding: Convergence or association? In G.A. Calvert, C. Spence & B.E. Stein (Eds.), *Handbook of Multisensory Processing* (pp. 203-223). Cambridge, MA: MIT Press.

Berryhill, M., Kveraga, K., Webb, L., & Hughes, H. C. (2007). Multimodal access to verbal name codes. *Perception & Psychophysics, 69*, 628-640.

Braida, L.D. (1991). Crossmodal Integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology, 43A*, 647-677.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C.R., McGuire, P.K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*, *593-596*.

Calvert, G.A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15*, 57-70.

Calvert, G.A., Campbell, R., & Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649-657.

Fowler, C.A., & Dekle, D.J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 816-828.

Fowler, C.A., & Rosenblum, L.D. (1991). Perception of the phonetic gesture. In I.G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory of Speech Perception* (pp. 33-59). Hillsdale, NJ: Lawrence Erlbaum.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding. *Journal of the Acoustical Society of America, 112, 30-33.*

Grant, K.W., Tufts, J.B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing impaired individuals. *Journal of the Acoustical Society of America, 121*, 1164-1176.

Green, K.P., & Miller, J.L. (1985). On the role of visual rate information in phonetic perception. *Perception and Psychophysics, 38*, 269-276.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception. *Cognition, 21*, 1-36.

Massaro, D.W. (1987). Speech perception by ear and eye. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D.W. (2004). From multisensory integration to talking heads and language learning. In G.A. Calvert, C. Spence & B.E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 153-176). Cambridge, MA: The MIT Press.

McGurk, H., & McDonald, J.W. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Meredith, M.A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research, 14*, 31-40.

Miller, J., Kuhlwein, E., & Ulrich, R. (2004). Effects of redundant visual stimuli on temporal order judgments. *Perception & Psychophysics, 66*, 563-573.

Puce, A., Allison, T., Bentin, S., Gore, J.C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18*, 2188-2199.

Rosenblum, L.D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78), Malden, MA: Blackwell Publishing.

Schroter, H., Ulrich, R., & Miller, J. (2007). Effects of redundant auditory stimuli on reaction time. *Psychonomic Bulletin & Review, 14*, 39-44.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica, 30*, 276-315.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26, 212-215*.

Townsend, J.T., & Fific, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Perception and Psychophysics, 66*, 953-962.

Townsend, J.T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39*, 321-359.

Townsend, J.T., & Wenger, M.J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111*, 1003-1035.

Wenger, M.J., & Townsend, J.T. (2000). Basic response time tools for studying general processing capacity in attention, perception, and cognition. *The Journal of General Psychology, 127*, 67-99.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Frequency of Use Leads to Automaticity of Production: Evidence from Repair in Conversation [1]

Vsevolod Kapatsinski

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] I would like to thank the NIH for financial support through Training Grant DC-00012 and Research Grant DC-00111 and to Adam Albright, Tessa Bent, Joan Bybee, Dan Jurafsky and the audiences at HDLS VII and the Workshop on Gradience and Frequency Effects in Phonology for helpful comments.

Frequency of Use Leads to Automaticity of Production: Evidence from Repair in Conversation

Abstract: Investigation of spontaneous replacement repairs found in the Switchboard Corpus (Godfrey et al., 1992) shows that low-frequency repaired words are more likely to be interrupted prior to replacement than high-frequency words are. These results provide novel empirical support to the hypothesis that the production of high-frequency words is more automatic than the production of low-frequency words (Bybee, 2002; Logan, 1982). The relationship between the effects of frequency on interruptibility is argued to be partially mediated by the effect of frequency on duration. In addition to testing the link between frequency and automaticity, the present paper reports that replaced words tend to be more frequent than the words that replace them, providing support for the hypothesis that high-frequency words are easier to access in word production, which has been criticized on the basis of not observing this frequency asymmetry in semantic substitution errors (Garrett, 2001). Finally, whether a word is interrupted is found to depend strongly on the length of the word, with long to-be-replaced words being more likely to be interrupted than produced completely. Thus, while speakers prefer to produce constituents with a continuous delivery (Clark & Wasow, 1998), the drive to produce a continuous constituent competes with the drive to interrupt as soon as possible (Main Interruption Rule, Levelt, 1983, 1989).

Introduction

Theoretical Background

Bybee (2002) suggests that the production of high-frequency words and phrases is more automated than the production of low-frequency words and phrases. Under this hypothesis, high-frequency words are more cohesive than low-frequency words: the parts forming a high-frequency word are more tightly linked together than the parts forming a low-frequency word.

Previous evidence for a link between cohesion and frequency has come from studies showing that high-frequency words are more likely to undergo reductive sound change (Bybee, 2002; Hooper, 1976). Mowrey and Pagliuca (1995; Pagliuca & Mowrey, 1987) go as far as claiming that all internally-motivated regular sound changes in progress that have been attested can be explained by an increase in gestural compression. Bybee (2001: 79-83) and Phillips (2001) suggest that there are other sources of sound change but that Mowrey and Pagliuca's claim holds for sound changes that involve lexical diffusion from high-frequency to low-frequency words.

An increase in the temporal overlap between successive gestures and temporal compression of the sequence of articulatory goals corresponding to a word is expected to result from automatization of word production (Bybee, 2002). Assuming that in a sequence of articulatory goals, a goal gains control of articulation when it is activated sufficiently, and that activation spreads from earlier goals to later ones, a goal will receive control of articulation earlier when it is strongly connected to the preceding goal. Thus, the preceding goal is less likely to be completely reached when the following goal is highly predictable in the context. In addition, when the gestures called for by successive goals do not interfere with each other, which could cause undershoot, articulatory overlap between gestures implementing successive goals is more likely in a high-frequency sequence. Under this account, a high-frequency word is a more cohesive unit than a low-frequency word.

However, the finding that reductive sound changes start in high-frequency words has also been interpreted as indicating that speakers do not expend as much articulatory effort in such words because of their high contextual predictability for the listener (e.g., Bybee 2002: 269; Gregory et al., 2000; Lindblom, 1990). Fowler (1988) shows that words that have already been mentioned in the course of the conversation are shorter than words that are mentioned for the first time (see also Fowler and Housum, 1987) but only if the two tokens are co-referential. Words are not shortened if a homonym has recently been pronounced but are shortened if preceded by a synonym. Fowler (1988: 317) writes that "production of a homophone of a target… is not sufficient to yield shortening… even though the word's articulatory routine has recently been used. Apparently the shortening reflects the talker's estimate that a listener has other information available to help identify the word". Gregory et al. (2000) support this interpretation by showing that semantic relatedness to the discourse topic influences word duration even when repetition is controlled: words related to the discourse topic are shorter than unrelated words.

Under this alternative interpretation, word frequency does not directly influence gestural compression, automaticity of production, or word cohesion. Rather, frequency is simply one of the factors that influences contextual predictability, which serves as a constraint on how much reduction the speaker thinks s/he can get away with.

In the present paper, we investigate a hitherto untested prediction of the hypothesis that the production of high-frequency words is more automatic than the production of low-frequency words. As Anderson (2000: 99) puts it, "automaticity occurs when practice eliminates most of the need for central cognition", which leads to the behavior becoming relatively impervious to cognitive influences. In particular, the more automatic a behavior, the harder it should be to interrupt. Thus, if the production of a high-frequency word is more automatic than the production of a low-frequency word, the production of a high-frequency word should be harder to interrupt than the production of a low-frequency word.

To address this issue, we will analyze a corpus of conversations among native English speakers (Switchboard, Godfrey et al., 1992), which has been tagged for disfluencies. The working hypothesis is that when the speaker interrupts his/her production to replace the word s/he has just produced or started producing, the interruption is more likely to be delayed until the end of the to-be-repeated or to-be-replaced word if the word is frequent than if it is rare.

The Phenomenon

In a replacement repair, the speaker replaces the word s/he has just produced or started producing by a different word. Examples of replacement repairs from the Switchboard Corpus are shown in (1)-(4). The **replaced word** is shown in bold while the **replacement** is italicized. We will call the observed part of the replaced word, e.g., *wa* in (3), the **remainder**, reserving the term **replaced word** for the inferred complete lexical item, e.g., *watch* in (3). Examples in (1)-(4) show that the speaker has a choice of producing the replaced word completely or interrupting its production. The present paper is restricted to cases of replacement repair in which the replaced word and the replacement word are semantically related because it is nearly impossible to guess the identity of an interrupted replaced word if it is not semantically related to the replacement.

(1) It was **pathe-**, I mean, it was *horrible*.
(2) That's why we were surprised to see 'Toyota' **written**, I mean, *imprinted* on the engine
(3) I will intentionally buy newspaper to **wa-**, to *look at* the news.
(4) They don't want to become a state for fear of losing **Spanish**, uh, *Hispanic* heritage.

Cohesion as an Influence on Disfluency Location

While there have been no studies of frequency effects in replacement repair, previous work on repetition repair and other disfluencies has shown that the location of interruption and how much material is repeated are influenced by constituency. Boomer (1965), Clark and Wasow (1998), Levelt (1983), and Maclay and Osgood (1959) found that interruption of speech production is more likely to occur at word boundaries than within words and between major syntactic constituents, such as subject and object, rather than within them. Beattie and Butterworth (1979), Goldman-Eisler (1958, 1968), Tannenbaum et al. (1965) and Cook (1969) demonstrated that hesitations tend to occur in between-word transitions of maximum uncertainty, as indicated by low transitional or Cloze probability. These results suggest that interruption is sensitive to cohesion: speech production is more likely to be interrupted at the boundary between cohesive units than within a cohesive unit. Thus, if high-frequency words are more cohesive than low-frequency words, speakers should be less likely to interrupt speech production in the middle of a high-frequency word than in the middle of a low-frequency word.

Several studies found that speakers tend to start repetition from the nearest major constituent boundary (Maclay & Osgood, 1959; DuBois, 1974; Nooteboom, 1980; Levelt, 1983; Fox & Jasperson, 1995; Clark & Wasow, 1998; Kapatsinski, 2005). Definitions of major constituent boundaries differ somewhat across studies, with most researchers taking such boundaries to include clause, object, and oblique boundaries (Clark and Wasow, 1998; Fox and Jasperson, 1995; Kapatsinski, 2005; Maclay and Osgood, 1959).[2] Based on this work, Clark and Wasow (1998: 206) proposed the **Continuity Hypothesis**, which states that speakers prefer to produce syntactic constituents with a continuous delivery. For instance, if speech production is interrupted somewhere in a prepositional phrase, speakers tend to repeat everything they have produced after starting the phrase as in (5) below.

(5) I was really familiar **with a lot**, **with a lot of, of** the AOR type music

In (5), the speaker repeats three words s/he has already produced despite an overall preference to repeat as little as possible (in the sample of Kapatsinski 2005, 79% of repetitions are one-word repetitions, 18% are two-word repetitions, and only 3% are three-word repetitions). The likely reason, according to the Continuity Hypothesis, is that the speaker wants to produce the entire prepositional phrase without interruption.[3] Importantly, while English speakers often repeat prepositions, Japanese speakers do not repeat postpositions, which would involve restarting speech from the middle of a postpositional phrase (Fox et al., 1996). Finally, the Continuity Hypothesis is supported by the fact that if word production is interrupted within a word, the speaker almost always restarts the word, rather than continuing from the point of interruption.[4]

Kapatsinski (2005) found that how much is repeated in a repair is influenced by between-word transitional probability. Speakers do not start repeating from the nearest constituent boundary if that constituent boundary is a high-probability transition. Kapatsinski tried to predict how many words will be involved in each repetition found in the Switchboard corpus depending on the location of the nearest constituent boundary and on which of the three nearest between-word transitions has the lowest transitional probability. The location of the nearest constituent boundary correctly predicted 44% of the three-word repetitions in the Switchboard corpus. Then transitional probability was added as a predictor. The two predictors were combined so that if transitional probability at some nearby word boundary is much lower than at the nearest constituent boundary, subjects were predicted to start repeating from the

[2] The status of the subject-verb boundary is questionable (Fox and Jasperson, 1995). In addition, Levelt (1983) argues for an alternative criterion for where disfluencies should occur, according to which one should be able to continue the constituent interrupted by the disfluency in such a way that it would be conjoinable with the constituent following the disfluency.
[3] Alternatively, speakers may have difficulty initiating production from the middle of a cohesive unit.
[4] I have been able to find only one example of the latter on Switchboard.

transition with the lowest transition probability. Otherwise, they were predicted to start from the nearest constituent boundary. This modification of the Continuity hypothesis improved the predictability of three-word repetitions to 57%, while maintaining 70% accuracy on one-word and two-word repetitions, where chance performance is 33% (Kapatsinski, 2005:490).

Thus prior findings suggest that between-word cohesion influences location of interruption and speakers tend to restart interrupted cohesive units, whether the cohesion is caused by syntactic constituency or probability of co-occurrence. In the present study we will examine whether location of the interruption is also sensitive to within-word cohesion and, more specifically, whether high-frequency words are more cohesive than low-frequency words.

The Possible Roles of Relative Frequency

Surprisingly, there has been only one study looking at word frequency as an influence on disfluency location. Biber et al. (1999:1059) observe that the indefinite article is less prone to being repeated than the definite article and propose that "perhaps, all other things being equal, the higher a word's frequency, the more likely it is to form repeats… It is easy for the speaker to utter a very frequent word, without having a clear plan of what words will follow it. Hence, such a word precedes a natural hesitation point in the utterance". Biber et al. support the hypothesis by pointing out that *an* is repeated very rarely since before choosing *an* the speaker must at least decide on a vowel-initial word to follow it. Otherwise, the speaker would choose the much more frequent variant *a*. Consequently, the sequence *a an* is much more frequent than the sequence *an a*. In addition, the authors find that frequent subject+verb contractions, those that involve 's, 're, 'm and 'll, are more likely to be repeated, per number of tokens of the contraction in the corpus, than less frequent contractions involving 've and 'd (Biber et al., 1999: 1061-2).

Biber et al.'s (1999) hypothesis provides a possible prediction for when a to-be-replaced word's production will be interrupted. The hypothesis is that a high-frequency word is likely to come to mind faster than a low-frequency word. Thus, if the replaced word is frequent and the replacement word is rare, the replaced word will come to mind long before the replacement word. Thus, the speaker will have enough time to produce the replaced word in its entirety before s/he becomes aware of the more appropriate alternative. On the other hand, if the replacement word is frequent relative to the replaced word, the appropriate replacement is likely to come to mind soon after the speaker starts to utter the less appropriate word, leading the production of the replaced word to be aborted before the entire word is produced. This theory predicts that, other things being equal, interrupted words should be replaced by high-frequency words while uninterrupted words should be replaced by low-frequency words. Thus in the present study, we examine both frequency of the replaced word and frequency of the replacement word as predictors of whether or not the replaced word is interrupted.

In addition, if a high-frequency word comes to mind faster than a low-frequency word, the case in which a frequent inappropriate word is replaced by a rare but more appropriate word should be more common than the case in which a rare word is replaced by a word that is both more appropriate and more frequent. Thus, the replaced word should tend to be more frequent than the replacement word. However, studies of semantic substitution errors have failed to find a difference between the erroneous word ('the intrusion') and the correct target (DelViso et al., 1991; Harley & MacAndrew, 2001; Hotopf, 1980; Silverberg, 1998). Garrett (2001) notes that this negative result is inconsistent with existing models of word production as well as experimental data from picture naming, which show that pictures with high-frequency names are faster than pictures with low-frequency names (e.g., Jescheniak & Levelt, 1994;

Oldfield & Wingfield, 1965).[5] In contrast to studies of semantic substitution errors, the present sample shows a small but reliable difference in frequency between replaced words and replacement words in the expected direction, closing the gap between naturalistic and experimental data observed by Garrett (2001) and supporting the role of token frequency in facilitating lexical access in production.

Main Interruption Rule and Error Detection

An assumption made by the model in the preceding section is that interruption is triggered by awareness of an alternative, rather than recognition of the inappropriateness of the word being produced. Alternatively, interruption could be triggered by detection of inappropriateness and the search for an alternative could be initiated by detection of inappropriateness. Under this hypothesis, the location of the interruption would be independent of how fast the alternative is accessed. Rather, a word would be likely to be interrupted if its inappropriateness is detected early relative to when the production of the word is initiated.

Levelt (1983, 1989) proposes that speakers interrupt production as soon as they detect inappropriateness of the word being produced (what he calls the **Main Interruption Rule**). Under the Main Interruption Rule, words may be interrupted if their inappropriateness is detected quickly. The speed of detection could plausibly depend on the severity of the error. Thus, a word that is merely inappropriate may be less likely to be interrupted than a word that is an outright speech error, as found by Levelt (1983). [6] There is some evidence that low-frequency words are more likely to be involved in speech errors (e.g., Harley and MacAndrew, 2001). If high-frequency words are less likely to be uttered in error and more likely to be merely inappropriate than low-frequency words, error detection may be slower in high-frequency words, making high-frequency words less likely to be interrupted than low-frequency words. We will return to this possibility in the analysis section.

Experimental Studies of Interruptibility in Language Production

There have been three previous studies that specifically examined how easy it is to interrupt language production and the factors influencing interruptibility (Ladefoged et al., 1973; Logan, 1982; Slevc & Ferreira, 2006). In all of these studies, on a small proportion of trials, the subject was presented with a stop signal, which indicated to the subject that they should stop production.

While none of these studies were specifically designed to test for frequency effects, Logan (1982, Experiment 3) observed that if the typists were told to stop typing immediately before they started typing the word ‘the’, they tended not to stop until after producing ‘the’, producing 2.72 letters on average. The same subjects produced fewer than 2 letters on average if the stop signal came before a content word (verb or noun). Logan showed that while the word ‘the’ was typed faster than other words, the time it took subjects to stop typing ‘the’ was longer than the time it took them to stop typing content words. He attributed the effect to word frequency, noting that ‘the’ is the most frequent word in English. The present study extends this finding by investigating a much larger range of words and word frequencies in naturalistic speech production.

[5] The lexical locus of this effect was confirmed by its disappearance in picture recognition (Jescheniak and Levelt 1994, experiment 2).

[6] Levelt himself (1989:481) seems to reject this possibility, writing “there is no reason to assume that the detection of error occurs more frequently within the troublesome word than the detection of inappropriateness”, suggesting instead that interruption is used by the speaker to tell the listener whether the replaced word is a speech error.

Method

The Corpus

For this study we collected all tokens of replacement repair in the Switchboard corpus (Godfrey et al., 1992) that satisfied our inclusion criteria. The Switchboard corpus is a collection of telephone conversations between native American English speakers on predetermined topics that are chosen by the participants from a fixed set of alternatives with no knowledge of the identity of their interlocutor-to-be. The version of Switchboard annotated for disfluencies contains about two million words. The corpus is annotated with a special symbol ('+') which marks the locations of repairs. Sound recordings of the conversations are available online from the LDC (https://online.ldc.upenn.edu/search/). To be included in the present sample, a token of repair had to be coded as one in the corpus. In addition, the author listened to the coded tokens of repair and excluded a number of cases based on the exclusion criteria outlined in the next section.

Exclusions

In the present paper, we concentrate on semantically motivated replacement repair. Thus instances of repair which involve word insertion as in (6) or (7), word deletion, or reordering as in (8), as opposed to replacement were excluded.

(6) It does give you a **good**, a *real good* workout.
(7) Just to see whether or not we're **falling**, you know, *getting ahead*, **falling** behind or staying even or what.
(8) They ought to, you know, go out of the way, I think, a little bit more to, to *help you get*, **help get you** rehabilitated

Since it is difficult to guess the identity of an interrupted replaced word when it is not semantically related to the replacement word, uninterrupted replaced words were excluded as well if they were not semantically related to the replacement. Thus, the example in (9) was excluded from the sample.

(9) I went to the bike **shock,** I mean, the bike *shop*.

The example in (9) would be excluded from the sample for another reason as well. In (9), the replaced word (*shock*) and the replacement word (*shop*) share beginnings. Therefore, if the replaced word were interrupted, it would be impossible to tell that the sentence involves replacement rather than repetition. Thus, all cases in which the replaced word and the replacement word share beginnings were excluded from the sample if they shared more than one segment. Repairs involving words shorter than three segments or longer than eight segments were excluded because there were very few such words in the sample.

In addition, instances of repair in which the replaced consisted of more than one word were excluded. These include cases of the type shown in (10), where *turned [out]* is abandoned in favor of *was*, as well as cases in which multiple words that are part of the replaced surface as in (11). Contractions like *can't* or *don't* and *going to* in the sense of *will* were considered single words and included in the sample.

(10) It **turned**, it *was* okay.
(11) The court systems need to be **more accurate** in, in, *stiffer* in their penalties.

Cases in which the replaced was a function word that was incompatible with what followed the replacement, as in (12) where *has* appears to be replaced by *is*, were also eliminated because it is likely

that in these cases repair is motivated by a desire to replace not the function word itself but some word downstream in the planning sequence or the syntactic construction itself (cf. Stemberger, 1984).

(12) **Has**, *is* this guy a convicted felon?

Uninterrupted replacement repairs included both cases in which the flow of speech was interrupted immediately after the replaced word and those in which it was interrupted later. Thus, cases like (13) were included in the sample.

(13) I haven't **had** a chance, I haven't *got* a chance to look at them yet.

Finally, there is a thin line between replacement repair and certain grammaticalized constructions, which should not be included into a sample of repairs because they disallow interruption. One such construction is the clarification construction in which the 'replacement' is a hyponym of the replaced. Thus one can argue that the example in (14) does not involve repair but rather clarification. However, example (15) in which the replaced word is interrupted, cannot be interpreted in this way. Thus, the speaker may prefer to say (14) instead of (15) regardless of the frequency of *same*. Thus, cases in which the replacement is a hyponym of the replaced were excluded from the sample.

(14) But, no, no real association with TI other than being in the **same** industry, the *electronics* industry.
(15) But, no, no real association with TI other than being in the **sa-**, the *electronics* industry.

The mirror image of the clarification construction illustrated in (14) is presented by subject topicalization in which the 'replaced' is a hyponym of the replaced. An example is presented in (16). To avoid inadvertently including such cases into the sample, all examples in which the replacement is a pronoun, the replaced is a noun phrase, and the two can be coreferential were excluded from the sample.

(16) **My husband and I**, *we* just sit there and cackle.

Another potentially grammaticalized case excluded from the sample is the use of interruption following subject+*just* followed by repetition of the same subject as in (17). Such cases are quite common, although more commonly *just* is either repeated or omitted and may involve an interruption that is preplanned for emphatic purposes rather than generated online when a decision to replace a word is made.

(17) He **just**… He *simply* doesn't care anymore.

Another case in which repair can be confused with a grammatical construction if the replaced word is not interrupted is when the 'replaced' and the 'replacement' are numbers and the second number is larger than the first (in terms of absolute value, as (19) shows). Thus, repairs involving numbers were included only if the second number was closer to zero than the first.

(18) It's taken them **ten**, *fifteen* minutes at a time.
(19) When it's minus **twenty-five**, minus *thirty* degrees…
(20) When you're **twenty**, *thirty* years old…
(21) He was there in nineteen eighty **four**, eighty *five*.

Finally, repairs are important to distinguish from lists. A specific problem is presented by lists of near-synonyms in which the following synonym is 'more intense' than the preceding one, e.g., *big giant trees* or (possibly) the example in (19). In these cases, the second word is not intended to replace the first

word, hence interruption is not an option. In addition, cases in which the speaker can't decide on the correct word and plans to indicate his lack of certainty by using a disjunction in advance are potentially problematic (a possible example is shown in (22).

(22) He's a computer **programmer**, or a computer *engineer*.

Fortunately, lists that do not have a conjunction usually have more than two elements and were excluded on the basis of this criterion. In addition, both listing constructions and disjunctions can be identified by intonation. The presence of emphasis on the replaced list intonation, or the absence of interruption was sufficient for exclusion. As Ladefoged et al. (1973) observed in their study of experimentally elicited interruption, interrupted words almost invariably end with a glottal stop or at least significant glottalization, while uninterrupted words do not. Thus, glottalization is a very reliable cue for whether the word was interrupted. Nonetheless, 31 tokens were excluded from the study because there was disagreement between the present author and the corpus coders on whether or not the word was interrupted or because the present author was not certain about the status of the word.

Exclusions Specific to Particular Analyses

While cases in which the replacement consisted of more than one word, as in (3) where *watch* is replaced by *look at* or (23) where *had* is replaced by *came out of*, were included in the complete sample, they were excluded for the purposes of comparing the frequency of the replaced word to the frequency of the replacement word both in terms of their absolute values and as predictors of interruption. Comparing word frequencies to a mix of word and phrase frequencies would be unfair because the frequency of a phrase is on average lower than the frequency of a word just because a phrase contains multiple words.

(23) I **had** a, I *came out of* a thirty-one hundred square foot two story house.

In order to assess whether low-frequency words are more likely to be uttered in error, rather than being merely inappropriate, we need to determine whether a given repair involves an error. Determining whether a repair involves an error in natural conversation is quite difficult and it is not clear that the distinction can be reliably made in all cases. Moreover, a large proportion of cases are similar to example (24) where what the speaker may consider a speech error, the listener, who does not know anything about the speaker's family, would surely not. Thus, the analysis will be restricted to unambiguous cases only.

(24) My **parents**, my *mother* is trying to let my grandmother stay in her house.

Levelt (1983:63), writes that "in an appropriateness repair… the reparandum is correct but needs some qualification". This suggests that a hyponymy relation is involved. Such cases were excluded from the present sample. In addition, the example shown in Levelt(1989:481) suggest that repairs of suboptimal choices can also involve synonymy as shown in (25).

(25) To the left of it a **blanc,** or a *white* crossing point.

While it is not clear whether this example would be included in the present sample because the replaced and the replacement are conjoined with *or* and it is not indicated that the example involved hesitation, the sample does include a number of cases in which the replaced and the replacement are synonymous, as in (26)-(27).

(26) I don't have the expertise to just hurry up and do it like **some**, *a* professional would.
(27) That's my **private**, you know, my *own* home.

These cases can be compared to tokens in which the replaced and the replacement are incompatible because of having demonstrably different referents as in (28)-(29). A common special case is the replacement of a quantifier by a quantifier with a different range of possible values as when *most* is replaced by *all*, *few* is replaced by *most*, *eleven* is replaced by *twenty-one*, *quite* being replaced by *not really*. On the other hand, cases in which the replaced and the replacement are quantifiers whose ranges of values are similar, as when *several* is replaced by *a few*, can be considered repairs not involving a speech error.

(28) A sixty-seven **Chev-**, uh, *Mustang*
(29) You may be able to take **care**, take *advantage* of that.

An additional class of repairs that can be said to involve speech errors are repairs in which the replaced word does not fit the preceding context as in (30)-(31).

(30) The person who is the line **own-**, the line *manager*.
(31) I was watching the **ra-**, the *TV* today.

Finally, repairs in which one form of a verb is replaced by a different form of the same verb, such as *is* being replaced by *was* can be considered repairs involving speech errors. This does not include cases like *was* being replaced by *has been* in which the two verb forms can have the same referent. Such cases were not included in the analysis.

Measuring Frequency, Duration, and Number of Segments

For each instance of repair included in the sample (N=1749), the duration of what remained of the replaced word (the remainder) and the duration of the replacement word were measured. In order to examine the extent to which any possible effects of frequency are mediated by the effect of frequency on duration (frequent words are shorter), I estimated the length the interrupted word would have if it were not interrupted. Several estimates were obtained. For each of these estimates, the duration of the word did not include the word-final segment. This is because one purpose for which we need estimates of word duration is to compare the durations of interrupted and uninterrupted words. Since a word may not be coded as interrupted if its final segment was perceived by the coders, the final segment is not a possible location for interruption, and the status of the preceding transition is questionable since it can contain strong cues to the final segment's identity.

First, a very crude estimate of word duration was obtained by multiplying the duration of the remainder by the ratio of the number of segments in the remainder to the number of segments in the complete word. Second, the duration of the complete word produced in isolation from text by an adult female native speaker of American English was obtained from LDC's American English Spoken Lexicon (http://www.ldc.upenn.edu/cgi-bin/aesl/aesl). Finally, ten samples of each of the replaced words were obtained from the Switchboard Corpus and their durations were measured. When ten tokens of the word were not available, all available tokens were used. When more than ten tokens were available, the ten samples used were randomly selected. This last measure proved to be the best of the duration measures in terms of predicting whether or not the word would be interrupted. Hence, the comparisons between frequency and duration as predictors of interruption reported below use this measure.

Durations were measured by hand in Praat. The principal difficulty in measuring duration came from cases in which the to-be-measured boundary fell between two stops or a stop and a pause. When a word began with a stop preceded by another unreleased stop or silence, the beginning of the word was taken to be the point at which the intensity track starts to increase sharply from the floor as shown in

Figure 1 for the word *trickles*. In the case of a stop-final word, the midpoint of the preceding segment was taken to be the end of the word.

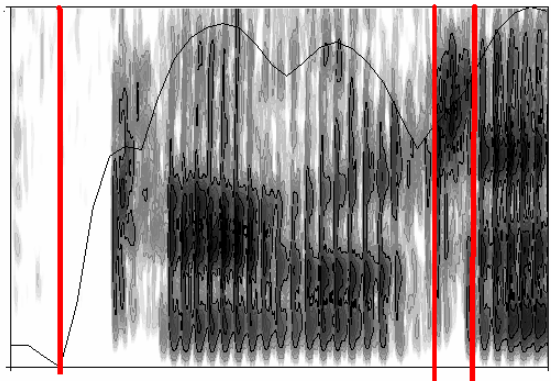


Figure 1. Measurement of duration of stop-initial words (spectrogram with a superimposed intensity track). Word boundaries are shown by the thick lines. The rightmost line shows the actual end of the word while the next rightmost line shows the end of the word as measured for the purposes of this study.

For the purposes of estimating word length in numbers of segments, affricates, diphthongs, syllabic nasals and liquids and /ɜr/ were coded as single segments. This decision was made because cases in which a diphthong was interrupted (e.g., [ha- haʊsɪz], [θɜrzdɛ- θɜrzdɛɪ]) and cases in which the schwa was produced without the following sonorant (e.g., [mɑðə- maðɚ], [ivə- ivən]) were exceedingly rare, and there were no cases in which an affricate was interrupted.

Word frequency was operationalized as frequency of occurrence within the Switchboard Corpus, the corpus under analysis in the present study. Since Switchboard consists of conversations on a limited range of topics, frequencies within the corpus may not correlate very well with frequencies elsewhere in the language. Since recent repetitions are likely to be more important for present behavior than earlier repetitions, and since the production of a word may be more automatic when it is related to the topic of conversation and therefore somewhat predictable, frequency within the corpus under analysis is still arguably a more appropriate measure than frequency within some other corpus (e.g., Francis and Kuçera 1982). Surface frequency rather than base frequency was used. That is, frequency was not aggregated across different inflectional forms of a particular word. This decision is based on a regression analysis of the effect of frequency on whether or not a to-be-repeated word is interrupted, which showed that surface frequency was a better predictor of interruption than base frequency.

For the purposes of analysis, frequency was logarithmically scaled since ease of lexical access in both perception (Howes & Solomon, 1951) and production (Oldfield & Wingfield, 1965) is correlated with log frequency better than with raw frequency. The basic idea behind the log transform is that the difference in frequency between a word that occurs only once in the corpus and a word that occurs ten times is much more psychologically significant than the difference between a word that occurs 1000 times in the corpus and one that occurs 1010 times. For 3-segment and 1-syllable words considered separately the distribution of log frequencies is skewed, violating the assumptions of standard statistical tests. For

this reason, frequencies were converted to ranks for the purposes of statistical tests involving the subsamples of three-segment and one-syllable words.[7]

Analysis

In this section, we first establish that interrupted words that are interrupted and words produced completely do in fact differ in token frequency, that longer words are more likely to be interrupted than shorter words, and that the frequency difference remains when number of segments is controlled. We then examine whether the speed of accessing the replacement can account for the results and establish that interrupted words do not tend to be replaced by high-frequency words while words produced completely are followed by low-frequency words. In the following section we confront the issue that the identity of the replaced word needed to be guessed and show that interrupted replaced words, which I guessed, are as frequent *relative to their replacements* as uninterrupted words, for which no guessing was involved. Therefore, my guesses are argued not to be biased in favor of the hypothesis in that, even if wrong, they tend to produce words that are as frequent as the interrupted words intended by the speakers. We then address the possibility that errors that occur in low-frequency words are more severe and thus easier to detect. Finally, we examine the interaction of frequency, interruptibility and duration, arguing that high frequency of use does not just shorten words but also makes interruption dispreferred.

Frequency and Number of Segments

Figure 2 shows that the longer the replaced word, in terms of number of segments, the more likely it is to be interrupted. The relationship between number of segments and likelihood of interruption is well approximated by a logarithmic curve. Figure 2 also shows that words longer than four segments are more likely to be interrupted than produced completely. In addition, it should be born in mind that replacements involving only one segment and replacements in which the identity of the replaced word could not be guessed are not included in the sample. As a result, Figure 2 is likely to underestimate the true likelihood of interruption in replacement repair.

Figure 2 indicates that it is not the case that all words are created equal in terms of the interaction of the Continuity Hypothesis (Clark and Wasow, 1998) with the Main Interruption Rule (Levelt, 1983, 1989). While in general, words are produced completely more often than they are interrupted in the present sample (61% of all words in the sample are not interrupted), Figure 3 suggests that this is an artifact of the fact that there are more short words than long words in the English lexicon. However, between-word transitions can still be privileged locations of interruption relative to word-internal segment-to-segment transitions. An eight-segment word maximally contains seven possible word-internal locations for interruption and one between-word location. Thus, if 40% of all interruptions involving an eight-segment replaced word occur in the between-word location, the between-word transition is privileged relative to the word-internal transitions as a location for interruption, as shown in Figure 3.

Figure 3 confirms that the data in Figure 2 do not contradict the Continuity hypothesis. As predicted by the Continuity hypothesis, for any word length, the between-word transition is a significantly more common location for interruption than any one of the within-word transitions according to the chi-square test (the closest contender among between-word transitions is the location after the third segment in eight-segment words that hosts 18 interruptions relative to 27 cases in which an eight-segment word is not interrupted; the difference is significant, $\chi^2(1)=7.2$, $p<.01$). On the other hand, since many interrupted cases of repair are not included in the sample because the identity of the interrupted word could not be

[7] Rank conversion of a set of numbers (e.g., frequencies) involves arranging the numbers from the highest to the lowest and replacing each number with its position in the sequence. For instance, if we have a sample of words that have frequencies of 1000, 2, 35, and 99, the corresponding ranks are 1, 4, 3, and 2 respectively. This is a standard way to deal with non-normal data.

guessed, this result does not provide strong evidence for the Continuity hypothesis. For that, we will have to turn to frequency effects.

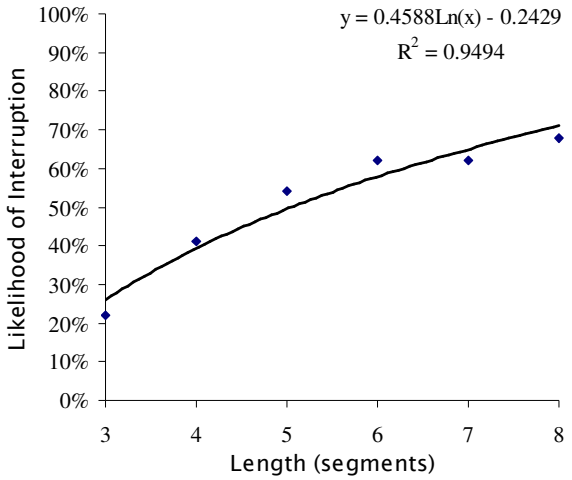


Figure 2. When a speaker intends to replace a word, s/he is more likely to interrupt it if it is long than if it is short.[8]

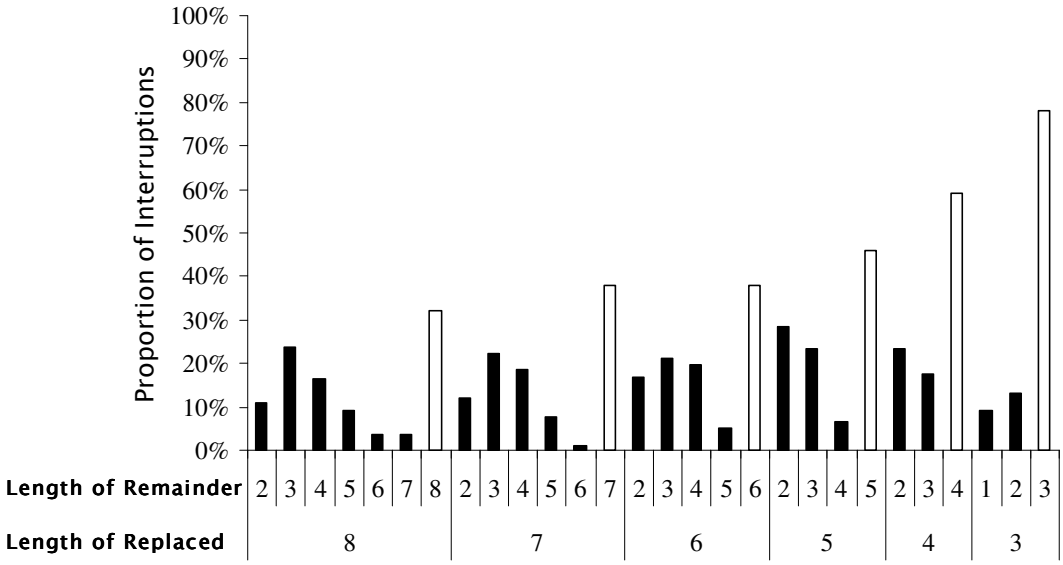


Figure 3. Interruptions are more likely to occur in a between-segment transition that spans a word boundary than in any between-segment transition within a word. The proportions shown are out of all interruptions involving replaced words of a given length. Thus, percentages within a bin defined by length of the replaced sum to 1.

High-frequency words tend to have fewer segments and Figures 2-3 show that words that have fewer segments are less likely to be interrupted. Therefore, for an effect of frequency on interruptibility to be established, it needs to be shown that it holds when number of segments is controlled. This is shown in

[8] Grouping the words by number of syllables rather than number of segments produces the same result.

Figure 4. For each word length, replaced words that are interrupted tend to be lower in frequency than words that are produced completely. The difference is statistically significant overall (in a multiple linear regression that also included log number of segments, interruption was a significant predictor of frequency, $t(1746)=9.934$, $p<.0005$; frequency is a significant predictor of interruption when frequency and length are entered into a binomial logistic regression as covariates, $p<.001$), as well as for three-, four-, five- and seven-segment words considered separately (for 3-segment words, $t(798)=7.821$, $p<.0005$[9]; for 4-segment words, $t(406)=4.092$, $p<.0005$; for 5-segment words, $t(190)=2.051$, $p=.042$; for seven-segment words, $t(131)=2.131$, $p=.035$). It is not significant for six-segment and eight-segment words.

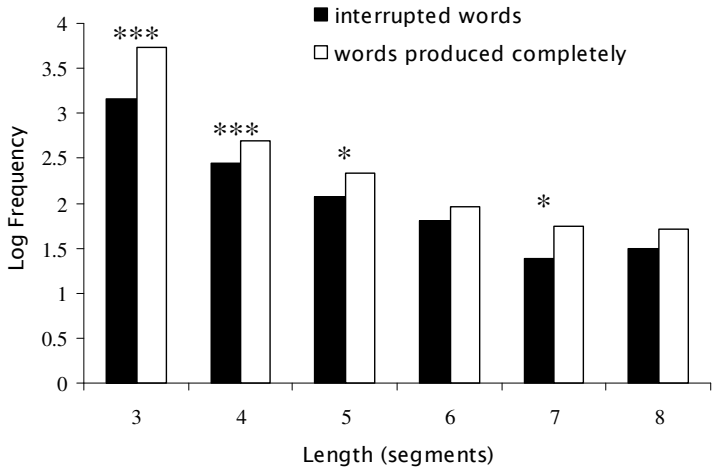


Figure 4. Words that are interrupted tend to be less frequent than words produced completely. One star indicates significance at the .05 level in a two-tailed t-test. Three stars indicate significance at the .005 level.

The data in Figure 4 support the hypothesis that high-frequency words are more cohesive than low-frequency words and their production is more automatized than the production of low-frequency words. However, alternative explanations are possible. The next section will consider an explanation based on speed of accessing the replacement, the following section explores possible observer bias, the one after that considers duration, and the one after that error detectability.

Frequency of the Replaced vs. Frequency of the Replacing

A possible alternative explanation is suggested by Biber et al.'s (1999) account of why frequent words are more likely to be repeated than rare words. It is possible that words that replace interrupted words are more frequent than words that replace uninterrupted words. If this were the case, interrupted words would be interrupted because the more appropriate alternative would come to mind more quickly. Assuming that the decision to interrupt production in single-word replacement repair is caused by activation of a more appropriate word, this decision would then be made earlier when the replacement word is frequent. And if the replaced word is rare and thus accessed slowly, the decision to replace the word would be made shortly after it accessed, giving speaker more opportunities to interrupt its production. Thus, this hypothesis predicts that interrupted words should be rarer *relative to their*

[9] For this analysis, frequencies were converted to frequency ranks as the distribution of log frequencies was highly skewed for three-segment words.

replacements than uninterrupted words without necessarily predicting a difference in absolute token frequency between interrupted and uninterrupted words.

However, the data are inconsistent with this prediction. Not only is there an absolute frequency effect in the data (as shown by Figure 4) but there is also no relative frequency effect. In order to derive estimates of relative frequency, the (log) frequency of each replaced word was divided by the sum of (log) frequencies of the replaced word and the corresponding replacement. Then mean relative frequency of interrupted replaced words was compared to mean relative frequency of uninterrupted replaced words. The mean relative frequency of interrupted words was .54 while the mean relative frequency of uninterrupted words was .53. This non-significant difference ($t(1029)<1$, $p=.4$) is in the opposite direction from the one predicted by the hypothesis.[10] Figure 5 shows that words that replace interrupted words tend to be less, rather than more, frequent than words that replace words that are produced completely.

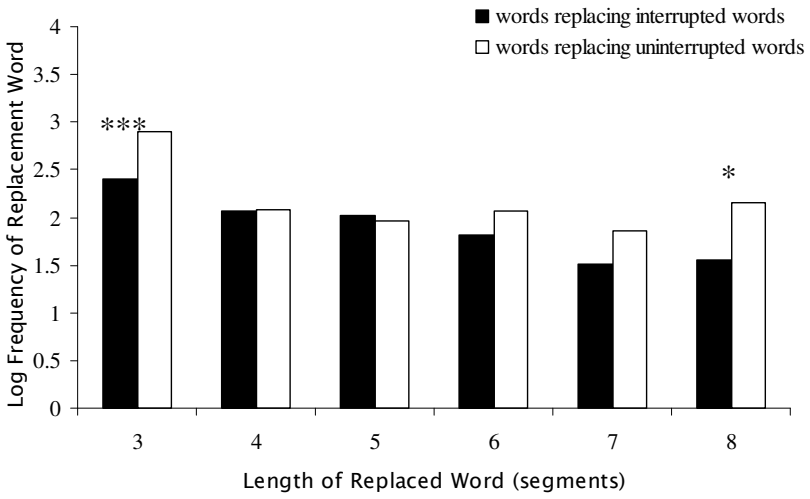


Figure 5. Words that replace interrupted words tend to be less frequent than words that replace words that are produced completely. One star indicates significance at the .05 level in a two-tailed t-test. Three stars indicate significance at the .005 level.

The reason for this result can be inferred from the data in Figure 6, which shows that the frequency of the replacement is positively correlated with the frequency of the replaced. Thus, the replaced and the replacement tend to be of similar frequency. This finding has also been observed in studies of lexical substitution errors (DelViso et al., 1991; Harley and MacAndrew, 2001; Hotopf, 1980; Silverberg, 1998). The correlation is very similar in magnitude to that obtained by Harley and MacAndrew (2001) in their study of lexical substitution errors: $r=.44$ in the present study, compared with $r=.4$ in Harley and MacAndrew (2001).

[10] For the purposes of the analyses reported in this section, multiple-word replacements and replacement words shorter than 3 or longer than 9 segments were eliminated to make the sample of replacements comparable to the sample of replaced words. Because of this, the sample only contains 1030 tokens.

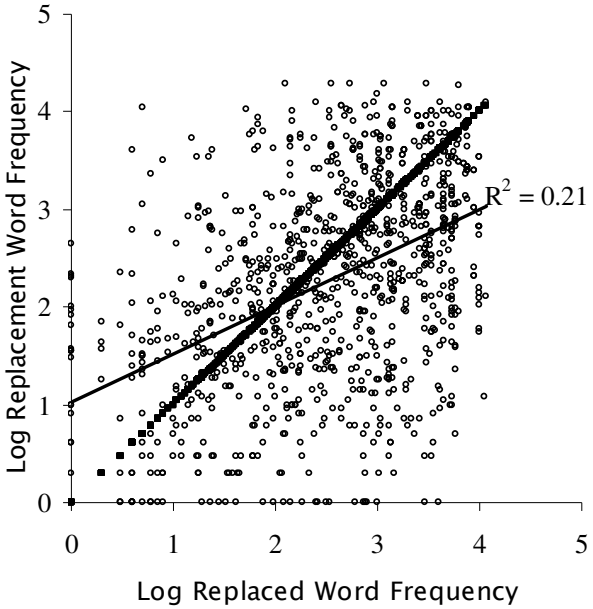


Figure 6. The relationship between the frequency of the replaced and the frequency of the replacement. A linear regression fit is shown as well as the diagonal for which the frequency of the replaced and the frequency of the replacement are equal. The data show a significant positive correlation between the frequency of the replaced and the frequency of the replacement. The fact that fewer points are above the diagonal than below indicates that the replacement tends to be less frequent than the replaced.

However, Figure 6 also shows that, unlike in studies of semantic substitutions, replaced words tend to be more frequent than the replacement words in the present sample. The replaced was more frequent than the replacing in 593 cases while the replacement was more frequent than the replaced in 428 cases. Thus, the replaced was more frequent than the replacement in 58% of the cases in which the two differed in frequency. While the effect is small, it is significant in both the chi-square test and the paired samples t-test ($\chi^2(1)=26.66$, $p<.0005$; $t(1029)=7.307$, $p<.0005$).

In order to make the data even more comparable to data obtained in studies of semantic substitution errors, all cases in which the replaced was interrupted were then eliminated from the sample, leaving 489 tokens. The frequency asymmetry was still observed. If anything, it became stronger: mean frequency of the replaced was 436 words/million while mean frequency of the replacement was 193 words/million, $t(488)=8.058$, $p<.0005$; the replaced was more frequent than the replacement in 63% of the cases, $\chi^2(1)=31.51$, $p<.0005$. The correlation between the frequency of the replaced and the frequency of the replacement was present as well ($r=.38$). Thus, the finding that the replaced tends to be more frequent than the replacement cannot be due to inclusion of interrupted replaced words in the present study and their exclusion from previous studies.

While Hotopf (1980: 100) and DelViso et al.'s (1991) report results that are in the same direction (in Hotopf, 1980, 56% of intrusions are more frequent than the corresponding targets; 53% in DelViso et al., 1991), albeit non-significant, Harley and MacAndrew (2001) do not. Furthermore, Harley and MacAndrew's (2001) sample is even larger than the present one (N=783 for Harley and MacAndrew

2001[11], vs. N=489 here). Furthermore, mean word frequencies in the full sample of repairs (199 for the replaced and 102 for the replacement) are similar to those in Harley and MacAndrew (2001) (153 and 165.5 respectively). It is possible that differences between methods of frequency estimation can account for the discrepancy between the present study and Harley and MacAndrew (2001). Frequencies used in Harley and MacAndrew (2001) are based on the written Brown Corpus (Francis and Kuçera, 1982) and the ones used here are based on the spoken Switchboard corpus (Godfrey et al., 1992). In addition to being spoken, the Switchboard corpus also has the advantage of being larger than the Brown corpus, more recent (the texts used in the Brown corpus were published in 1961), and being the same corpus as the one from which the disfluency tokens are drawn.

Coder Bias

Identification of an interrupted replaced word necessarily involves guessing whereas identification of a word that has been produced completely does not. Thus a possible explanation for why interrupted words tend to be of lower frequency than uninterrupted words is that my guesses are biased in favor of the hypothesis. That is, it is possible that I tend to come up with words that are lower in frequency than the words the speaker intended to produce. As the results of the previous section show, the frequency of the replaced word is correlated with the frequency of the replacement word. We can use this finding to assess the hypothesis of observer bias. If the frequency of the interrupted replaced words is lower relative to the corresponding replacement words than the frequency of the uninterrupted replaced words is relative to their corresponding replacements, the hypothesis of observer bias would be confirmed.

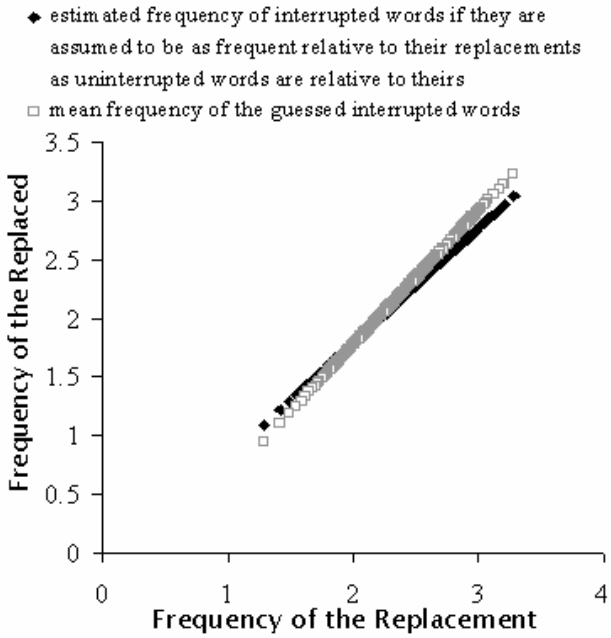


Figure 7. Frequencies of replaced interrupted words can be objectively estimated from the frequencies of the corresponding replacement words based on the relationship between the frequencies of replaced words and replacement words observed with uninterrupted tokens. These estimated frequencies of interrupted replaced words are then compared to the frequencies of guessed interrupted replaced words. If guesses are biased in favor of low-frequency words, the frequencies of guessed words would be lower than expected (the gray line would be below the black line). The present figure shows that this is not the case.

[11] Inferred from the df of the t-test comparing semantic targets and intrusions (Harley & MacAndrew 2001: 408).

First, the strength of the correlation between the frequency of the replaced and the frequency of the replacement does not depend on whether the replaced word is interrupted (r=.39 when the word is interrupted vs. .38 when it is not). More importantly, as Figure 7 shows, the replaced word, if anything, tends to be more frequent relative to the corresponding replacement when I had to guess its identity than when I did not. In other words, frequencies of guessed words are not lower than expected based on the frequencies of the corresponding replacement words, contradicting the hypothesis of observer bias in favor of low-frequency words.

The average frequency for interrupted words used in the sample (and guessed by me) was 47.5 words/million while the average estimated frequency based on the relationship between the frequencies of replaced uninterrupted words and the corresponding replacements was 46.5 word/million. Thus, the hypothesis that the difference in frequency between replaced and replacement words are due to observer bias is disconfirmed.

Erroneous vs. Suboptimal Choices and the Main Interruption Rule

Replacement repair does not always involve an outright speech error. Levelt (1983, 1989) proposed that repairs involving speech errors are very different from repairs that involve lexical choices that are considered suboptimal but not erroneous by the speaker. Levelt (1989: 481) proposes that "words that are not errors themselves tend to be completed before interruption… By interrupting a word, the speaker signals to the addressee that the word is an error. If a word is completed, the speaker intends the listener to interpret it as correctly delivered." Levelt (1983: 63) shows that in his corpus, 32% of immediate repairs of erroneously uttered words (91/284) involve interrupting the word, while only 11% of immediate repairs of suboptimal lexical choices involve interruption of the repaired (20/175). Furthermore, there is evidence that semantic substitution errors are more likely to involve low-frequency words than high-frequency words (Harley & MacAndrew, 2001). If high-frequency replaced words are more likely to be merely inappropriate rather than erroneous low-frequency replaced words, the frequency effect observed in the present study could be ascribed to the error severity effect observed by Levelt. (Of course, this argument cuts both ways. One could also argue that the error severity effect is a frequency effect in disguise.)

For maximum coding reliability, only instances of repair in which the replaced was not interrupted were analyzed. There was no tendency for repairs involving speech errors to involve less frequent words than repairs involving suboptimal lexical choices. The results are shown in Table 1. No differences in frequency between erroneous and suboptimal words are significant. Thus, we can reject the hypothesis that high-frequency words are less likely to be interrupted because they are less likely to be uttered in error for the present sample.

Length	Erroneous Words	Suboptimal Words
3	3.42 N=107	3.35 N=112
4	2.62 N=41	2.63 N=67
5	2.26 N=8	2.34 N=18
6	2.05 N=14	2.00 N=13

Table 1. Log frequencies of uninterrupted replaced erroneous vs. suboptimal words: erroneous words do not tend to be less frequent.

In addition to the problem of accounting for the present data, the hypothesis that high-frequency words are less likely to be interrupted than low-frequency words because high-frequency words are less likely to be produced as errors runs into problems with Logan's (1982) data. In his study, replacement was triggered by an external stop signal, rather than erroneous or inappropriate production. Nonetheless, a frequency effect was present.

A way to maintain the Main Interruption Rule (Levelt, 1983, 1989), which states that the speaker interrupts speech production as soon as the occasion for repair is detected, in the face of the present data would be to say that speakers are slower to detect that a high-frequency word is wrong or inappropriate than they are to detect the incorrectness of a low-frequency word. However, this hypothesis cannot account for Logan's (1982) experimental data where there is no error to be detected.[12] To account for why 'the' is typed completely after the stop signal is presented while less frequent words are truncated without invoking a preference to maintain constituent continuity would be to say that the detection of the stop signal is slowed down when a high-frequency word is being produced. It is not clear why this should be the case. If anything, production of a high-frequency word should be less taxing and demand fewer cognitive resources than the production of a low-frequency word, leaving more cognitive resources free to be used in perceiving the stop signal. Thus, if anything, we would predict the perception of the stop signal to be faster while a high-frequency word is being produced than while a low-frequency word is under construction.

Frequency and Duration

There is a strong negative correlation between word frequency and word duration (r = -.72 in the present sample), which remains even when number of segments is controlled (in the present sample, r = -.63 for three-segment, -.55 for four-segment, -.53 for six-segment, -.35 for seven-segment, and -.65 for eight-segment uninterrupted replaced words; no correlation is observed for five-segment words, r=-.04). Thus, frequent words tend to be shorter than rare words even when number of segments is controlled (as previously found in corpus studies by Gregory et al., 2000, and Jurafsky et al., 2001). This finding is predicted by the hypothesis that high frequency leads to automatization of production but it suggests that the effect of frequency on interruptibility may be accounted for by the effect of frequency on duration.

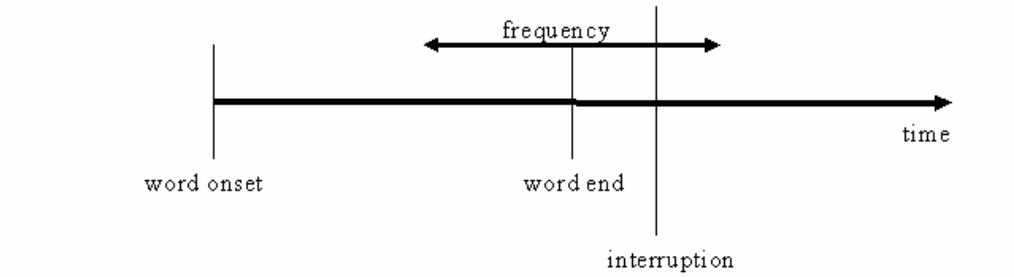


Figure 8. A model in which the only variable affected by frequency is word duration. Time since word onset is indicated by the thick line. Vertical lines mark important points in time, such as the end of the word and the location of interruption. The arrows attached to a vertical line indicate the extent to which variation in frequency can influence the location of the vertical line. In this model, the only point in time whose location is influenced by frequency is the end of the word, which can fall after or before the fixed location of interruption.

[12] One could claim that the 'error' being detected is the fact that production is still continuing. Then one could say that detection of this fact is more difficult when the word being produced is more frequent. However, this would mean that the production of high-frequency words is less cognitively penetrable than the production of low-frequency words, which is precisely the claim of the automaticity hypothesis.

The simplest model of the effect of frequency on interruptibility is that there is no effect. How long it takes a speaker to reach and carry out the decision to interrupt a word is independent of frequency. Rather, all that frequency influences is word duration. As Figure 8 shows, the location of interruption relative to the onset of the word is fixed in this model. The only variable affected by frequency is the duration of the word. When frequency is low enough, the word becomes so long that the decision to interrupt speech is carried out before the word is produced completely.

Under this model, words are interrupted only if they are sufficiently long. Therefore, there should be no difference in duration between the remainders of interrupted replaced words and uninterrupted replaced words. This is not the case in the data. Overall, remainders of interrupted words (mean duration = 217 ms) are shorter than uninterrupted replaced words (mean duration = 316 ms): $t(1136)=15.97$, $p<.001$. Figure 9 shows the results broken down by the length of the replaced in segments. This result indicates that interruption comes earlier in time, relative to the beginning of the to-be-replaced word, when the word is interrupted than when it is not, contrary to the predictions of the model in Figure 8. Thus, there is something about uninterrupted words that delays interruption when these words are produced. This is consistent with Logan's (1982) results regarding the very frequent word 'the': while typers took less time to type 'the' than other words, the hypothesis that the difference in typing speed accounted for the result was ruled out because the time it took typists to stop while producing 'the' was longer than the time it took them to stop while producing other words.

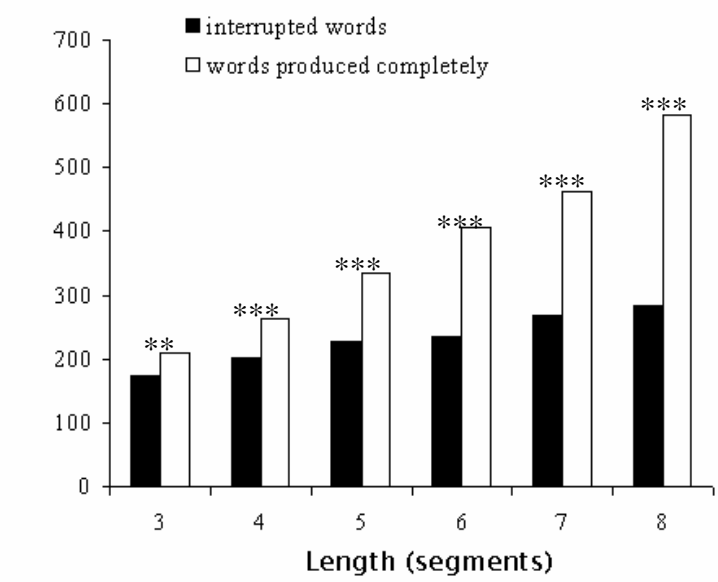


Figure 9. Remainders of interrupted words are shorter than uninterrupted replaced words.

An objection that could be raised to our interpretation of the data in Figure 9 is that the coder could sometimes erroneously code words that are interrupted very late during their production as uninterrupted. This is presumably not a problem with Logan's (1982) data because typing involves a discrete response while speech is continuous and involves extensive co-articulation (e.g., Coleman, 2003; Öhman, 1966), which means that the cues for the final segment can be present much earlier in the word. Furthermore, there may be more coarticulation in high-frequency words than in low-frequency words (Yun, 2006). This is in fact suggested by the data in Figure 3 where the probability of interrupting the word drops off just before the word is completed.

However, the data in Figure 3 can be interpreted in multiple ways, including misperception, a tradeoff between the speaker's desire not to interrupt the word and the desire to interrupt as soon as possible, and generally early detection of errors with interruption being sometimes delayed until the end of the word. In addition, within-word interruption is reliably accompanied by a particular cue, the presence of glottalization (Ladefoged et al., 1973). There was very little disagreement between the present author and the Switchboard corpus coders (only 31 words were eliminated based on this criterion). In addition, the word lengths for uninterrupted words used in the present study do not include the word-final segment. It is highly unlikely that the coders would have coded a word as uninterrupted without perceiving the final segment. Finally, the correlation between word duration and frequency is negative. Therefore, if we were to project the durations of the remainders of interrupted words from the relationship between frequency and duration found in uninterrupted words, we would expect durations of remainders of interrupted words to be longer than the durations of uninterrupted words because the frequencies of interrupted words are lower than frequencies of uninterrupted words.

The data presented so far are sufficient to reject the simple model in Figure 8. The differences in duration between the remainders of interrupted and uninterrupted words are too great to be ascribed to differences in duration between the corresponding complete words. However, the data presented thus far and Logan's (1982) results for typing 'the' are consistent with the model is shown in Figure 11. This model relies on the assumption that the closer the speaker is to the end of the word when s/he reaches the decision that the word is to be replaced, the less likely s/he will be to stop immediately. One can think of the speaker as choosing the better of two evils: to stop immediately, interrupting a cohesive constituent, or to continue producing material that will need to be replaced. In other words, the speaker can be thought of as choosing between violating the Continuity hypothesis (Clark & Wasow, 1998) vs. violating the Main Interruption Rule (Levelt, 1983, 1989). The smaller the amount of material that remains to be produced to avoid interrupting the word, the more likely the speaker is to choose producing the word to the end. Since frequency influences word duration, the amount of material that needs to be produced to complete the word will be smaller in a high-frequency word than in a low-frequency word.

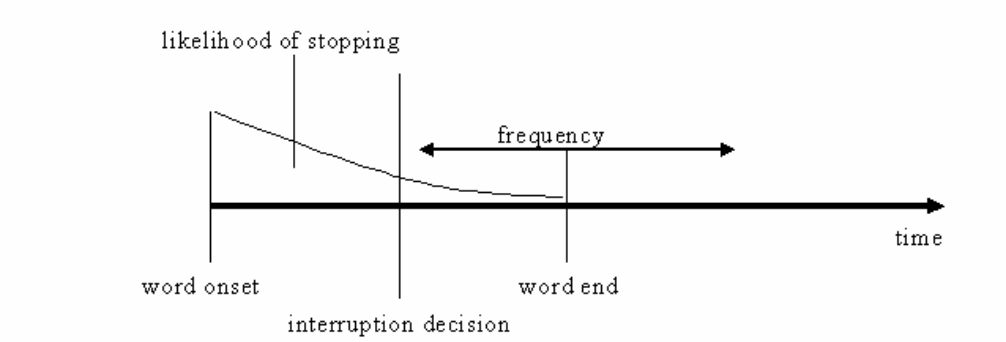


Figure 11. A model in which likelihood of interrupting the speech stream immediately is lower if the amount of material that remains to be produced or time that it takes to complete the word is small. The likelihood of stopping immediately is shown by the height of the curved line. The higher the curve at a certain point in time, the higher the likelihood that the word will be interrupted immediately if the decision to interrupt is made at that point in time. In this model, the closer a speaker is to the end of the word, the less likely s/he is to interrupt speech production immediately. Word duration is influenced by frequency, so a speaker is more likely to be close to the end of the word when deciding to interrupt speech production if the word is frequent than if it is rare. Thus, in a frequent word, the interruption decision is likely to occur at a point when likelihood of stopping immediately is low.

An alternative model is presented in Figure 12. Here, the speaker's reluctance to interrupt a word is simply greater if the word is frequent than if the word is rare, regardless of how much linguistic material remains to be produced and how much time it would take to complete the word. There may be an effect of duration but frequency has an effect on likelihood of stopping that is independent of duration.

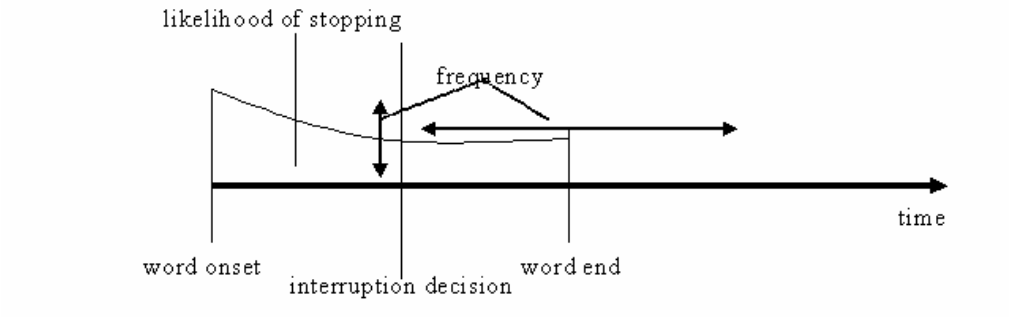


Figure 12. A model in which interruption is dispreferred in frequent words. Frequency in this model influences both the duration of the word and the likelihood of stopping immediately if the interruption decision is reached during word production as indicated by the arrows being attached to the curve indicating likelihood of stopping immediately (the curve is not crucial for this model and could be replaced by a horizontal line).

The difference between the two models lies in whether frequency has any effect on interruption when duration is controlled (both models can account for an independent effect of duration since duration is uncontroversially influenced by factors other than frequency such as speaking rate and number of segments). In order to examine this issue, a binomial logistic regression was conducted. Duration, frequency, and number of segments were entered into the analysis as covariates. Number of segments was subsequently excluded because it was not statistically significant as a separate predictor. Thus, the analyses presented below included only duration and frequency as covariates. Both were significant at the .0001 level on the full sample. The sample was then split by number of syllables so that monosyllabic and multisyllabic words were submitted to the regression analysis separately. Both frequency and duration were significant in both analyses. Frequency was significant with p=.001 for multisyllabic and p<.0001 for monosyllabic words. Duration was significant with p=.014 for multisyllabic and p=.01 for monosyllabic words, N=717 for multisyllabic words, N=1032 for monosyllabic words. Thus we can tentatively conclude that frequency has some effect on interruptibility that is not mediated by the effect of frequency on duration.[13]

Conclusion

When a speaker intends to replace a word s/he has started producing, s/he has the choice of stopping immediately, obeying Levelt's (1983) Main Interruption Rule, or delaying interruption until the word is completed, obeying Clark and Wasow's (1998) Continuity Hypothesis. The present study has argued that the speaker's choice is influenced by word duration and word frequency. Speakers prefer not

[13] A necessary caveat for this conclusion is that our estimates of frequency and duration are imperfect. The full model achieved only 61% accuracy in predicting whether the word was broken when the word was multisyllabic and 75% accuracy when the word was monosyllabic, suggesting that there is much room for improvement in modeling interruptibility. Perhaps, frequency would not account for any variance that duration does not account for as well if our estimate of duration were better. However, the fact that including number of segments or number of syllables as an additional predictor does not reduce the significance of frequency suggests that this is unlikely.

to interrupt high-frequency words. This effect provides novel empirical support for the hypothesis that the production of high-frequency words is more automatic, being both faster and less susceptible to conscious control than the production of low-frequency words (Bybee, 2002; Logan, 1982). Thus Bybee's (2002) hypothesis that reductive sound change starts with high-frequency words because the production of such words is more automatic is at least psychologically plausible. In addition, the present study found that speakers tend to replace suboptimal lexical choices by less frequent but more appropriate words, supporting the idea that high-frequency words are accessed faster than low-frequency words (e.g., Jescheniak and Levelt, 1994). Thus, frequent words are easier to access, faster to produce, and harder to interrupt than rare words.

References

Anderson, J.R. (2000). *Cognitive psychology and its implications*. New York: Worth.

Beattie, G.F., & Butterworth, B.L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*, 201-11.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, *8*, 148-58.

Bybee, J.L. (2002). Word frequency and context of use in the lexical diffusion of phonemically conditioned sound change. *Language Variation and Change*, *14*, 261-90.

Bybee, J.L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1430-50.

Clark, H.H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*, 201-42.

Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics*, *31*, 351-72.

DelViso, S., Igoa, J.M., & Garcia-Albea, J.E. (1991). On the autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research*, *20*, 161-85.

DuBois, J.W. (1974). Syntax in mid-sentence. *Berkeley Studies in Syntax and Semantics*, *1,* III:1-25.

Fowler, C. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech, 28,* 47-56.

Fowler, C.A., & Housum, J. (1987). Talkers' signaling of "old" and "new" words in speech and listeners' use of this distinction. *Journal of Memory and Language*, *26*, 489-504.

Fox, B.A., Hayashi, M., & Jasperson, R. (1996). Resources and repair: A cross-linguistic study of syntax and repair. In E. Ochs, E. Schegloff, & S. Thompson (eds.), *Interaction and grammar* (pp. 185-237). Cambridge, UK: Cambridge University Press.

Fox, B.A., & Jasperson, R. (1995). A syntactic exploration of repair in English conversation. In P.W. Davis (ed.), *Alternative linguistics: Descriptive and theoretical modes* (pp. 77-134). Amsterdam: John Benjamins.

Francis, W.N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston, MA: Houghton Mifflin.

Garrett, M. (2001). Now you see it, now you don't: Frequency effects in language production. In E. Dupoux (ed.), *Language, brain and cognitive development* (pp. 227-40). Cambridge, MA: MIT Press.

Godfrey, J.J., Holliman, E.C., & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. *IEEE ICASSP*, I517-20.

Goldman-Eisler, F. (1958). Speech production and predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*, 96-106.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* San Diego: Academic Press.

Gregory, M., Raymond, W.D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (2000). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society*, *35*, 151-66.

Harley, T.A., & MacAndrew, S.B.G. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research*, *30*, 395-18.

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (ed.), *Current progress in historical linguistics* (pp. 96-105). Amsterdam: North Holland.

Hotopf, W. (1980). Semantic similarity as a factor in whole-word slips of the tongue. In V. Fromkin (ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 97-110). New York: Academic Press.

Howes, D.H., & Solomon, R.L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41,* 401-10.

Jescheniak, J.D., & Levelt, W.J.M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 824-843.

Jescheniak, J.D., Meyer, A.S., & Levelt, W.J.M. (2003). Specific word frequency is not all that counts in speech production: Comments on Caramazza, Costa et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 432-438.

Jurafsky, D., Bell, A., & Girand, C. (2002). The role of lemma in form variation. In C. Gussenhoven & N. Warner (eds.), *Papers in Laboratory Phonology VII* (pp.1-34). Berlin/New York: Mouton de Gruyter.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W.D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J.L. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure* (pp. 229-54). Amsterdam: John Benjamins.

Kapatsinski, V.M. (2005). Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *Berkeley Linguistics Society*, *30*, 481-92.

Ladefoged, P., Silverstein, R., & Papçun, G. (1973). Interruptibility of speech. *Journal of the Acoustical Society of America*, *54*, 1105-1108.

Levelt, W.J.M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41-104.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (eds.), *Speech production and speech modeling* (pp. 403-39). Dordrecht: Kluwer.

Logan, G.D. (1982). On the ability to inhibit complex movements: A stop-signal study of typewriting. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 778-792.

Maclay, H., & Osgood, C.E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*, 19-44.

Mowrey, R., & Pagliuca, W. (1995). The reductive character of articulatory evolution. *Rivista di Linguistica*, *7*, 37-124.

Nooteboom, S. (1980). Speaking and unspeaking: detection and correction of phonological and lexical errors in spontaneous speech. In V. Fromkin (ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 87-95). San Diego: Academic Press.

Öhman, S.E.G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151-168.

Oldfield, R.C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*, 273-281.

Pagliuca, W., & Mowrey, R. (1987). Articulatory evolution. In A.G. Ramat, O. Carruba, & G. Bernini (eds.), *Papers from the VIIth International Conference on Historical Linguistics* (pp. 459-72). Amsterdam: John Benjamins.

Phillips, B.S. (2001). Lexical diffusion, lexical frequency, and lexical analysis. In J.L. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure* (pp. 123-36). Amsterdam: John Benjamins.

Slevc, L.R., & Ferreira, V.S. (2006). Halting in single word production: A test of the perceptual loop theory of speech monitoring. *Journal of Memory and Language*, *54*, 515-540.

Stemberger, J.P. (1984). Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology*, *1*, 281-313.

Tannenbaum, P.H., Williams, F., & Hillier, C.S. (1965). Word predictability in the environments of hesitation. *Journal of Verbal Learning and Verbal Behavior*, *4*, 134-40.

Yun, G. (2006). The effects of lexical frequency and stress on coarticulation. Paper presented at Berkeley Linguistics Society Meeting 32, Berkeley, CA.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Development of Lexical Connectivity in Pediatric Cochlear Implant Users[1]

Thomas M. Gruenenfelder and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This research was supported by NIH Grant DC00111 to Indiana University.

Development of Lexical Connectivity in Pediatric Cochlear Implant Users

Abstract. The present study examined the performance of pediatric cochlear implant (CI) users on easy (high frequency words from low density neighborhoods) and hard (low frequency words from high density neighborhoods) words on the monosyllabic Lexical Neighborhood Test (LNT) and Multi-syllabic Lexical Neighborhood Test (MLNT). The easy—hard effect (the superior performance on easy words compared to hard words) increased slightly on the LNT for Oral Communication (OC) users but not for Total Communication (TC) users as lexicon size increased. The easy—hard effect was invariant as a function of lexicon size for both OC and TC users on the MLNT. Similarly, the word length effect (the superior performance for long words on the MLNT compared to short words on the LNT) did not vary as a function of lexicon size for either OC or TC users. The size of the easy—hard effect was not correlated with the size of the word length effect for either class of users. When lexicon size was controlled for, OC users performed better on both the LNT and MLNT than did TC users. These results were discussed in terms of how the mental lexicon of CI users develops over time and the role of lexical connectivity in spoken word recognition.

Introduction

Because traditional word recognition tests underestimate the ability of hearing impaired children to comprehend spoken words, Kirk, Pisoni, and Osberger (1995) developed two new spoken word recognition tests, the Lexical Neighborhood Test (LNT) and the Multisyllabic Lexical Neighborhood Test (MLNT). Both tests involve open-set identification of words spoken in isolation (but see Eisenberg, Martinez, Holowecky, & Pogorelsky, 2002 for an extension of these tests to recognizing words in sentences). The LNT uses only monosyllabic words; the MLNT uses two and three syllable words.

Two major criteria were used when selecting items from the test. First, each word used in the test should have a relatively high probability of being in the child's lexicon. Kirk et al. (1995) noted, for example, that fewer than 1/3 of the words on the commonly used Phonetically Balanced Kindergarten (PB-K) test are found in Logan's (1992) computational analyses of the CHILDES database (MacWhinney & Snow, 1985). A child could, in theory, accurately reproduce an unknown word, if all the word's phonetic details are accurately perceived. If, however, the child perceives only some information for a portion of the word, then he/she needs to make an educated guess concerning what the word is. That educated guess is likely to be confined to words in the child's lexicon and not to include words unfamiliar to the child. The result would be lower scores on a test containing a higher proportion of words unknown to the child. To avoid this bias, in their new tests, Kirk et al. used only words produced by children from 3 to 5 years of age (Logan, 1992).

Second, Kirk et al. (1995) selected their test words in accordance with the assumptions of prevailing theories of spoken word recognition (Auer & Luce, 2005; Luce & Pisoni, 1998, Marslen-Wilson 1987, 1989; McClelland & Elman, 1986; Norris, 1994). In particular, they included in each of their test sets, two types of words—some predicted by these theories to be easy and some predicted to be hard. Words with a high frequency of occurrence in the language (e.g., Kucera & Francis, 1967) are typically easier to recognize than words with a lower frequency of language (e.g., Andrews, 1989; Elliot, Clifton, & Servi, 1983; Howes, 1957; Pollack, Rubenstein, & Decker, 1959; Savin, 1963). In addition, words phonetically similar to few other words are generally easier to recognize than words phonetically similar to many other words (e.g., Luce & Pisoni, 1998; McClelland & Elman, 1986; Treisman, 1978a,

1978b; Vitevitch & Luce, 1999). The single phoneme Deletion, Addition, and Substitution (DAS) rule is frequently used to operationally define phonetic similarity (Eukel, 1980; Greenberg & Jenkins, 1964; Landauer & Streeter, 1973). Two words are considered phonetic neighbors (and hence phonetically similar to one another) if one can be changed into the other by the deletion, addition, or substitution of a single phoneme. Words with many neighbors are said to come from high density neighborhoods; those with fewer neighbors to come from low density neighborhoods. The finding that words from low density neighborhoods are easier to recognize than words from high density neighborhoods is referred to as the neighborhood density effect.

The word frequency effect and the neighborhood density effect are both compatible with the general predictions of several current models of spoken word recognition (e.g., Luce & Pisoni, 1998, Marslen-Wilson 1987, 1989; McClelland & Elman, 1996; Norris, 1994, Treisman, 1978a). Indeed, they are two fundamental facts which such models must account for. The Neighborhood Activation Model (NAM) of Luce and Pisoni (1998), for example, assumes that a spoken word activates its representation in the mental lexicon and also the representations of its phonetic neighbors. A probabilistic decision rule is then used to select among the activated representations. The activation levels themselves are adjusted multiplicatively by the word's frequency of occurrence, biasing the recognition process towards more common words.

In developing their LNT, Kirk et al. (1995) selected half of their words to be "easy" and half to be "hard." Easy words were a) above the median frequency of usage in Logan's (1992) corpus of the speech of 3 to 5 year olds, and b) below the median neighborhood density in Logan's corpus. Hard words, in contrast, were below the median frequency of usage and above the median neighborhood density. Words in the LNT were restricted to monosyllabic words. Similarly, in the MLNT, easy words were words above the median frequency of usage for multisyllabic words and below the mean neighborhood density for multisyllabic words. Hard words were below the median frequency and above the median neighborhood density for multisyllabic words. Words in the MLNT were restricted to multisyllabic words.

Note that because monosyllabic words tend to be used more frequently and come from higher density lexical neighborhoods (Gruenenfelder & Pisoni, 2005) than multisyllabic words, the cutoff values used in the LNT differed from those used in the MLNT. In particular, for monosyllabic words, the median frequency in the Logan (1992) corpus was four occurrences and the median density was four neighbors. These were the cutoff values used in constructing the LNT. In contrast, for multisyllabic words, the median frequency was two occurrences and the median density was zero neighbors, and these were the cutoffs used in constructing the MLNT.

Not surprisingly, at least in open-set tests, when lists constructed in this manner are used, easy words are recognized more easily by normal hearing adults than hard words (e.g., Sommers, Kirk, & Pisoni, 1997). Sommers et al. found the same result for CI users who were apparently deafened as adults and were implanted as adults. Kirk, Pisoni, and Miyamoto (1997) found similar results with mildly to moderately impaired adult listeners. Similarly, normal hearing children listening to amplitude-reduced speech, normal hearing children listening to spectrally degraded stimuli, and hearing impaired children using CIs are better able to recognize easy words than hard words, both when the words are spoken in isolation and when they are part of a meaningful (though not overly semantically constrained) sentence (Eisenberg, Martinez, Holowecky, & Pogorelsky, 2002). Bell and Wilson (2001) reported similar results for normal hearing adults listening to sentences in noise. These findings replicate the word frequency and neighborhood density effects mentioned above.

More significantly, Kirk et al. (1995) found a similar easy-hard effect in hearing impaired children using cochlear implants. Pediatric cochlear implant (CI) users more accurately identified easy words than hard words in both the LNT and the MLNT. In addition, a word length effect was also found: pediatric CI users performed better on the MLNT (multisyllabic words) than on the LNT (monosyllabic words). Similar results were found by Kirk, Eisenberg, Martinez, and Hay-McCutcheon (1999). Kirk et al. (1995) interpreted the easy-hard effect to mean that CI users, like normal hearing adults, organize their mental lexicon into similarity neighborhoods of words and that they use this organization when identifying spoken words.

The present paper uses the LNT and MLNT developed by Kirk et al. (1995) to address three different issues. The first concerns the "representational specificity" of the phonetic categories used by pediatric CI users when recognizing spoken words. To the extent that CI listeners use relatively broad phonetic categories, we might expect relatively robust easy-hard effects. In a test involving the recognition of isolated words, broad phonetic categories force the CI user to make educated guesses concerning the identity of each test word based on what amounts to limited phonetic input. Given an easy word—a high frequency word with few neighbors—that guess is relatively likely to hone in on the correct word. In the case of an easy word, there are, in effect, fewer phonetically similar distractors competing for recognition. If the guess is primarily confined to the actual word's neighbors, then the fewer the neighbors, the more likely is the guess to be correct, resulting in an easy-hard effect. Further, to the extent that the CI user is biased towards guessing higher frequency words (cf. Luce & Pisoni, 1998), easy words are more likely to be guessed correctly than hard words, resulting in an even larger easy-hard effect.

Over time, as phonetic representations become more refined and detailed, such educated guessing is less necessary, and the easy-hard effect should thus become smaller. In fact, if the phonetic categories were fine enough to discriminate all the words in the CI user's mental lexicon, then performance for both easy and hard words would be at ceiling and no easy-hard effect would occur. Exactly such a phenomenon is evident with normal hearing listeners. These listeners perform near ceiling on isolated words heard in the clear, i.e., at high Signal-to-Noise ratios. Only when the words are spoken in noise or are distorted in some way does the easy-hard effect emerge (e.g., Sommers et al., 1997). This reasoning suggests that the easy-hard effect should be relatively large shortly after the listener has received a CI (but long enough after so that the CI user is adequately perceiving some phonetic information) and then gradually diminish as the CI user gains experience with the device. Accordingly, the first purpose of the present paper was to examine the development of the easy-hard effect as a function of the time after implant.

Consistent with this reasoning, Eisenberg et al. (2002) found a somewhat larger easy-hard difference in a group of low performing CI users (N = 3) than in a group of high performing CI users (N = 9). However, the sample sizes were small. Consequently, no statistical analyses comparing the two groups were performed and the difference between groups is unlikely to have reached statistical significance if such analyses had been performed. Eisenberg et al. did find a larger easy-hard effect for normal-hearing children listening to spectrally degraded speech (The speech was reduced to four spectral channels.) than for normal-hearing children listening to speech in the clear albeit at reduced intensity (25 and 30 dBA). Overall, percent correct for children listening to intensity-reduced speech was higher (~65% correct) than for children listening to spectrally-degraded speech (~55% correct), suggesting that the latter group was extracting broader phonetic categories from the stimuli than was the former group. This finding is consistent with the hypothesis above that easy-hard effects should become smaller as phonetic information becomes more refined. On the other hand, the high performing CI users in Eisenberg et al.'s study showed an easy-hard effect of the same magnitude as the children listening to

intensity-reduced speech, even though their overall performance level was considerably higher (~85% correct).

An alternative view suggests a quite different time course of the development of the easy-hard effect. At early test intervals after receiving an implant, CI users' mental lexicon of **spoken** words may be quite small. When the mental lexicon is small, performance on these tests is likely to be relatively poor, simply because the child does not know many of the words on the tests. Further, because the lexicon is small, statistically any given word is unlikely to have many neighbors and hence no neighborhood density effects emerge. As the lexicon grows, words acquire neighbors and neighborhood density effects begin to emerge. The result should be that, to the extent that the easy-hard effect is at least partially due to neighborhood density (and not entirely to word frequency), the easy-hard effect should grow over time, until easy words begin to reach ceiling and the hard words catch up. Another way of stating this prediction is that over time, performance on easy words should improve faster (until it reaches ceiling) than performance on hard words, and, to the extent that the LNT/MLNT difference also reflects neighborhood density, performance on the MLNT should improve faster than performance on the LNT.

A second issue addressed in this paper concerns the origin of the word length effect. There are at least two possible reasons why multisyllabic words are recognized more easily than monosyllabic words by CI users (as well as by normal hearing adults). First, CI users may be sensitive to word length, in terms of number of syllables, as a word recognition cue. More specifically, it may be easier for them to extract word length information from a spoken word than it is to extract fine phonetic information. Given partial phonemic information, word length may help CI users choose one lexical representation from among several competing representations. That is, if the listener knows the word contains two syllables, then any competing representations of monosyllabic or trisyllabic words can be eliminated.

The second possible reason for the word length effect concerns differences in neighborhood density of the words used in the MLNT and LNT. In Logan's (1992) corpus, multisyllabic words tended to have fewer neighbors (and to occur with lower frequency) than monosyllabic words. (Similarly, if a lexicon more representative of that of college students is used, for example, Nusbaum, Pisoni, & Davis, 1984, multisyllabic words have fewer neighbors and a lower frequency of occurrence than do monosyllabic words.) For monosyllabic words, the median number of neighbors in Logan's corpus was 4 (range 0 – 19). The median frequency of occurrence was also 4 (range 1 – 519). In contrast, for multisyllabic words, the median number of neighbors was 0 (range 0 – 7). The median frequency of occurrence was 2 (range 1 to 100). To the extent that the effects of neighborhood density are stronger than those of word frequency, the word length effect may occur not because CI users are sensitive to word length but because MLNT words come from less dense neighborhoods than do LNT words and are therefore less confusable with other phonetically similar words.

To summarize, the word length effect may be due to CI users' sensitivity to the syllabic structure of words or it may be due to neighborhood density differences between shorter and longer words. We made a preliminary attempt to disentangle these two hypotheses by comparing the time course of development of the easy-hard effect with that of the word length effect. To the extent that these time courses parallel one another, the hypothesis that both effects are due to the same underlying variable is supported. To the extent that the two effects develop with different time courses, the hypothesis that they have different causes—viz. neighborhood density for the easy-hard effect and number of syllables for the word length effect—would be supported.

We realize that this test is less than ideal, especially given that MLNT words have lower frequency of occurrences than LNT words, which should work against a word length effect. Nevertheless, we think that this approach could provide at least a starting point in better understanding of the word length effect.

The third issue we explored in the current paper concerned the effects of early experience on spoken word recognition in deaf children with CIs. More specifically, we examined differences in the structure of the mental lexicon of oral communication (OC) CI users and total communication (TC) CI users. Although Kirk et al. (1995) used both types of users in their original report, they did not report results separately for the two groups. We might expect that TC users extract broader phonetic categories from the acoustic stimulus than do OC users, simply because TC users rely more on non-phonetic cues for understanding language than do OC users. If so, then TC users should show larger easy-hard effects than do OC users. Alternatively, it may be that TC users organize their mental lexicons in an entirely different manner than normal hearing adults and OC CI users. What is phonetically similar to TC users may not be phonetically similar to OC users, and vice versa. In that case, we might expect to see a greatly reduced easy-hard effect in TC users.

Method

Participants

The participants were 138 children receiving services at the Indiana University Medical Center who had provided informed consent allowing the use of their test results for research purposes. Different analyses included different subsets of these 138 participants. Hence, this group is referred to as the master group. All participants had CIs and all test results reported here were collected after the implant had been received. Testing was done as part of the participant's regular post-implant clinical appointments. Most children were tested during multiple appointments, each appointment being approximately an integer multiple of 6 months post-implant. A test interval of 0 corresponds to as near as possible immediately after implant, a test interval of 1 to 6 months post-implant, a test interval of 2 to 12 months post-implant and so on. Table 1 shows some characteristics of the participants. Ninety of the participants were users of Oral Communication (OC) and 48 were users of Total Communication (TC).

	Mean (mos.)	**SD (mos.)**
Age at Onset	6.45	19.22
Age at Implant	47.20	26.50
Age at First Test Interval	92.47	32.11

Table 1. Characteristics of the master group of 138 participants.

Procedure

Each child received a battery of tests at each interval. The tests a particular child received at a given interval were not necessarily the same as those received by another child at the same interval, nor were they necessarily the same as those that child had received at the previous interval. We created a master file by selecting all intervals for each child in which the child had received the LNT test. At no interval did a child receive the MLNT without also receiving the LNT. Hence, this procedure includes all the LNT and MLNT data collected from these children. Since most children received the LNT at multiple

test intervals, most contributed multiple data points in this master file. The master file included data from 443 test intervals across the 138 children. Hence, the mean number of intervals on which each child was tested was 3.22 (S.D.=1.91). The mean interval after implant at which testing occurred was 9 (S.D.=4.47).

Appropriate subsets of data were then selected from this master file according to the specific hypotheses being tested. For example, when testing a hypothesis involving only LNT scores, data from all 443 test intervals were included. In contrast, when testing a hypothesis concerning the relation of LNT and MLNT scores, only those test intervals were included where a child had contributed both LNT and MLNT scores. The specific subsets of data included in each analysis are described in the results section. The general procedures for administering the LNT and MLNT at the Indiana University Medical Center are described in Kirk et al. (1995).

Results

Changes in the Easy-Hard Effect with Lexicon Size

The first issue that we investigated concerned changes in the easy—hard effect as a function of the size of the child's lexicon. Overall performance on the LNT was used as a measure of lexicon size. Figure 1 is a scatter plot of the easy-hard effect as a function of overall LNT percent correct. When all samples from our master group were included, the size of the easy-hard effect correlated positively with LNT performance, $r = 0.28$, $t(441) = 6.13$, $p < .001$, indicating that as the child's lexicon grew so did the easy-hard effect. This analysis, however, is susceptible to ceiling and especially floor effects. A child with no lexical knowledge at all would score 0% on the LNT, showing an easy-hard effect of 0. In contrast, a child with some lexical knowledge would show a positive easy-hard effect. Mixing scores from two such populations would result in an overall positive correlation between the easy-hard effect and LNT performance.

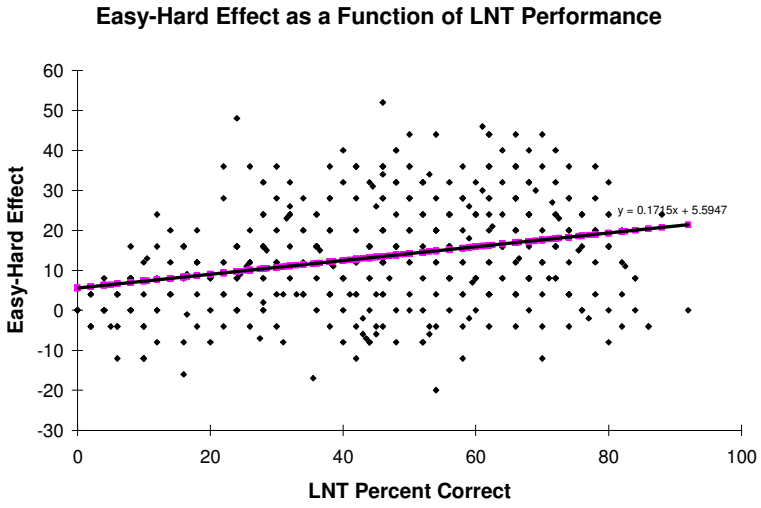


Figure 1. Size of the easy-hard effect as a function of percent correct on the LNT.

To control for floor and ceiling effects, we repeated the correlational analysis after first eliminating all tests in which the child scored 20% or less on the LNT (a total of 73 tests) and all tests in which the child scored above 80% correct on the LNT (a total of 12 tests). These corrections reduced the

correlation to 0.11. Although statistically significant, $t(356) = 2.09$, $p < .05$, the correlation is quite small and explains less than 1.25% of the total variance. Essentially, this result suggests that the size of the easy-hard effect is not correlated with overall LNT performance.

To help ensure that we were not missing a more subtle relation between the easy—hard effect and overall LNT performance, we selected from our master group all test intervals in which overall LNT performance was 21-40% correct (N=96), all those in which overall performance was 41-60% correct (N=142), and all those in which overall performance was 61 –80% correct (N=120). Figure 2 plots performance on LNT easy words and LNT hard words for these three performance intervals. The easy-hard effect for these three performance intervals was 13.38, 15.42, and 17.93, respectively. An analysis of variance showed that the increase in the effect with performance interval was reliable, $F(2, 355) = 3.19$, $p < .05$. Individual t tests showed that the easy-hard effect was smaller in the 21-40% correct group than in the 61-80% correct group, $t(214) = 2.62$, $p < .01$. All other pair-wise comparisons were non-significant.

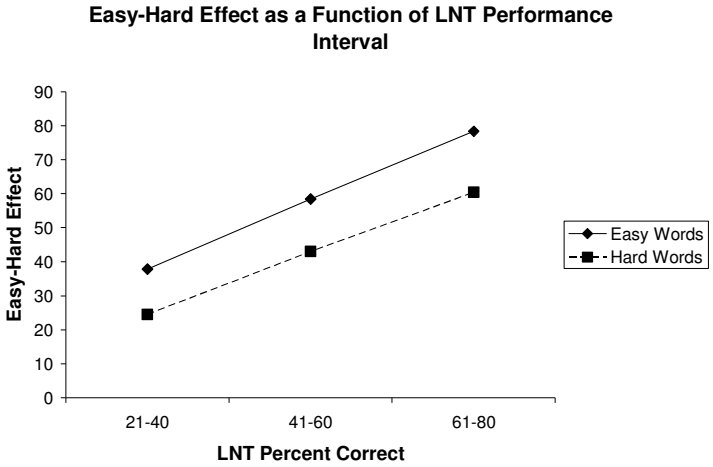


Figure 2. The easy—hard effect as a function of LNT performance interval.

The easy-hard effect is not, of course, arithmetically independent of the LNT Percent Correct, since the overall percent correct is simply the mean of the percent correct on easy words and the percent correct on hard words. Therefore, we also examined the relation between the easy-hard effect and performance on a different measure of lexicon size, PPVT raw scores. We first selected from our master group of 453 test intervals, the 358 on which the child had scored between 21 and 80% correct, inclusive, on the LNT. From this group, we then selected for additional analyses the 350 test intervals for which PPVT scores were also available. The overall correlation between LNT percent correct and PPVT raw scores was 0.177. Although statistically significant, $t(348)=3.35$, $p < .001$, the magnitude of the correlation is surprisingly low given that both tests purport to measure the size of a child's vocabulary.

Figure 3 shows a scatter plot of the easy—hard effect as a function of PPVT performance. The correlation between PPVT raw score and the size of the easy—hard effect was 0.116, $t(348) = 2.18$, $p < .05$. To further examine this effect, test records were divided into four sub-groups based on overall PPVT performance. The bottom 25 percent of records, based on PPVT performance, were assigned to Quartile 1, records in the next poorest performing 25% were assigned to Quartile 2, and so on, with the restriction that when the same PPVT score occurred in multiple test records, those records could not be

split across quartiles. The number of records in Quartiles 1-4 was 88, 90, 85, and 87, respectively. The mean PPVT raw score for Quartiles 1-4, respectively, was 40.35, 63.34, 83.99, and 125.39. Figure 4 shows percent correct on LNT easy words and LNT hard words as a function of PPVT quartile. The easy-hard effect for the four quartiles was 13.68, 14.00, 15.98, and 18.67, a significant increase across quartile, $F(3,346) = 2.66$, $p < .05$. Individual t-tests revealed that the easy-hard effect was significantly smaller in both Quartiles 1 and 2 than in Quartile 4, $t(173) = 2.65$, $p < .01$, and $t(175) = 2.40$, $p < .02$, respectively. The overall results of the PPVT analysis agree with the results of the overall LNT performance analysis. There is a small but statistically significant increase in the easy-hard effect as vocabulary size increases.

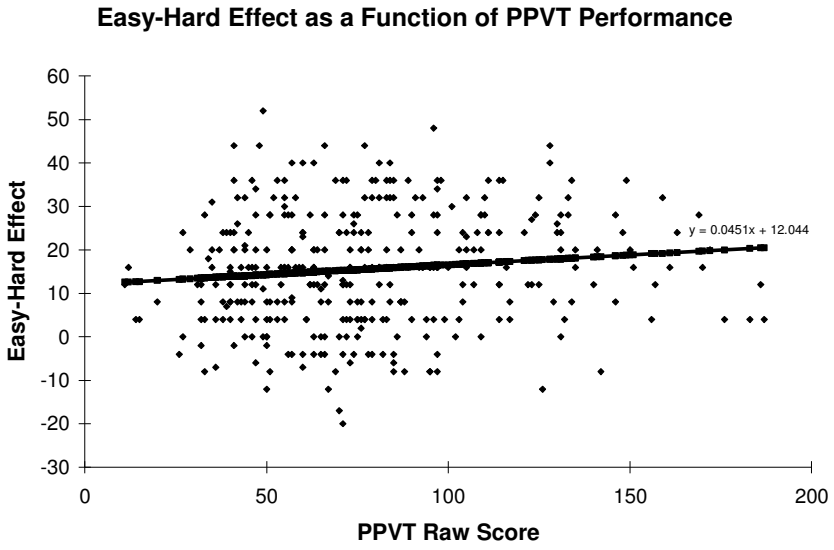


Figure 3. Scatter plot of the LNT easy—hard effect as a function PPVT raw score.

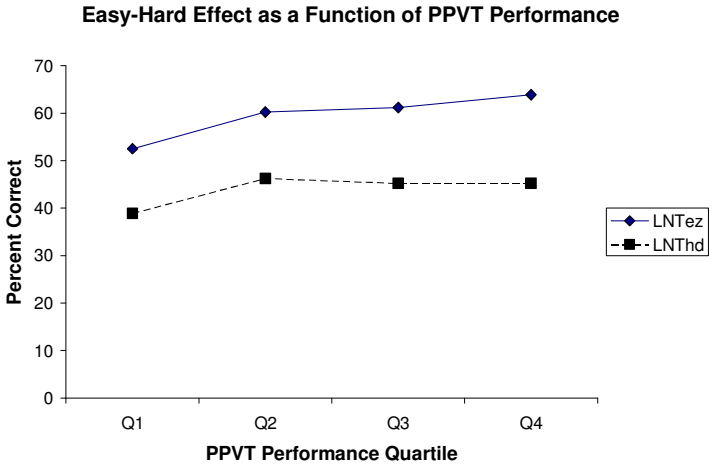


Figure 4. Percent correct on LNT easy and LNT hard words as a function of PPVT raw score performance quartile.

We also examined the easy-hard effect on the MLNT. Our master file of 443 test intervals included 213 on which the child was tested on the MLNT. Eighty-one individual children contributed these data. Figure 5 shows a scatter plot of the easy-hard effect on the MLNT as a function of overall performance on the LNT. Analyzing only those test intervals on which the child scored between 21% and 80% correct (N=162), inclusive, on the MLNT, we found no significant correlation between the MLNT easy-hard effect and overall performance on the LNT, r = -0.052, $t(160)$ = -0.66.[2]

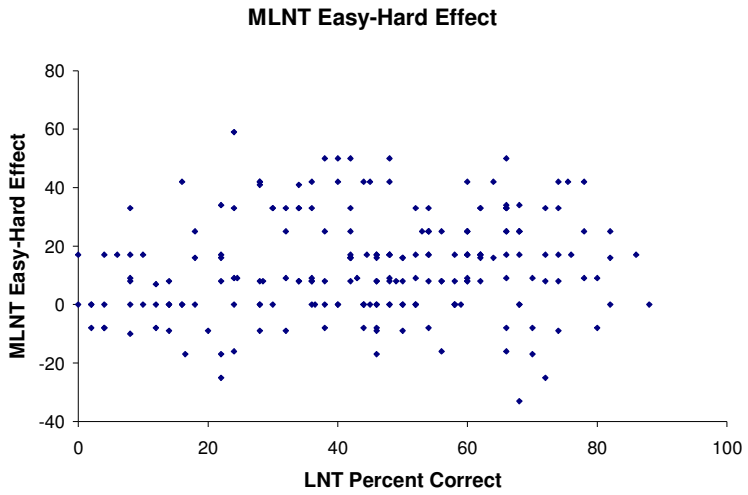


Figure 5. Scatter plot of the easy-hard effect as a function of overall percent correct on the LNT.

Figure 6 shows MLNT performance separately for easy words and hard words, broken down into four quartiles based on LNT performance percentile, with the restriction that test records with identical LNT scores could not be assigned to overlapping quartiles. An analysis of variance found no significant change in the easy-hard effect across these four performance intervals, $F(3, 158) = 1.34$. The MLNT easy-hard effect across the four quartiles, from lowest LNT performance to highest, was 15.32 N=40), 14.08 (N=39), 11.45 (N=40), and 19.05 (N=43).

We also analyzed the MLNT easy-hard effect contingent on PPVT performance in a manner similar to the analyses done for the LNT data. There were 158 test intervals for which both MLNT and LNT data were available. As was the case for the LNT data, the correlation between PPVT raw scores and MLNT percent correct was small but statistically significant, r = 0.189, $t(156)$ = 2.40, p <.02. The MLNT easy-hard effect did not significantly correlate with PPVT performance, r = 0.103, t(156) = 1.29. An analysis of variance of the easy—hard effect by PPVT quartile (N = 37, 41, 39, and 41 for Quartiles 1-4, respectively; mean PPVT raw score = 39.87, 63.46, 85.21, and 126.15 for Quartiles 1-4, respectively) also showed no significant effect of PPVT performance on the easy—hard effect, $F(3,154) < 1$. Across the four quartiles, the easy-hard effect was 11.86, 15.46, 15.41, and 17.61.

[2] Note that we are trimming the data to the 21% to 80% correct range using MLNT scores (in order to avoid floor and ceiling effects in the MLNT easy-hard effect, but we are using LNT scores as the basis for estimating the size of the child's lexicon. In our overall sample of MLNT tests (N = 214), performance on the LNT correlated extremely highly with performance on the MLNT, r = 0.894.

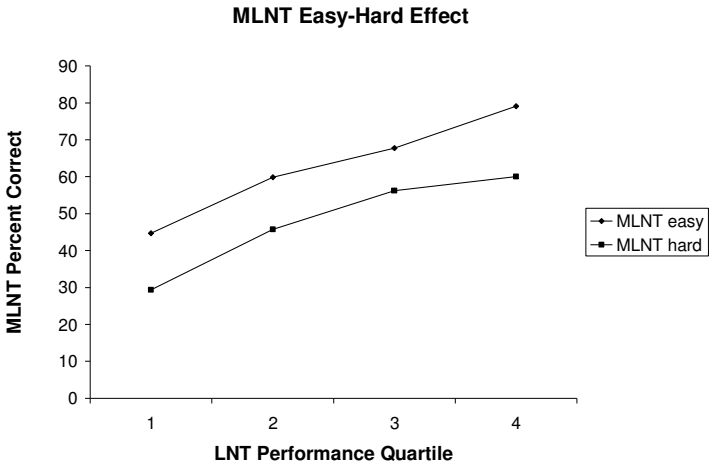


Figure 6. Percent correct on the MLNT easy words and MLNT hard words as a function of MLNT percent correct.

All of the above analyses included both children who used OC and children who used TC. It is quite possible that communication mode can affect the structure of the developing lexicon. In such a case, the easy-hard effect may be quite different in OC users than in TC users. Accordingly, we repeated the above analyses separately for each of these two communication modes. For TC users (N=113) performing in the overall range of 21 to 80% correct on the LNT, there was no significant correlation between the easy-hard effect and overall LNT performance, $r = 0.07$, $t(111) = 0.76$. For OC users (N = 245) performing in the same range, there was a small but statistically significant correlation, $r = 0.14$, $t(243) = 2.22$, $p < .05$. Note that the difference between these two correlations was itself non-significant.

Figure 7 shows LNT performance separately for easy words and hard words for OC and TC users, broken down into three bins of overall LNT performance: 21-40% correct (N = 43 for the OC group; N = 53 for the TC group), 41-60% correct (N = 104 for the OC group; N = 38 for the TC group), and 61-80% correct (N = 98 for the OC group; N = 22 for the TC group).

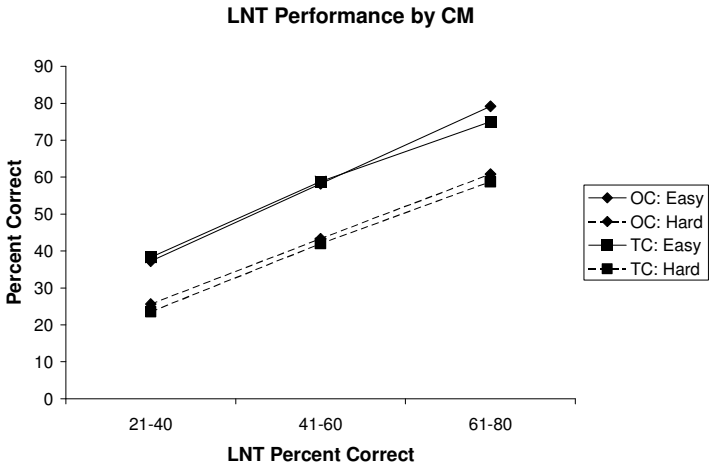


Figure 7: Percent correct as a function of LNT performance bin for easy and hard words for OC users and TC users.

Figure 8 explicitly shows the easy hard effect for the two CM groups. Overall, for the TC group, the easy-hard effect did not change size as a function of performance interval, $F(2, 110) < 1$. However, for the OC group, the easy-hard effect increased as LNT performance improved, $F(2, 242) = 3.98$, $p < .02$. Individual *t*-tests comparing performance bins in the OC group indicated that the easy-hard effect was marginally smaller in the 41-60% performance bin than in the 61-80% performance bin, $t(200) = 1.77$, $p < .10$, and significantly smaller in the 21-40% performance bin than in the 61-80% performance bin, $t(139) = 2.87$, $p < .005$. Comparisons of the OC and TC group at each individual performance bin were all non-significant, smallest $p = .21$.

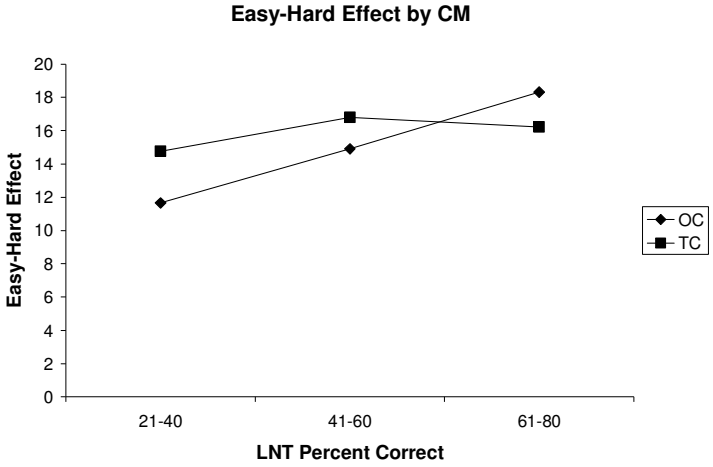


Figure 8. The easy-hard effect as a function of LNT performance bin for OC and TC users.

We also examined the relation between PPVT and LNT performance separately for OC children and TC children. For TC users, PPVT performance did not correlate with LNT performance, $r = 0.110$, $t(110) = 1.16$, or with the LNT easy-hard effect, $r = 0.056$, $t(110) = 0.59$. In contrast, for OC users, PPVT performance did significantly correlate with LNT performance, $r = 0.306$, $t(236) = 4.94$, $p < .001$, and with the LNT easy-hard effect, $r = 0.141$, $t(236) = 2.19$, $p < .05$. Both the OC group and the TC group were broken down into PPVT raw score performance quartiles. Table 2 shows PPVT performance for each of these quartiles for each CM group. Figure 9 shows the percent correct for LNT easy and LNT hard words as a function of PPVT performance quartile separately for the OC and TC groups.

	Q1	Q2	Q3	Q4
		OC Users		
N	60	59	60	59
Mean	37.53	58.54	79.20	114.20
Standard Deviation	9.13	6.32	6.03	41.07
		TC Users		
N	28	27	28	28
Mean	49.61	71.39	91.16	131.86
Standard Deviation	8.25	5.71	7.50	24.74

Table 2. N, Mean, and Standard Deviations for each of the four PPVT raw score quartiles for OC and TC users used in the LNT analysis.

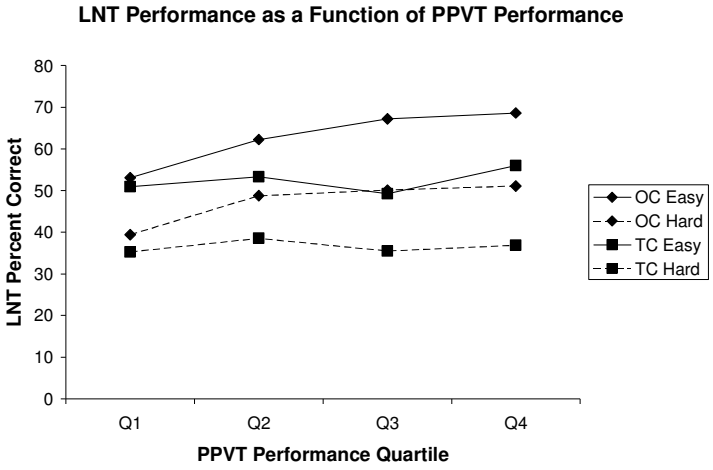


Figure 9. Percent correct on LNT easy and LNT hard words as a function of PPVT Performance Quartile for the OC users and the TC users.

Figure 10 shows the easy-hard effect separately for the OC and TC groups as a function of PPVT performance quartile. The easy-hard effect did not change significantly as a function of PPVT performance quartile for either the TC group, $F(3, 108) < 1$, or the OC group, $F(3, 234) = 1.53$ (despite the small positive correlation between PPVT performance and the easy-hard effect in the OC group).

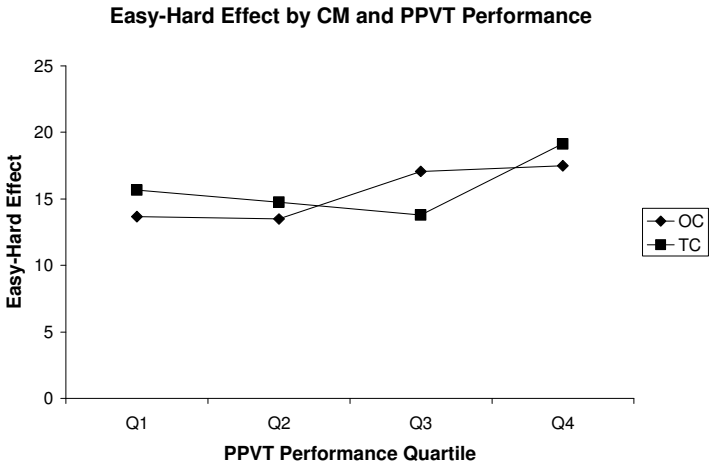


Figure 10. The Easy—Hard effect for OC and TC users as a function of PPVT Performance Quartile.

We also analyzed the easy-hard effect in the MLNT data separately for OC and TC users. MLNT data were available from 97 test intervals involving OC users whose overall MLNT percent correct was in the 21-80% range, and from 65 test intervals involving TC users whose overall MLNT percent correct was in the 21-80% range. For OC users, the correlation between the easy-hard effect for MLNT words and overall MLNT percent correct was -.08; the corresponding correlation for TC users was .01. Neither correlation was significant. Similarly, the easy-hard effect did not significantly change as a function of

performance bin (21-40%, 41-60%, 61-80%) for either the OC or TC group. For the OC group, the easy-hard effect across the three performance intervals was 16.44, 18.42, and 12.54, $F(2,94) = 1.19$. For the TC group, the easy-hard effect across the three performance intervals was 13.92, 18.89, and 12.61, $F(2,62) < 1$.

The MLNT data were also analyzed using PPVT performance as a measure of lexicon size. For OC users (N=93), PPVT performance correlated significantly with percent correct on the MLNT, $r = 0.321$, $t(91) = 3.23$, $p < .01$, but only marginally with the size of the easy-hard effect, $r = 0.178$, $t(91) = 1.78$, $.05 < p < .10$. For TC users (N=65), PPVT performance did not correlate significantly with either percent correct on the MLNT, $r = 0.168$, $t(63) = 1.35$, or with the size of the easy-hard effect, $r = 0.010$, $t(63) = 0.08$. The size of the easy-hard effect was also examined as a function of PPVT performance bin. Characteristics of PPVT raw scores for each performance bin are shown in Table 3, separately for OC users and TC users. To keep cell sizes reasonably large, PPVT performance was divided into thirds rather than quarters. For OC users, the size of the easy-hard effect from the worst performing to best performing PPVT tertile was, respectively, 11.87, 16.74, and 17.41, a non-significant effect, $F(2,90) < 1$. For TC users, the size of the easy-hard effect from the worst performing to best performing PPVT tertile was, respectively, 16.24, 13.05, and 15.17, also a non-significant effect, $F(2,90) < 1$.

	T1	T2	T3
	OC Users		
N	30	31	32
Mean	40.33	70.07	113.34
Standard Deviation	12.43	8.73	25.90
	TC Users		
N	21	21	23
Mean	51.29	78.71	123.17
Standard Deviation	8.99	7.62	31.39

Table 3. N, Mean, and Standard Deviations for each of the three PPVT raw score tertiles for OC and TC users used in the MLNT analysis.

Summary of the Effect of Lexicon Size on the Easy-Hard Effect. To the extent that either the LNT and MLNT or the PPVT are legitimate measures of lexicon size, there is no evidence that the size of the easy-hard effect **decreases** as the size of pediatric CI users' lexicons increases. On the contrary, there is some evidence that on the LNT the size of the easy-hard effect increases to a small amount with lexicon size. This effect appears to be limited to OC users. No such effect is evident in the data for TC users. For TC users, the size of the easy-hard effect seems to be fairly constant across difference lexicon sizes. On the MLNT, the size of the easy-hard effect also appears to be constant across different lexicon sizes for both OC and TC users.

Comparing the Easy-Hard Effect with the Word Length Effect

The second issue that we investigated concerned the origin of the word length effect. More specifically, we attempted to determine whether the word-length effect has the same underlying causes as the easy-hard effect. Accordingly, we examined how the word length effect varies with the easy-hard effect. In the overall sample of 213 test intervals for which data were available from both the LNT and MLNT tests, the correlation between the LNT easy-hard effect and the word length effect (MLNT

percent correct minus LNT percent correct) was 0.063, a non-significant relation. The overall correlation between the MLNT easy-hard effect and the word length effect was -.014, also a non-significant relation. In an effort to control for floor and ceiling effects, we repeated these correlations after eliminating all tests on which either LNT or MLNT overall percent correct was below 20% or above 80%. A total of 151 test intervals remained after eliminating these results. For this more restricted sample, the correlation between the LNT easy-hard effect and the word length effect was -0.072, a non-significant relation. The correlation between the MLNT easy-hard effect and the word-length effect was -0.135, a marginally significant **negative** correlation, $t(149) = 1.66$, $p < .10$.

We divided the sample of 151 tests where both LNT and MLNT performance was in the range 21-80% correct into those children using OC (N=91) and those using TC (N=60). For the OC group, the correlation between the easy-hard effect on the LNT and the word length effect was -0.151, $t(89) = 1.44$, n.s., and the correlation between the easy-hard effect on the MLNT and the word length effect was also -0.147, $t(89) = 1.40$, n.s. For the TC group, the easy-hard effect also failed to significantly correlate with the word length effect, $r = 0.077$, $t(58) = 0.59$ for the LNT easy-hard effect, and , $r = -0.117$, $t(58) = -0.89$ for the MLNT easy-hard effect.

To examine the question of whether the word length effect changes with lexicon size, changes in the effect with changes in PPVT raw scores were analyzed. The sample of 151 tests in which both LNT and MLNT performance was in the range of 21 to 80% correct included 147 tests which also included PPVT scores. There was no significant correlation between PPVT raw score and the word length effect, $r = 0.059$, $t(145) = 0.71$. Further, no correlation emerged when the data were analyzed separately for OC users, $r = 0.043$, $t(85) = 0.40$, and TC users, $r = 0.105$, $t(58) = 0.80$. Finally, the word length effect did not change significantly as a function of PPVT performance tertile in either the OC group, $F(2,84) < 1$, or in the TC group, $F(2,57) = 1.21$. Across PPVT performance tertile, in the OC group, the word length effect was 7.07, 10.35, and 8.24. In the TC group, the word length effect was 4.73, 9.45, and 8.08 across PPVT tertile.

Summary of Results on the Word Length Effect. In summary, there is little evidence that the easy-hard effect and the word length effect are correlated with one another. In addition, there is little evidence that the word length effect changes with growth in the size of the lexicon.

Comparing OC and TC Users

Our primary interest in comparing OC and TC users was to determine if the two classes of users showed different patterns of easy—hard or word length effects at a given lexicon size. Accordingly, from our master set of test records we selected a sample that matched lexicon size in OC and TC users. We used PPVT raw scores as a measure of lexicon size. From our master group of 443 test records, we first eliminated those for which no PPVT data were available, leaving 152 TC users and 281 OC users. For each test record contributed by a TC child, we then selected a test record contributed by an OC child with an identical PPVT raw score. If no OC child had an identical PPVT raw score, we selected that record with the nearest PPVT raw score. If multiple OC records had an identical PPVT raw score, we chose a record from amongst those with the identical score in a pseudo-random fashion. Records were selected from OC users without replacement. Table 4 shows demographic characteristics of the two groups. Note that TC children were marginally older than OC children, were fitted with a CI implant at a later age, and had a later age of onset of deafness. These characteristics were true of our master sample as well.

	OC	TC
PPVT Raw Score	82.33	82.02
Interval	9.78	9.54
CA	105.25	112.05#
Fitted	46.48	54.41***
Onset	6.95	13.26*

Table 4. Demographics of the OC and TC matched groups. Interval refers to the 6-month period after implant at which the test was conducted. (Interval 1 is 6 months post-implant, interval 2 is 12 months post-implant, and so on.) CA is chronological age, Fitted is age at which the child received the cochlear implant, and Onset is the age at onset of deafness. CA, Fitted and Onset are all in months. #: $p < .10$. * $p < .05$. *** $p < .005$.

Figure 11 shows overall performance of the OC children and TC children on the LNT, separately for easy and hard words. Overall, OC users performed better than TC users on the LNT, $F(1, 302) = 84.46$, $p < .0001$. Percent correct was higher on easy words than on hard words, $F(1,303) = 336.36$, $p < .0001$. The size of the easy—hard effect (15.36 for the OC group; 12.86 for the TC group) was at best marginally larger in the OC group than in the TC group, $F(1, 302) = 2.66$, $p = .10$.

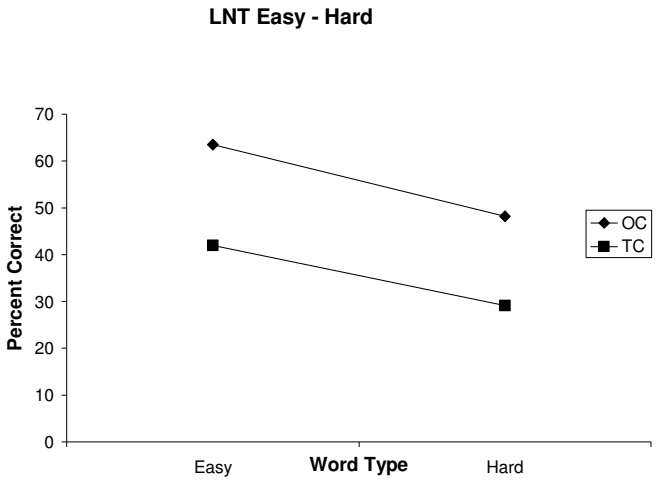


Figure 11. LNT performance as a function of word type (easy vs. hard) for OC users and TC users.

In order to avoid possible misleading results due to floor or ceiling effects, we repeated the above analyses but including only matched records where the overall LNT test score was between 21 and 80 percent correct, inclusive. We began with the 112 TC users who scored in this range and then matched each of those users with an OC user who also scored within that range, using the same matching procedures used for the larger group. Table 5 shows some demographic characteristics of these sub-groups. Although the two groups did not significantly differ on chronological age, TC users were fitted with a CI at a later age than OC users and were marginally older at the age of onset of deafness than OC users.

	OC	TC
PPVT Raw Score	86.63	86.45
Interval	10.39	10.13
CA	109.48	115.93
Fitted	47.08	54.74*
Onset	8.36	14.00#

Table 5. Demographics of the OC and TC matched groups who performed between 21% and 80% correct on the LNT. Interval refers to the 6-month period after implant at which the test was conducted. (Interval 1 is 6 months post-implant, interval 2 is 12 months post-implant, and so on.) CA is chronological age, Fitted is age at which the child received the cochlear implant, and Onset is the age at onset of deafness. CA, Fitted and Onset are all in months. #: $p < .10$. * $p < .05$

Figure 12 shows performance on the LNT for these sub-groups. OC users scored higher on the LNT than TC users, $F(1, 222) = 42.64$, $p < .0001$. Percent correct was higher on easy words than on hard words, $F(1,223) = 326.89$, $p < .0001$. The size of the easy-hard effect for the two groups did not differ, $F(1,222) < 1$.

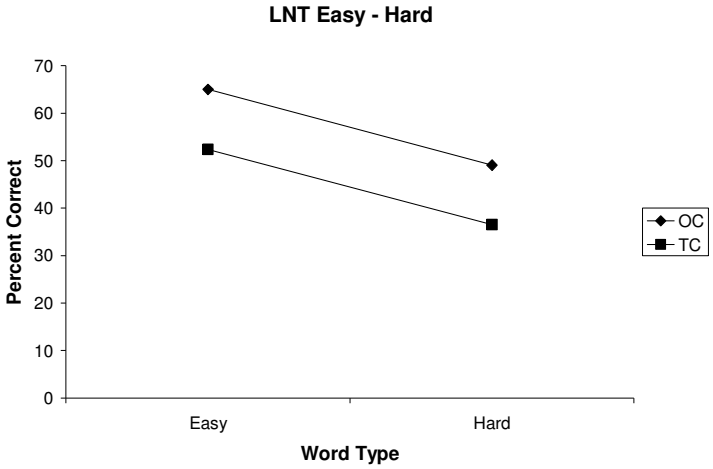


Figure 12. LNT performance as a function of word type (easy vs. hard) for OC users and TC users scoring between 21 and 80 percent correct on the LNT..

As mentioned, our intent had been to control for lexicon size by matching on PPVT raw score. This procedure seemed preferable to matching on LNT percent correct and then examining variations in the easy-hard effect since PPVT raw scores are arithmetically independent of the easy-hard effect but LNT overall performance is not. However, given that the OC children consistently performed at a higher percent correct on the LNT than did the TC children, an argument could be made that, to the extent that the LNT is a better measure of lexicon size than is the PPVT, we had not in fact matched the two sub-groups for lexicon size. Accordingly, we conducted an initial analysis in which we matched OC and TC children on overall LNT percent correct. To avoid spurious results due to floor and ceiling effects, this analysis was limited to those children who scored between 21 and 80 percent correct, inclusive, on overall on the LNT. Because of the overall lower performance of the TC children, it was not possible to match all TC test records with a corresponding OC test record. We first selected for inclusion in this sub-sample all TC test records for which there was an OC test record with an identical overall percent correct

on the LNT. We then selected from the remaining records all those pairs of OC and TC test records within 2% of each other on overall percent correct on the LNT, with the stipulation that for every pair in which the OC record had a higher score, another pair had to be included in which the TC record had the higher score. The final sub-sample consisted of 184 records, 92 from OC children and 92 from TC children.

Table 6 shows some demographic characteristics of this sub-sample, as well as the size of the easy-hard effect for the two CM groups. The important result was that the easy-hard effect for the OC children did not significantly differ from that of the TC children, $F(1, 182) < 1$.

	OC	TC
LNT Percent Correct	47.49	47.49
Interval	9.29	9.93
CA	103.92	113.60#
Fitted	48.55	53.63
Onset	7.25	15.27*
Easy-Hard Effect	14.89	15.86

Table 6. Demographics of the OC and TC matched groups who performed between 21% and 80% correct on the LNT. These groups were matched on overall LNT Percent Correct. Interval refers to the 6-month period after implant at which the test was conducted. (Interval 1 is 6 months post-implant, interval 2 is 12 months post-implant, and so on.) CA is chronological age, Fitted is age at which the child received the cochlear implant, and Onset is the age at onset of deafness. CA, Fitted and Onset are all in months. #: $p < .10$. * $p < .05$

We performed a similar set of analyses on the MLNT data, beginning with an analysis where we matched records based on PPVT raw scores. For those children who contributed MLNT data, we matched each TC child with an OC child using the same procedure as used for the LNT data. Each resulting sub-group had a total of 83 test records. Table 7 shows some demographic characteristics of this group. Figure 13 shows the percent correct of these children on the MLNT, broken down by CM and Easy and Hard words. The MLNT results paralleled those for the LNT. OC children correctly identified more words on the MLNT than did the TC children, $F(1, 164) = 28.09$, $p < .0001$. Easy MLNT words were identified correctly more frequently than hard MLNT words, $F(1, 165) = 108.62$, $p < .0001$. The size of the easy-hard effect did not differ significantly between the two CM groups, $F(1, 164) < 1$.

	OC	TC
PPVT Raw Score	83.93	83.83
Interval	9.35	9.42
CA	110.64	113.35
Fitted	54.36	56.25
Onset	7.02	18.66***

Table 7. Demographics of the OC and TC MLNT groups matched on PPVT raw scores. Interval refers to the 6-month period after implant at which the test was conducted. (Interval 1 is 6 months post-implant, interval 2 is 12 months post-implant, and so on.) CA is chronological age, Fitted is age at which the child received the cochlear implant, and Onset is the age at onset of deafness. CA, Fitted and Onset are all in months. *** $p < .005$.

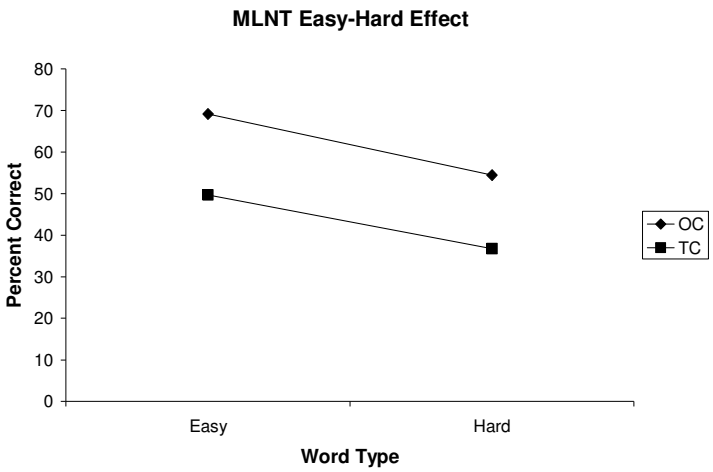


Figure 13. Percent correct for easy and hard MLNT words for OC and TC users matched on PPVT scores.

Figure 14 shows the word length effect for these two sub-groups. Overall, percent correct was higher on the MLNT than on the LNT, $F(1, 330) = 9.18$, $p < .005$. Although this word length effect was slightly larger in the OC group than in the TC group, the difference was not statistically significant, $F(1, 164) = 1.53$.

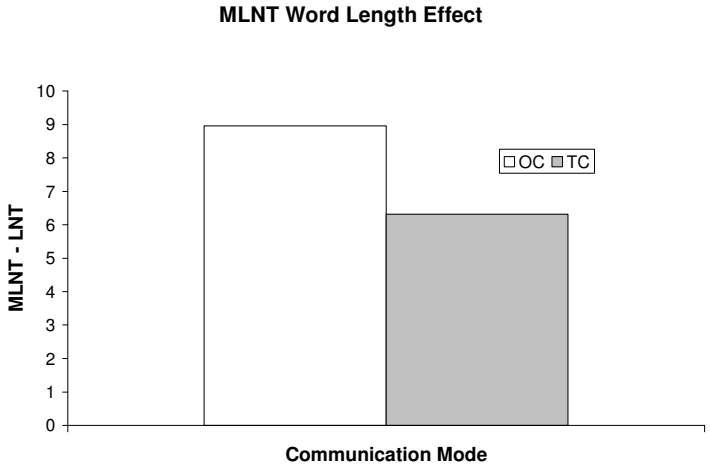


Figure 14. Word length effect for OC and TC users matched on PPVT scores.

Following the same rationale as for the LNT tests, we also matched MLNT test records on LNT scores instead of PPVT scores, following the same procedure as for LNT scores. This procedure resulted in 63 test records in the OC group and 63 in the TC group. Demographic characteristics of this sub-group, as well as the word length effect, are shown in Table 8. When matched on LNT performance, OC children did not score significantly higher on the MLNT than did TC children, $F(1, 124) < 1$. Easy words were identified more accurately than hard words, $F(1, 125) = 77.79$, $p < .001$. The size of the easy-hard effect on the MLNT, however, did not differ significantly between the two CM groups, $F(1, 124) = 1.10$. Overall, percent correct was higher on the MLNT than on the LNT, $F(1, 250) = 10.18$, $p < .005$.

Although this word length effect was slightly larger in the OC group than in the TC group, the difference was not statistically significant, $F(1, 124) = 1.10$.

	OC	TC
LNT Percent Correct	40.48	40.51
Interval	49.78	54.03
CA	105.79	111.92
Fitted	56.02	57.89
Onset	5.62	21.51***
Word Length Effect	10.01	7.96

Table 8. Demographics of the OC and TC groups contributing MLNT data and matched on overall LNT percent correct. Interval refers to the 6-month period after implant at which the test was conducted. (Interval 1 is 6 months post-implant, interval 2 is 12 months post-implant, and so on.) CA is chronological age, Fitted is age at which the child received the cochlear implant, and Onset is the age at onset of deafness. CA, Fitted and Onset are all in months. *** $p < .001$.

We also analyzed the MLNT data matched on LNT scores after eliminating all pairs of test records in which either the OC record or the TC record had an LNT percent correct outside the range of 21-80, inclusive. The two resulting sub-groups each had 50 test records. This manipulation did not change the basic pattern of results. Overall percent correct on the MLNT did not differ between OC and TC children, $F(1, 98) < 1$. MLNT easy words were identified more accurately than hard words, $F(1, 99) = 166.05$, $p < .001$. The size of the easy-hard effect, however, did not differ significantly between the two CM groups, $F(1, 98) < 1$. Overall, percent correct was higher on the MLNT than on the LNT, $F(1, 198) = 20.79$, $p < .0001$. Although this word length effect was slightly larger in the OC group (Mean = 11.16) than in the TC group (Mean = 9.43), the difference was not statistically significant, $F(1, 98) < 1$.

Summary of Results on Communication Mode. OC children and TC children both show robust easy—hard and word length effects. When test records are matched on lexicon size, the size of these effects does not significantly differ across users of the two communication modes. When PPVT is used to match on lexicon size, OC children perform better than TC children on both the LNT and MLNT.

Discussion

The main results of the present study can be summarized as follows:

First, for OC users, the size of the easy-hard effect on the LNT increases, albeit by a small amount, as lexicon size increases. The increase occurs regardless of whether lexicon size is measured by LNT overall percent correct or PPVT raw scores. In contrast, for TC users, the size of the easy—hard effect does not change as lexicon size increases.

Second, unlike the results for the LNT, the size of the easy-hard effect on the MLNT does not change as lexicon size increases for either OC or TC users. This result holds regardless of whether lexicon size is measured by overall LNT percent correct or by PPVT raw scores.

Third, the size of the easy-hard effect does not significantly correlate with the size of the word length effect for either OC users or TC users.

Fourth, the size of the word length effect does not appear to change as lexicon size increases, for either OC users or TC users, where lexicon size was measured by PPVT raw scores.[3]

Fifth, when matched on PPVT raw scores, OC children perform better than TC children on both the LNT and MLNT.

We begin by first attempting to explain the pattern of results concerning the easy-hard effect. Suppose that CI children assume that what they are hearing in the LNT (and MLNT) are spoken words, and that they therefore repeat back the closest matching word in their lexicon. Suppose further that, due to the acoustic "noise" introduced by the CI, CI children begin with relatively undifferentiated, coarse phonetic representations. Many words that are similar sounding yet clearly different words to an adult normal hearer are in fact mapped to the same representation in the mental lexicon of CI users. The CI user, for instance, might map the words /kæt/, /kæp/, and /kæn/ to a single representation, say that corresponding to /kæp/. When presented with any of these three words in a word repetition task, the child will respond with /kæp/. Note that this process essentially amounts to a guessing strategy. Such a guessing strategy is more likely to be correct for words with few phonetically similar words (i.e., easy words) than for words with many phonetically similar words (i.e., hard words). The result is a robust easy-hard effect. Over time, as the child gains experience with the CI, finer phonetic discriminations are learned and the mental lexicon becomes correspondingly more differentiated. The words /kæt/ and /kæp/ may still map to the same lexical representation, but the word /kæn/ now maps to its own representation. In effect, on a word repetition task, the child is now guessing from amongst fewer alternatives, resulting in overall improved performance. When a new discrimination is learned, it is more likely to split a group of hard words into two smaller equivalence classes than it is to split a group of easy words into two smaller groups, simply because the group of hard words is larger to begin with. The result is that a) overall lexicon size should appear larger since the child is making more phonetic discriminations, and b) performance on hard words should improve faster than performance on easy words. Thus, as lexicon size increases, the size of the easy-hard effect should decrease. This prediction is clearly inconsistent with our data.

An alternative perspective may make these data more comprehensible. While the child's lexicon is developing, it may simply be the case that many words that in the more mature lexicon would have multiple neighbors have in fact many fewer neighbors. Statistically, the less mature lexicon, with fewer overall words, is likely to show reduced neighborhood density differences between easy and hard words. This reduction in turn would mitigate easy-hard effects. As the lexicon grows in size and connectivity, so do neighborhood density differences and therefore so does the easy-hard effect. That is, the easy-hard effect should increase as the size of the mental lexicon increases. This prediction is consistent with our findings for OC children on the LNT.

This second alternative can be stated somewhat differently. The key insight of word recognition models such as those proposed by Luce and Pisoni (1998) and Auer and Luce (2005) is that words are recognized in the context of other, similar sounding words. According to these models, a listener's mental lexicon is organized into similarity neighborhoods of interconnected words, with some neighborhoods being more densely populated than others. The easy-hard effect is a consequence of this organization into similarity neighborhoods. To the extent that a listener has no interconnections among the words in the lexicon, no easy-hard effect emerges. As those lexical connections develop and become

[3] Parallel analyses using overall LNT percent correct as a measure of lexicon size were not performed. Since the word length effect is defined as MLNT percent correct minus LNT percent correct, there will be a tendency for a negative correlation to emerge with LNT even if MLNT and LNT performance are completely independent.

richer, the mental lexicon organizes itself into similarity neighborhoods and easy-hard effects emerge. Hence, larger easy-hard effects would be expected as the CI user's lexicon matures.

Why then is the increase so small and limited to OC children on the LNT? The answer to the second part of the question—why the increase is limited to OC children—may well be that only these children, because they rely much more on phonetic information for language communication, form the more richly interconnected lexical space necessary for similarity neighborhoods, and hence easy-hard effects, to emerge. The answer to the first part of the question—why the increase is so small—may well be that both the above alternatives may be operating, with observed performance simply the averaged effects of these two opposing processes. Increased phonetic differentiation may be causing decreased easy-hard effects while growth in lexicon size is causing increased easy-hard effects. To the extent that the two opposing trends are of approximately equal magnitude, they cancel one another out and the overall effect is little or no change in the easy-hard effect as the lexicon grows.

There is one aspect of our data which is not, at least superficially, entirely consistent with this hybrid proposal. In particular, it is only with OC children that we observed an increased easy-hard effect, indicating that growth in lexicon size was playing a larger role with them than phonetic differentiation. However, since it is precisely OC children who in their daily lives are presumably forced to make greater use of auditory cues for spoken word recognition, it is precisely these children that we would expect to show increased phonetic differentiation and hence smaller easy-hard effects, since words like /kæt/ and /kæp/ would now have separate representations. Similarly, when we matched OC and TC users on overall LNT percent correct or PPVT raw scores, we might have expected reduced easy-hard effects in the OC group. It is, of course, possible that by the time we began testing these children, the OC users' powers of phonetic discrimination were already relatively advanced. Consequently, these children were in fact more likely to show the effects of a growing lexicon. There is, however, nothing in our data that would support this *ad hoc* explanation.

There is a second limitation to this hybrid proposal. It posits a coincidence (The two effects just happen to nearly balance one another out.) and it runs the risk of non-falsifiability (In any given study, any pattern of change in the easy-hard effect could be explained by appealing to different strengths of the two processes.). Clearly, before it can claim to have received strong support, additional research is necessary that attempts to tease apart the two effects. Do, for example, children who show a poor ability to discriminate phonemes (as measured, for example, by performance on an ABX discrimination task using minimal-pair nonsense syllables) also show larger easy-hard effects?

Our second group of results concerns the word length effect and its relation to the easy-hard effect. Like the easy-hard effect in most cases, the word length effect did not show significant change as the size of the lexicon grew. The more important result is that the easy-hard effect and the word length effect do not correlate with one another. That these two effects appear independent of one another suggests that they have different causes. Easy words come from sparser neighborhoods than hard words and neighborhood density is thought to at least in part account for the easy-hard effect (e.g., Kirk et al., 1995). Longer words also come from sparser neighborhoods than shorter words. Hence, neighborhood density could also underlie the word length effect. That it is independent of the easy-hard effect, however, suggests that we may need to look elsewhere than density for an explanation of the word length effect. The obvious alternative is that the CI users are sensitive to the syllabic structure of words, and use the number of syllables as a word recognition cue, at least when identifying isolated words. Because there are fewer multi-syllabic words than monosyllabic words in the child's lexicon, (Logan, 1992), knowing the number of syllables in the word limits possible identifications to a smaller universe for

longer words than it does for shorter words. The result would be better performance in recognizing multi-syllabic words, independent of neighborhood density.

The final result meriting comment concerns the finding that, when OC and TC children are matched on PPVT scores, the OC children outperform the TC children on both the LNT and the MLNT. These results are reminiscent of those of Sommers et al. (1997). One obvious difference between the PPVT, on the one hand, and the LNT and MLNT on the other hand is that the PPVT is a closed-set task whereas the LNT and MLNT are both open-set tasks. Sommers et al. found that normal hearing adults listening in the clear, normal hearing adults listening under conditions of noise, and adult CI users all showed effects of talker variability (identifying words in lists with multiple talkers compared to identifying words in lists spoken by a single talker) and lexical difficulty (easy words and hard words as defined here) only on open-set tasks. Closed set tasks were not sensitive to these manipulations even when the response alternatives were designed to be maximally similar to the correct response. These results were recently confirmed by Clopper, Tierney, and Pisoni (2003) under somewhat different listening conditions. A similar effect is likely happening in the present study. The closed-set PPVT is simply not sensitive enough to reveal differences between the OC and TC populations of CI users. This result emphasizes the importance of including open-set tasks in test batteries investigating differences in the ability of different CI user populations in identifying spoken words.

In summary, we found some evidence that for OC users the size of the easy-hard effect on the LNT increases as the size of the child's lexicon increases. This result could reflect increased connectivity in the child's lexicon. As the lexicon grows, similarity neighborhoods begin forming and neighbors begin competing with one another for recognition. This result was not apparent for TC users. Likewise, we found little or no change in the easy-hard effect on the MLNT as lexicon size increased, and little or no change in the size of the word length effect. Overall, both the size of the easy-hard effect and the size of the word length effect seem remarkably stable over a wide range of lexicon sizes.

References

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*, 802-814.

Auer, E.T., Jr., & Luce, P.A. (2005). Probabilistic phonotactics in spoken word recognition. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception* (pp. 610-630). Oxford: Blackwell.

Bell, T.S., & Wilson, R.H. (2001). Sentence recognition materials based on frequency of word use and lexical confusability. *Journal of the American Academy of Audiology, 12,* 514-522.

Clopper, C.G., Tierney, A.T., & Pisoni, D.B. (2003). Effects of response format on speech intelligibility in noise: Results obtained from open-set, closed-set and delayed response tasks. *Journal of the Acoustical Society of America, 113*, 2254.

Eisenberg, L.S., Martinez, A.S., Holowecky, S.R., & Pogorelsky, S. (2002). Recognition of lexically controlled words and sentences by children with normal hearing and children with cochlear implants. *Ear & Hearing, 23*, 450-462.

Elliot, L.L., Clifton, L.A.B., & Servi, D.G. (1983). Word frequency effects for a closed-set identification task. *Audiology, 22*, 229-240.

Eukel, B. (1980). A phonotactic basis for word frequency effects: Implications for automatic speech recognition. *Journal of the Acoustical Society of America, 68*, S33.

Greenberg, J.H. & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157-177.

Gruenenfelder, T.M., & Pisoni, D.B. (2005). Modeling the mental lexicon as a complex system: Some preliminary results using graph theoretic measures. In *Research on Spoken Language Processing Progress Report No. 27* (pp. 27-47). Bloomington, IN: Speech Research Laboratory, Indiana University.

Howes, D.H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America, 29*, 296-305.

Kirk, K.I., Eisenberg, L.S., Martinez, A.S., Hay-McCutcheon, M. (1999). Lexical neighborhood test: Test-retest reliability and interlist equivalency. *Journal of the American Academy of Audiology, 10*, 113-123.

Kirk, K.I., Pisoni, D.B., & Miyamoto, R.C. (1997). Effects of stimulus variability on speech perception in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research, 40*, 1395-1405.

Kirk, K.I., Pisoni, D.B., & Osberger, M.J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16*, 470-481.

Kucera, F. & Francis, W. (1967). *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Landauer, T.K. & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*, 119-131.

Logan, J.S. (1992). *A computational analysis of young children's lexicons.* (Research on Spoken Language Processing Technical Report No. 8). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1-36.

MacWhinney, B., & Snow, C. (1995). The child language data exchange system. *Journal of Child Language, 12*, 271-296.

Marslen-Wilson, W.D. (1987). Parallel processing in spoken word recognition. *Cognition*, *25*, 71-102.

Marslen-Wilson, W.D. (1989). Access and integration: Projecting sound onto meaning. In W.D. Marslen-Wilson (Ed.), *Lexical access and representation*. Cambridge, MA: Bradford, 3-24.

McClelland, J.L. & Elman, J.L (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189-234.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words* (Research on Spoken Language Processing Technical Report No. 10). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31*, 273-279.

Savin, H.B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America, 35*, 200-206.

Sommers, M.S., Kirk, K.I., & Pisoni, D.B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear & Hearing, 18*, 89-99.

Treisman, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review*, *85*, 525 – 570.

Treisman, M. (1978b). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, *17*, 37-59.

Vitevitch, M.S. & Luce, P.A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374-408.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Effects of Clustering Coefficient on Spoken Word Recognition[1]

Nicholas A. Altieri and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This study was supported by the NIH Grants DC-00111 and DC00012. I would like to acknowledge Thomas Gruenenfelder for helpful comments, ideas and data analysis.

Effects of Clustering Coefficient on Spoken Word Recognition

Abstract. Since the late 1960's, researchers have explored how the structure of the mental lexicon affects spoken word recognition. The proposal that words are recognized relationally in the context of other words in lexical memory has encouraged the use of complex systems to describe connectivity in both the phonological and semantic lexicon. The present study assessed the role of the graph theoretical measure of *Clustering Coefficient* (CC) using two experimental paradigms: same-different discrimination, and perceptual identification to explore how global lexical variables affect spoken word recognition. In Experiment 1, listeners judged whether two words were the same or different. Longer response latencies were obtained for high CC words than low CC words. In Experiment 2, the stimuli were processed with an 8-channel noise vocoder to degrade the signal and a new group of listeners also performed the same-different task. In contrast to findings obtained in Experiment 1, listeners discriminated high CC words more accurately than low CC words. In Experiment 3, an open-set perceptual identification task was carried out to examine correct and incorrect responses using 8, 10 and 12 channels. Listeners identified low CC words more accurately than high CC words in the 10 and 12 channel conditions, but not in the 8-channel condition. Detailed analysis of the incorrect responses revealed that listeners used different perceptual strategies as the number of channels increased. These results suggest that global, emergent factors reflecting the structural organization and connectivity of words in the mental lexicon affect spoken word recognition and should be included in current models of spoken word recognition and lexical access.

Introduction

Structural analyses of the mental lexicon have been used to explore how listeners process and store spoken words. Numerous researchers have focused their efforts on understanding the structural properties and topology of the phonological mental lexicon following the publication of Oldfield's seminal article "Things, Words, and the Brain" (1966). One of the findings that motivated inquiries into the structure of the mental lexicon were word frequency effects; that is, effects demonstrating a systematic relationship between the frequency of a word's occurrence in the language and its intelligibility in noise (Howes, 1957). In order to explain word frequency effects, Oldfield (1966) hypothesized that the lexicon had a structure where words are organized into separate compartments or bins depending on their frequency of occurrence in the language. During the recognition process, the compartments are searched using a binary search algorithm. Oldfield assumed that memory bins containing high frequency words were searched first, which could explain why response latencies were shorter for common words in both *naming* and *same-different* judgment tasks.

Other models of the mental lexicon have also been proposed to account for frequency effects. Morton's Logogen Theory (1979) assumed that high frequency words have lower recognition thresholds than low frequency words in the lexicon. Logogens are not words, but are hypothetical units that act as evidence accumulators during the word recognition process. The term *logogen* literally means "word birth" (*logos* meaning words, *genus* for birth). According to Morton, evidence from both auditory and visual input is analyzed by logogens. Both bottom-up sensory evidence and top-down contextual information interact to bring the evidence accumulators above threshold. In general, the more contextual information that is available, the less bottom-up sensory information is required to bring the logogen above the critical threshold for word recognition. Once sufficient evidence has been accumulated to surpass a specified threshold, word recognition is assumed to occur. Logogen theory accounts for word-

frequency effects by assuming that less sensory evidence is required for high frequency words than low frequency words during the recognition process.

One shortcoming of Logogen theory is its inability to explain how pseudowords and non-words are recognized. In Logogen theory, the lexicon consists only of words with critical threshold values, and counters that accumulate evidence from features. However, the lexicon does not contain information about non-words, and exactly how logogens can accumulate evidence required for the recognition of phonologically possible non-words is not addressed.

In an effort to mathematically formalize predictions regarding the structure of the lexicon with respect to word frequency effects, Treisman (1979) considered three ways words might be represented in memory: as a *tree*, as a *collection*, or as a multidimensional *space*. An example of a tree structure is the way that words are organized in the dictionary. Locating a word in the lexicon involves finding the branch corresponding to the first sound of the word, then the second branch corresponding to the second sound of the word, and so on until a word is identified on a terminal twig. Degraded stimuli might cause the search to finish before ending at a terminal twig corresponding to an entire word in the lexicon because details of certain features might be missing.

Treisman described the "urn model" of Pollack, Rubenstein and Decker (1960) as an analog of a *collection* model of the lexicon. Each ball in the urn represents a word, and the number of balls in an urn that correspond to a word is proportional to the frequency of occurrence of that word in the language. The words organized in a collection have no fixed relation to one another. A model that describes how spoken word recognition occurs over a collection of lexical representations is the *universal forced choice model* (e.g., Luce, 1959). Luce's choice model assumes that decisions about word recognition are made using information from all of the words represented in the mental lexicon. According to this model, high frequency words will be identified more accurately because the probability of selecting a word is proportional to its frequency of occurrence. However, since high frequency words have more representations in the lexicon, they will also have a higher probability of being generated as error responses than low frequency words. Thus, the model assumes that the correlation between frequency of occurrence in the language and the frequency with which a word is generated as an error, will be strong because both correct and incorrect responses are selected based on a frequency weighted bias.

Words residing in a *multidimensional space*, on the other hand, are classified along continuous acoustical dimensions and are represented as points in the space, whereas non-words are represented as holes in the space. Identifying a word in the continuous space involves finding unique values on the relevant dimensions; if a unique value cannot be identified on a particular dimension, a range of uncertainty remains. Treisman outlined several predictions derived from *partial identification theory*, which hold for a model of the mental lexicon as a multidimensional space, but not a collection or a tree. Partial identification theory follows Luce's choice rule in assuming that words are identified in a forced-choice decision process, but differs inasmuch as the choice is taken to be limited to a subset of words lying in the "acoustic sub-volume defined by the stimulus" (Treisman, 1979).

The idea of an acoustic sub-volume is similar to the notion of a phonological neighborhood, which is a portion of the subspace of the lexicon containing phonologically similar words. The predictions derived from partial identification theory assert that the correlation between word frequency, and the frequency with which a word is generated as an error will be weak, and that words which are infrequently given as errors will be more easily recognized than words that are more frequently given as errors. To see why this is the case, consider a high frequency word with few similar sounding words residing in its sub-space. Since it has few neighbors, it is less likely to be generated as an error compared to words that have many similar sounding neighbors. This principle becomes even stronger as the signal-

to-noise ratio increases and the sub-volume of similar sounding words becomes smaller and more refined. Treisman provides evidence supporting all three hypotheses and concludes that a multidimensional space is a more accurate representation of the mental lexicon than either a tree or a collection model.

The urn and tree models of the mental lexicon also rely on the principle of "structural equivalence"— which assumes that the phonological characteristics of high and low frequency words are basically the same (Pollack et al., 1960). The principle of structural equivalence assumes that since the phonemic compositions of high and low frequency words are identical, the only difference between them is experienced frequency in the language. If the lexicon is modeled as continuous space with varying degrees of lexical density (the number of non-word holes differs depending on where we are in the space) then the composition of high and low frequency words could differ with regard to the number of similar sounding words in their acoustical subspaces.

Connectivity and Sub-Lexical Components

Landauer and Streeter (1973) assessed the widely held assumption of "structural equivalence" by showing that high and low frequency words differ in the number of similar sounding phonological neighbors they have, and the distribution of component phonemes. In a computational study using 260 high frequency and 260 low frequency words, they found that high frequency words tend to have more lexical neighbors than low frequency words. That is, high frequency words were phonetically similar to many other high frequency words. Landauer and Streeter defined "lexical neighbor" as a word that can be created from a target word by a single deletion, addition or substitution (DAS) of a letter. In a second computational study involving 150 four, five, and six-letter high frequency words, and 150 four, five, and six-letter low frequency words, they also found systematic differences between the frequency distributions for phonemes and letters across different levels of word frequency. For example, the phonemes /n/, /l/, /t/, and /z/ represented a total of 23 percent of the phonemes in high frequency words but only 18 percent of the phonemes in low frequency words. Thus, high and low frequency words not only differ in their experienced frequencies, but they also differ in their structural properties. Landauer and Streeter's findings influenced subsequent models of spoken word recognition because they showed that sub-lexical segments might affect how words are organized in lexical memory.

Following Landauer and Streeter (1973), several researchers began to study the relations between the phonological properties of words and the structure of the lexicon (Eukel, 1980). Eukel had a group of subjects listen to a recorded list of 25 CCVC nonsense words varying in Greenberg and Jenkins' measure of phonological similarity, and 58 real words with a wide range of objective frequencies. Greenberg and Jenkins' metric measures the phonological similarity of nonsense words to real words of English, and constitutes an indirect way of measuring the phonotactic probability of sequences in the lexicon (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1996).

The subjects in Eukel's study were asked to make subjective judgments about the frequency of occurrence of real words as well as nonsense words. His findings showed that participants subjective judgments of word frequency are highly correlated with both objective measures of experienced word frequency, and Greenberg and Jenkins' computational measure, indicating that word frequency effects for non-words are due at least in part to probabilistic phonotactics. The finding of subjective frequency effects for non-words also adds converging evidence to Landauer and Streeter's hypothesis that the principle of "structural equivalence" does not hold, because high and low frequency words differ not only by virtue of how often they occur in language, but also in their the segmental composition.

The idea that words are organized into similarity spaces based on phonotactic probability and word frequency has motivated the assumption that words are recognized relationally in the context of

other similar sounding words rather than in isolation (Luce &Pisoni, 1998; Marslen-Wilson, 1984). The belief that words are recognized relationally has led to several models concerning the way the lexicon might be organized. One important milestone in the field was the development of Cohort theory (Marslen-Wilson, 1984; Marslen-Wilson, 1990). In Cohort theory, word recognition is assumed to take place one phoneme at a time from the beginning of word in real time. During the initial stages of word recognition, a set of potential word candidates, or what Marslen-Wilson called a "word initial cohort", becomes activated based on information in the stimulus. As additional phonemes are perceived during the recognition process, more words are eliminated from the cohort due to deactivation of words that are no longer compatible with the input signal. This process continues until there is only one possibility. For example, when perceiving an utterance of the word "catapult", the cohort is first reduced to words beginning with the phoneme /k/, and then to words composed of word initial /ka/ (for example, "can", "cap", "catapult", etc.). Finally, once enough phonemes are perceived to the point where the input diverges from all other possible word candidates, "catapult" is recognized.

A connectionist model of word recognition sharing some design features of Cohort theory is the TRACE model of speech perception (McClelland & Elman, 1986; also see Protopapas, 1999). Two versions of TRACE have been developed to model different data in the literature: TRACE I was the initial implementation of the model and was built to model phoneme perception, while TRACE II was designed to explain data concerning lexical access. The TRACE I model consists of three layers of nodes: the feature, phoneme, and word layer. The representations in TRACE are localist where each node, or independent processing unit, represents a particular unit at each layer. Thus, at the phoneme layer, each node represents a phoneme, and at the word layer, each node represents a word in the lexicon. When input activates a set of features, activation spreads bi-directionally between layers, activating items consistent with the input on relevant dimensions. Activation flows between layers in a process termed "interactive activation"; phonemes activate words, and words containing activated phonemes send the proportional amount of activation back down to the phoneme and feature layers.

As sensory evidence accumulates, nodes begin to inhibit activated items that are mismatched with the input. Unlike activation, inhibition only operates between nodes within a particular layer. In order to build time into the model, the units in TRACE are reproduced and represented multiple times, with one representation at each time slice. As different representations become active at different time slices, the list of possible word candidates changes, which is an important feature shared with Cohort theory—where activation is a key assumption (Marslen-Wilson, 1984).

Using an activation framework that is similar to Cohort theory and the TRACE model, the Neighborhood Activation Model (NAM) (Luce & Pisoni, 1998) describes how words are recognized "relationally." NAM assumes that similar sounding words compete with or inhibit the input word during the recognition process. The metric for similarity used by NAM differs from Cohort theory. In Cohort theory, perceptually similar words are organized into cohorts based on shared features from the beginning to the end of the word. The metric of phonological similarity used by NAM is the deletion, addition, substitution rule (DAS) originally used by Landauer and Streeter (1973), where two words are "neighbors" if one word can be changed into the other via the deletion, addition, or substitution (DAS) of a single phoneme. Using this metric, "cat" and "cab" are neighbors because they differ only in the coda position and "bat" and "sat" are neighbors because they differ only in the onset position.

NAM assumes that words are organized into similarity spaces in lexical memory, which can be quantified by the number of similar sounding neighbors a word has, referred to as *neighborhood density*. If a word has a large number of phonologically similar neighbors based on the DAS rule, then the word is classified as *high-density* because it resides in a high-density neighborhood. If a word has few phonological neighbors, then it is classified as a *low-density* word. NAM makes several specific

predictions about spoken word recognition including neighborhood density effects: words with many phonological neighbors are inhibited more by their phonological neighbors and consequently are recognized more slowly and less accurately than words with fewer phonological neighbors.

Luce and Pisoni (1998) tested NAM in several experimental paradigms including: lexical decision, word repetition (naming), and perceptual identification. The Hoosier Mental Lexicon, a database of 20,000 words and their phonological transcriptions, was used to compute "similarity neighborhoods" (Nusbaum, Pisoni, & Davis, 1984). Data from lexical decision and word repetition experiments showed that words in *dense neighborhoods* had longer response latencies than words in *sparse neighborhoods*. Data from identification experiments demonstrated that listeners identified low-density words over a range of different signal-to-noise ratio more accurately than high-density words.

Luce and Pisoni's findings inspired a series of follow-up studies that revealed competitive inhibition at the lexical level and facilitatory probabilistic phonotactic effects at the sub-lexical level (Vitevitch & Luce, 1998; Vitevitch & Luce, 1999). For example, using a same-different discrimination experiment, Vitevitch and Luce (1999) showed that response latencies were longer for high-density words consisting of high phonotactic probability segments (i.e., high probability words) than low-density words consisting of low phonotactic probability segments. The opposite result was observed for non-words; non-words with high phonotactic probability segments were recognize more quickly and more accurately than non-words consisting of low phonotactic probability segments.

Graph Theory and Complex Systems

While Cohort theory, TRACE, and NAM have provided several novel insights about how words are recognized in the context of other similar sounding words in memory, the models are incomplete because they fail to provide a description of how the global properties of the mental lexicon might affect spoken word recognition. If Treisman's (1979) hypothesis is correct with regard to conceptualizing lexical representations of words as trajectories in a multi-dimensional acoustical space, then it is important to explore the effects of the global topology and connectivity among words in this space. In order to accomplish this goal, new tools and new variables are needed to describe the structure of the lexicon and quantify structural relationships between words.

Recently, several researchers have applied tools used in the analysis of complex systems to the study of the mental lexicon (Gruenenfelder & Pisoni, 2006; Steyvers & Tenenbaum, 2005; Vitevitch, 2004). Complexity theory conceptualizes how the separate parts of a system interact with one another to produce emergent behaviors (Barabási, 1999). Complex systems can be represented graphically with individual components represented as nodes or vertices, and relationships between nodes represented as links. Graph theory provides a means of describing the patterns of connectivity among words in the mental lexicon, and offers a different theoretical approach for modeling lexical growth and development (see Vitevitch, 2004).

The study of complex systems is interdisciplinary, spanning several fields of scientific inquiry. Barabási described the network structure, including the *scale-free* distribution of links, in various complex systems including social networks, the world-wide-web, and the Internet (Barabási, 1999; Albert & Barabási, 2002). The degree distribution of a network refers to the number of links a node has. If a node is randomly selected, the probability that it has k neighbors is $p(k) = c/k^{\alpha}$, where c and α are constants.[2] The scale-free degree distribution follows a power law; which is a linear trend if plotted on Log-Log coordinates, and approximates an exponential distribution when plotted on linear coordinates. A

[2] The degree of the exponent in complex networks following the power law degree distribution of links is between 2 and 3.

"scale-free degree distribution" also means that no single descriptive parameter such as a mean or median can accurately describe the number of links attached to a randomly selected node. One property of scale-free networks is that a few nodes are highly connected hubs, since they contain many links, while the majority of nodes have few links.

Many real world and man made complex systems share the *small world* property (see Barabási, 1999). A short path length where the distance between any two nodes is small (i.e., six degrees of separation) and higher clustering coefficient (CC) than a random graph of comparable size characterize the small-world structure. CC is the probability that any two neighbors of a given node are connected (Watts & Strogatz, 1998). In small world networks, the average CC is several magnitudes larger than the CC of a random graph.

Small World and Scale Free Structure of the Mental Lexicon

In a recent computational study, Steyvers and Tenenbaum (2005) modeled three types of semantic networks as complex graphs. The semantic networks they examined consisted of *word associations*, *WordNet*, and *Roget's Thesaurus*. The word association database consisted of stimulus words given to participants and words they wrote down that were associates of the stimulus. For example, if a subject was given the word "dog", they might have generated the associative response "fetch". WordNet, the second database, is a modern version of Roget's thesaurus consisting of words along with their synonyms and antonyms. Steyvers and Tenenbaum reported that all three networks displayed a small world structure and exhibited a scale-free degree distribution, suggesting that semantic representations in memory can be modeled as a complex system.

Vitevitch (2004) conducted a phonological analysis of the mental lexicon, modeling it as a complex network using standard graph theoretical measures. Vitevitch used the DAS metric to construct the graph. Two words (nodes) were connected in the graph if they differed by a single phoneme. The results of the analysis of 19,340 words in the Hoosier Mental Lexicon database suggested that the lexicon shared a variety of general properties with natural and artificial complex systems. Vitevitch found a short path length and a CC that was several magnitudes larger than would be predicted from a random graph of similar size, as well as a scale-free degree distribution.[3] Based on the earlier suggestions of Albert and Barabási, Vitevitch argued that these findings were indicative of a mechanism for "preferential growth" and "attachment". These terms refer to the notion that during lexical development, words that are already highly connected are more likely to acquire new connections as novel words are learned and added to the lexicon. The more time a word remains in the lexicon, the more connections it forms to other words already in the lexicon. That is, as the mental lexicon grows, children are more likely to learn words that are similar to ones they already know.

Recently, Gruenenfelder and Pisoni (2006) replicated and extended Vitevitch's (2004) work. They conducted a reanalysis of data collected by Luce and Pisoni (1998) in their word repetition and identification studies. Graph theory measures including CC were computed for each word in the database in order to also replicate Vitevitch's findings. When comparing the reaction times with the CC of the stimuli in an open set word repetition task, Gruenenfelder and Pisoni observed a significant positive correlation between CC and response latency. However, no significant effects were observed in their analysis correlating percent correct scores and CC from the perceptual identification experiment.

[3] The complex graph constructed by Vitevitch (2004) is problematic because out of nearly 20,000 words, over 10,000 words are isolated hermits without any phonological neighbors. The lack of connectivity, particularly among multi-syllabic words, suggests that the metric used for constructing the graph (i.e., the DAS rule) might be flawed in some respects.

The findings of Gruenenfelder and Pisoni (2006) suggest that CC affects spoken word recognition. As with the case of neighborhood density, high CC words appear to inhibit the processing of spoken words producing increased response latencies relative to low CC words, indicating that global variables of the lexicon might have behavioral effects on spoken language processing. Their study, however, has several weaknesses. The study was a post-hoc investigation, and neighborhood density and CC were confounded in the original stimuli used by Luce and Pisoni. There are no behavioral studies where these variables are controlled and analyzed separately from one another.

Given these weaknesses, the present study was carried out to experimentally investigate the effects of CC on spoken word recognition to determine if non-local properties of the lexicon affect the word recognition process. Graph theoretic analyses of the mental lexicon are in their infant stages and behavioral evidence is needed in order to begin to support the hypothesis that the organization of words in memory shares structural properties with other complex systems. In the present study, we replicated the results reported by Gruenenfelder and Pisoni (2006) and extended the findings into new domains by using the same-different discrimination, and perceptual identification paradigms.

Models and Hypotheses

What would previous models of word recognition predict about CC effects? Most models do not address how differences in CC would affect spoken word recognition. The "urn model" and Logogen theory, for instance, would not be able to make predictions about the effects of graph theoretical measures like CC, since the structure of these models does not presuppose lexical connectivity. Likewise, since Cohort theory assumes that words are recognized from beginning to end and words not in the cohort are discarded, it would not make predictions about CC.

A connectionist model like TRACE could, in principle, make predictions about the effect of CC on word recognition. Recall that activation in TRACE spreads between layers, while inhibition functions within layers. Thus, words send inhibition to other similar sounding words as activation reaches the word level from the phoneme level. As one word node begins receiving more activation than other words, it begins to inhibit those words—making them less likely to reach threshold. The connectivity pattern through which inhibition is sent is irrelevant. What matters is which word most closely matches the activation of phonemic units. Once inhibition from a more activated word is sent to its neighbors, the connectivity, or similarity between the neighbors (phonemically similar words) should not affect recognition of the stimulus. Therefore, TRACE would predict null results with respect to CC because there are no connections beyond local lexical interactions.

NAM, like the TRACE model of speech perception, also does not make any predictions about the effects of CC on spoken word recognition. While NAM describes how words are related to one another through the DAS metric, and how words in this sub-volume of the lexicon affect recognition, it does not describe how neighborhood connectivity affects response times or identification of words in degraded listening conditions. Thus, NAM assumes that all neighborhoods of a given size and neighborhood frequency have similar effects on spoken word recognition.

The purpose of this study was to determine if the CC effects on spoken word recognition and performance observed by Gruenefelder and Pisoni (2006), i.e., where shorter response latencies were observed for low CC stimuli in word repetition tasks, can be replicated when controlling for word frequency. Based on their earlier findings, we expected to find shorter response latencies for low CC words in a same-different discrimination task and more accurate identification of low CC words relative to high CC words under degraded listening conditions. Observing effects of CC on spoken word recognition in both same-different discrimination and perceptual identification tasks would provide

converging support for the hypothesis that global properties of the mental lexicon affect spoken word recognition processes.

Experiment 1: Effects of CC on Same-Different Discrimination

The purpose of Experiment 1 was to test the hypothesis that differences in CC affect the processing time of spoken words in a same-different discrimination task. An analysis of Vitevitch and Luce's (1999) high probability/high density stimuli revealed that the words they used also had high CC. They found that the reaction times for high probability/high-density words were lower than the reaction times for low probability/low-density words. Since probability and CC were confounded in the stimuli used in their study, the pattern of results means that low CC words were responded to more quickly than high CC words. In order to replicate both Vitevitch and Luce (1999) and Gruenenfelder and Pisoni (2006), we created a new set of stimuli that controlled for phonotactic probability and word frequency while manipulating CC and neighborhood density separately.

In the same-different task used in Experiment 1, listeners were presented with two spoken words over headphones and were asked to determine whether the two stimuli were the same or different words. The independent variables of CC and neighborhood density were moderately correlated (r ~ .49) in the Hoosier Mental Lexicon database, but the stimuli used in these experiments controlled for CC and neighborhood density. In addition to the prediction that high CC stimuli would be processed more slowly than low CC stimuli, we also predicted that high-density words would be processed more slowly than low-density words.

Method

Design

A 2 x 2 within subject design with CC and neighborhood density as independent variables was used. The dependent variable was reaction time measured in milliseconds. One hundred and sixty words were evenly divided into four experimental conditions: low-density and low CC; low-density and high CC; high-density and low CC; and high-density and high CC. Table 1 shows an outline of the basic design used here and in the following experiments.

High CC High Density (n = 40)	Low CC High Density (n = 40)
High CC Low Density (n = 40)	Low CC Low Density (n = 40)

Table 1. Experimental design: 2x2 within subject factors. Neighborhood Density and Clustering Coefficient are the independent variables and 40 words are used in each cell.

Participants

All of the participants were native speakers of Midwestern American English, who reported no history of speech and hearing disorders at the time of testing. The nineteen participants were recruited from the undergraduate introductory psychology subject pool at Indiana University in Bloomington. None of the subjects participated in more than one experiment reported here.

Stimulus Materials

All words in this experiment were selected from a subset of 938 monosyllabic words obtained from the Hoosier Mental Lexicon database. The stimuli in this experiment consisted of one hundred sixty words used as same pairs, and three hundred twenty words used in one hundred sixty different pairs. The same pairs consist of two different recordings of the same word. We did this in order to encourage listeners to process the stimuli lexically rather than attend to fine acoustic details.

Three hundred twenty words were selected for the different pairs. The words within each pair were phonological neighbors and differed by only one phoneme. The one hundred sixty pairs were counterbalanced for the location of the phonemic differences. A male talker recorded all of the stimulus tokens. The list of stimulus words was presented to the talker and the words were recorded one at a time on a PC using the SAP program. The words were subsequently digitized and edited into individual files using the PRAAT waveform editor. The Level-12 program was used to level the sound level of all the words at 65 dB.

Clustering Coefficient

The program Pajek was used to compute the CC for each of the stimulus words. Since CC is a measure of probability, it ranges from 0 to 1. We selected words in the upper 40 percent of the probability distribution as high CC words, and the words in the lower 40 percent of the distribution as low CC words. The average CC for high CC stimuli was .235, and the average CC for low stimuli was .126.

Neighborhood Density

We took similar steps in the selection process of high and low-density words. Neighborhood density was defined as the number of neighbors based on the DAS rule computed in the Hoosier Mental Lexicon database consisting of n = 19,340 words. High and low density items were selected from the upper and lower 40 percent of the distribution respectively using the same subset of 938 monosyllabic. The average neighborhood density for high-density words was 19.9 neighbors, and 11.1 neighbors for low-density words. Density was not weighted by frequency as we only computed raw scores.

Phonotactic Probability

Phonotactic probability was controlled for across levels of CC. The average phonotactic probability was .15 for the high CC words, and .14 for the low CC words. These probabilities refer to the frequency that a certain phoneme or segment occurs in a word. Thus, the segments in the high and low CC words occur at approximately the same rate in the English language. Since phonotactic probability was correlated with lexical density (Vitevitch & Luce, 1999), phonotactic probability varied across the high and low-density stimuli. High-density words had an average phonotactic probability of .168 and low-density words had an average phonotactic probability of .124.

Word Frequency

Frequency of occurrence in the language obtained from Kucera and Francis (1967) was matched for each of the four conditions. Word Frequency scores were transformed using a logarithmic function in order to compress frequencies at the high end of the distribution. The mean log frequency was 2.37 for the low frequency words, and was 2.32 for the high frequency words ($F_{(1,79)} = .095$, $p < 1.0$). For density, the mean log frequencies were 2.35 and 2.34 for low and high densities respectively ($F_{(1,79)} = 2.0$, $p = .665$).

Procedure

Testing took place in individual booths with up to three listeners being tested at a time. Each listener was seated in front of a PC equipped with *Beyer Dynamic DT 100* headphones and a two-button response box. Presentation of stimuli and collection of listener's responses were controlled by the PC. At the beginning of each trial, a light was illuminated at the top of the button box before each pair of words was played. The stimuli were presented over the headphones fixed at a comfortable listening level. There was a delay of 500 ms between the offset of the first word and the onset of the second word in the pair. The next trial began after the listener made a response on the button box. The instructions given at the beginning of the experiment asked listeners to make responses as quickly and as accurately as possible as soon as they were sure of their decision. Listeners responded *same* with their right hand and *different* with their left hand. Same and different trials were randomized for each subject. All listeners received ten practice trials at the beginning of the experiment that were not included in the final data analysis. No feedback was given on any of the trials.

Results

Analysis of Same Pairs

The mean response times for the *same* responses are shown in Figure 1 separately for CC and neighborhood density. Figure 2 shows the results for CC collapsed across neighborhood density. In each figure, mean response time in milliseconds is plotted on the ordinate. A one way ANOVA shows that the observed mean accuracy for the task approached ceiling levels and was not significant across conditions (($F(1,18) = 1.47$ and ($F(1,18) = 3.99$) for neighborhood density and CC respectively). Error responses were omitted from the data analysis of the response latencies.

Figures 1 and 2 show that the mean response times on the same trials were faster for low CC words across the two levels of lexical density. A series of ANOVAs were carried out on the response latencies for the *same* trials. F values were computed across both subjects and items; that is, both subjects and stimulus words were treated as random variables (Clark, 1973). We chose a .05 (two-tailed) level of significance for each test.

The results from the ANOVA across subjects demonstrated a significant main effect for CC. High CC words (mean = 987.14 ms) were responded to more slowly than low CC words (mean = 950.71 ms) ($F(1,18) = 10.007$, $p < .01$). The ANOVA for CC over items also showed a significant main effect for CC ($t(39) = 5.45$, $p = .022$). A binomial test was also significant; 16 out of 19 listeners showed longer response latencies for the high CC word pairs ($p = .004$). That is, the effects of CC on *same* pairs were robust across listeners.

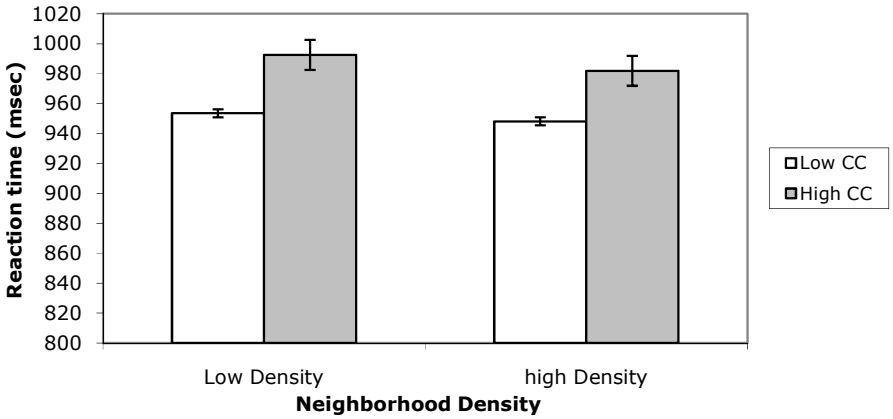


Figure 1. Mean reaction times across density and CC levels in the same-difference matching experiment. The data do not show a significant interaction, or significance across density levels. A significant main effect was observed for CC, where high CC words have longer response latencies than low CC.

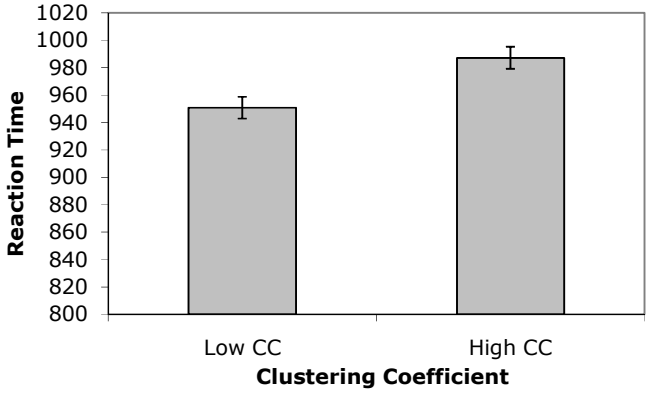


Figure 2. Mean reaction times for CC collapsed across density levels. The results show that high CC words were responded to more slowly than low CC words.

The effect of neighborhood density was not significant (mean HD = 964.9; mean LD = 972.62) ($F_{(1,18)} = .420$, $p = .525$). The only instance where neighborhood density reached significance was when the high CC low-density and low CC and high-density conditions were compared (mean HCC LD = 992.45 ms; mean LCC HD = 948.03 ms) ($t_{(39)} = 3.091$, $p < .01$). The interaction between density and CC was not significant ($F_{(1,18)} = .072$, $p < .720$).

Discussion

The results from the same-different discrimination task confirm the hypothesis that words with high CC produce longer response latencies than words with low CC. The hypothesis about the effects of neighborhood density was that high-density stimuli would show longer response latencies than low-density stimuli. Vitevitch and Luce found longer response latencies for high-density words in a same-

different discrimination task (1999), although they manipulated phonotactic probability directly rather than neighborhood density. One explanation for our failure to find significant effects for neighborhood density is that CC might be responsible for such effects, but in previous studies, neighborhood density was confounded with CC.

In order to understand the effects of CC more completely, it is necessary to obtain converging evidence. The stimuli used in Experiment 1 were spoken in the clear and presented under optimal listening conditions. In Experiment 2, we explored how listeners would respond to words with different levels of CC under degraded listening conditions using a same-different discrimination task. An analysis of the error patterns across experimental conditions was also carried out in order to further understand how CC affects spoken word recognition.

Experiment 2: Same-Different Discrimination. What Errors do Listeners Make?

In the first experiment, we analyzed reaction time data and found a significant main effect for CC. Experiment 2 was designed to investigate the effects of signal degradation on spoken word recognition. The stimuli used in Experiment 2 were degraded by processing them with a noise vocoder, which is used to simulate speech sounds that cochlear implant users are exposed to. The Tiger Speech Cochlear Implant Simulator version 1.01.07 was used to degrade the signal. Filtering the signal into a specific number of frequency bands in the first step in creating vocoded speech—in this experiment we used 8 bands. Once the speech was filtered into a specified number of bands, the amplitude envelope for each band was extracted with a low pass filter. Frequency was then replaced in each band with noise. Shannon, Fan-Gang, Kamath, Wygonski, and Ekelid. (1995) found that 4 channels of vocoded speech, using either noise or sine wave carriers, provided sufficient information for word identification in meaningful sentences. Shannon et al.'s findings suggest that this might be the minimal spectral information required for recognition, provided that temporal cues are also present in the signal. In order for subjects to have a better chance of perceiving words in isolation, we used 8 spectral channels instead of 4-channels.

Listeners in Experiment 2 carried out the same discrimination task used in Experiment 1; only now the stimuli were degraded using noise vocoded speech. Recall that error data were analyzed in Experiment 1 for the purpose of ruling out the possibility that significant differences in accuracy were found across conditions, and we observed that listeners made few errors with no difference in accuracy across conditions. Our prediction for Experiment 2 was that low CC pairs of words would be identified by listeners as *same* more accurately than high CC words.

Method

Design

This experiment also used a 2 x 2 design with CC and neighborhood density as independent variables. The dependent variable was percent correct since the degraded signal should decrease response accuracy. The one hundred sixty stimuli were evenly divided into the same four conditions used in Experiment 1.

Participants

The participants were twenty native speakers of Midwestern American English who reported no history of a speech or hearing disorder at the time of testing. The participants were recruited from the undergraduate psychology paid subject pool from Indiana University in Bloomington. Subjects were paid seven dollars and none of them served in the previous experiment.

Stimulus Materials

The same stimuli used in Experiment 1 were used in Experiment 2. In this experiment, both the same and different words were processed using the Tiger Speech Cochlear Implant Simulation version 1.01.07.

Procedure

Listeners were informed that they would hear pairs of English words. This point was emphasized in order to encourage listeners to engage in lexical access and prevent them from interpreting the vocoded speech as noise. Listeners were instructed to respond *same* or *different* as accurately as possible while also moving quickly in the experiment. The procedure was otherwise identical to the task described in Experiment 1.

Results

A series of repeated measures ANOVAs calculated the main effects of CC, neighborhood density, and the interaction between the two variables. A significant main effect was observed for CC (mean percent correct for low CC = 83.7 and the mean percent correct for high CC = 87.6) ($F_{(1,19)} = 19.696$ and $p < .001$). 17 of the 20 listeners recognized high CC words more accurately than low CC words ($p = .004$). The results are shown in Figure 3. The ANOVA of the items analysis showed a trend where high CC words were recognized more accurately than low CC words ($F_{(1,39)} = 3.89$, $p = .056$).

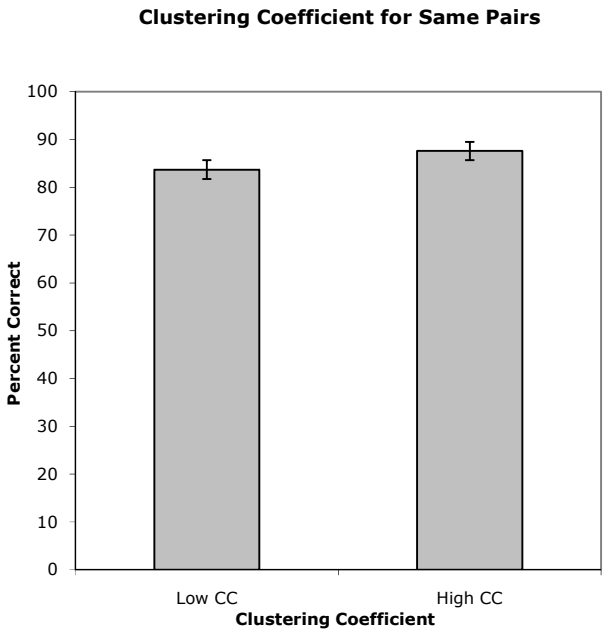


Figure 3. An analysis of the *same* pairs from Experiment 2. We collapsed across the variable neighborhood density and analyzed percent correct as a function of CC. The figure shows that high pairs of CC words were recognized more accurately than low CC pairs of words.

The main effect of neighborhood density showed a marginal but non-significant trend toward low-density words being recognized more accurately than high-density words. The marginal effects of density replicated findings reported in previous studies (Luce & Pisoni, 1998; Vitevitch & Luce, 1998). The mean percent correct for low-density words was 86.625, and the mean percent correct for high-density words was 84.675 ($F(1,19) = 2.144$, $p = .159$). The marginal main effect for neighborhood density points in the opposite direction as the effect of CC.

We also analyzed the responses to the different pairs. The purpose of this analysis was to determine if listeners were biased to respond *same* or *different*. In other words, how well were listeners able to discriminate the same pairs from the different pairs? To answer this question, the overall d' (sensitivity) for the *same* trials was calculated giving a value of 1.72: where d' = Z False Alarm – Z Hit. The value of 1.72 corresponds to an ROC curve indicating that listeners were able to discriminate the same pairs from the different pairs above chance. We also computed the mean d' for each independent variable across subjects and analyzed this data using one-way ANOVAs. Data showed a significant main effect for CC with a mean d' of 1.711 for low CC words and 1.96 for high CC words ($F(1,19) = 19.381$ and $p < .001$), suggesting that high CC words were more discriminable than low CC words. Effects were non-significant for density ($F(1,19) = 1.094$), and non-significant for the CC x density interaction ($F(1,19) = .634$).

Discussion

The results of manipulating CC in Experiment 2 revealed a different pattern of results from Experiment 1. In Experiment 1, we observed faster reaction times for low CC words relative to high CC words in the same-different task. When we used the 8-channel noise vocoder to degrade the stimuli in Experiment 2, we observed that low CC words were responded to less accurately across same pairs than high CC words, which was the opposite of what we predicted.

One explanation for the surprising and anomalous results from Experiment 2 was that listeners might have used pattern-matching strategies to complete the task without accessing words in their lexicon. That is, listeners carried out the same-different task without recognizing words even though they were told in the instructions that the stimuli consisted of English words. Although the d' analysis showed that listeners could discriminate the same pairs from the different pairs, it is doubtful that they were discriminating them on a lexical basis. In order to test this hypothesis, it is necessary to measure the accuracy of word recognition in a perceptual identification experiment carried out under the same degraded listening conditions. Recall that in Experiment 1, listeners heard words spoken in the clear and were instructed to make same-different judgments based on different tokens of the same word. This provides evidence that listeners were accessing their lexicon in Experiment 1.

In order to analyze listener's error responses while encouraging them to engage in lexical access, we used a perceptual identification experiment with degradation level as a between subjects variable. Unlike the same-different task used in Experiments 1 and 2, perceptual identification provided a more detailed description of how listeners perceived spoken words since they had to access their lexicon and respond with the word they heard. And, since an important aspect of this study is the investigation of the structural relationship of spoken words in the phonological mental lexicon, the perceptual identification paradigm provides a useful means for examining these relationships. Another reason for using perceptual identification in Experiment 3 was that the procedure allowed for the systematic study of listener's error patterns (Savin, 1963), as well as a detailed analysis of the component segments of the correct and incorrect responses.

Experiment 3: CC and Perceptual Identification of Spoken Words

Experiment 2 demonstrated that discriminated high CC words more accurately than low CC words under degraded listening conditions. Because the stimuli were degraded, listeners might not have perceived words and accessed representations from their lexicon, but instead could have relied on an auditory pattern matching strategy. No significant main effect was observed for neighborhood density, although the data indicated a trend toward a higher percentage of correct responses in the low-density condition compared to the high-density condition.

While the data from the same-different discrimination task used in Experiment 1 and 2 provided some indication of error patterns across conditions, they tell us little about how the stimulus words and their sublexical components were perceived by listeners. Which segments do listeners perceive most accurately when exposed to different levels of degraded speech? Also, how does the perception of words and component segments vary as a function of CC and density?

Both correct and incorrect responses were analyzed in Experiment 3 since both types of responses contain phonological and frequency related information. We also measured the word frequency of listener's responses as a function of CC and neighborhood density. Pollack, Rubenstein, and Decker (1960) conducted an analysis of incorrect responses to spoken words presented in noise. Hypothetically, because listeners were more accurate in judging high CC words as *same* in Experiment 2, we expected a similar pattern of results from the identification task in the between subject condition in which listeners were presented with 8-channel vocoded speech. That is, listeners should be more accurate in identifying high CC words relative to low CC words. Another prediction was that since listeners might not have recognized words in Experiment 2, we expected that percent correct identification in the 8-channel condition would be near the floor. Because listeners responded to low CC words more quickly than high CC words in Experiment 1, we expected low CC words to be recognized more accurately as the number of channels increased and the listening conditions improved.

Method

Design

A 3 x 2 x 2 design was used in Experiment 3. The between subject variable was the number of channels (8,10,12) and the within subject variables were CC and neighborhood density. The dependent variables were percent correct response, word frequency of listener's response, and response entropy. The stimuli were evenly divided into four conditions.

Participants

The participants in Experiment 3 were sixty-three native speakers of Midwestern American English, who reported no prior history of speech or hearing disorders at the time of testing. Twenty-one participants were recruited for each of the three between subject conditions from the undergraduate psychology pool at Indiana University in Bloomington. Listeners were either assigned course credit or paid seven dollars for their participation.

Stimulus Materials

The same set of one hundred sixty words used in the *same* pairs in Experiment 1 and 2 were used in Experiment 3. The stimuli were degraded using the Tiger Speech Cochlear Implant simulation version 1.01.07 described under the stimulus materials section under the Experiment 2 heading. The stimuli were degraded using 8-channels, 10-channels, and 12-channels.

Procedure

Experiment 3 used an open-set word identification task. Words were played over *Beyer Dynamic DT 100* headphones connected to a Macintosh computer at a comfortable listening volume. Listeners were instructed to listen to the words and use the keyboard to type in what they thought they heard as accurately as possible. Subjects were also instructed to listen carefully and take their time during the procedure.

Each trial began with the presentation of a plus sign on the center of the screen displayed for 500 milliseconds. After the plus sign disappeared from the screen, a degraded word was played over the headphones at a comfortable volume. After the word finished playing, a dialogue box was displayed on the screen asking the subject to type in what they heard. There was a 1,500 ms pause before the next trial began. The next trial did not begin until the subject finished typing in the response.

Results

In the data analysis, both the target word and response were phonetically transcribed using the alphabet form the CMU dictionary. If the transcription of the target and response matched in the onset, nucleus, and coda positions, the response was scored as correct. If listeners typed in a homonym of a target word, for example, by typing in the word *sea* instead of *see*, the response was scored as correct since the phonetic transcriptions are identical. In the analysis of correct responses, we first analyzed the percentage of words correctly identified as a function of number of channels, CC, and neighborhood density. We also looked at the number of correct responses in the onset, nucleus, and coda position across the two levels of CC and lexical density.

Words

In the first analysis, we computed the percentage of words identified correctly in each condition. The percentage of words identified correctly as a function of channel is shown in Figure 4(a). The overall percentage of words correctly identified as a function of number of channels and CC is plotted in Figure 4(b).

As the level of degradation decreased by increasing the number of channels from 8 to 10 to12, listeners identified words more accurately. The largest increase in performance occurred when the number of channels was increased from 8 to 10. The ANOVA results show a significant effect of the number of channels ($F(2,60) = 4213.732$, $p < .001$), clustering coefficient ($F(1,60) = 1399.787$, $p < .001$), and neighborhood density ($F(1,60) = 4213$, $p < .001$). There were also significant interactions between CC and the number of channels ($F(1,20) = 1402.87$, $p < .001$), neighborhood density and CC ($F(1,20) = 769.893$, $p < .001$), density and the number of channels ($F(1,20) = 1345$, $p < .001$), and a three way interaction between neighborhood density, CC, and the number of channels ($F(1,20) = 763$, $p < .001$). The reason for these interactions is that high CC and high-density words were recognized more accurately than low CC and low-density words in the 8-channel condition, while the opposite pattern was the case in the 10 and 12-channel condition.

Significant differences were observed for CC in the 8-channel condition, as indicated in Figure 4(b). The mean percent correct words identified for high CC stimuli was 9.3 percent, and for low CC stimuli, 7.5 percent ($F(1,20) = 5.1$, $p < .05$). The effects of neighborhood density showed a similar difference where the mean number of correctly identified high-density words was 9.34 percent, and 7.5 percent for low-density words ($F(1,20) = 5.5$, $p < .05$). The CC x density interaction was also significant ($F(1,20) = 11.182$, $p < .01$). The data show that the mean percent correct identification for low-density and low CC words was 8.2 percent, low-density and high CC was 6.8 percent. The reverse pattern was true for

high-density stimuli; listeners recognized high CC words more accurately relative to low CC words (11.85 and 6.9 percent respectively). We observed the highest number of words scored correctly when the level of neighborhood density and CC were either both high or both low. A subsequent items analysis revealed no significant differences for CC, density or the interaction—forcing us to question the main effects for the 8-channel condition ($F(1,20) = .015$, $p < 1.0$).

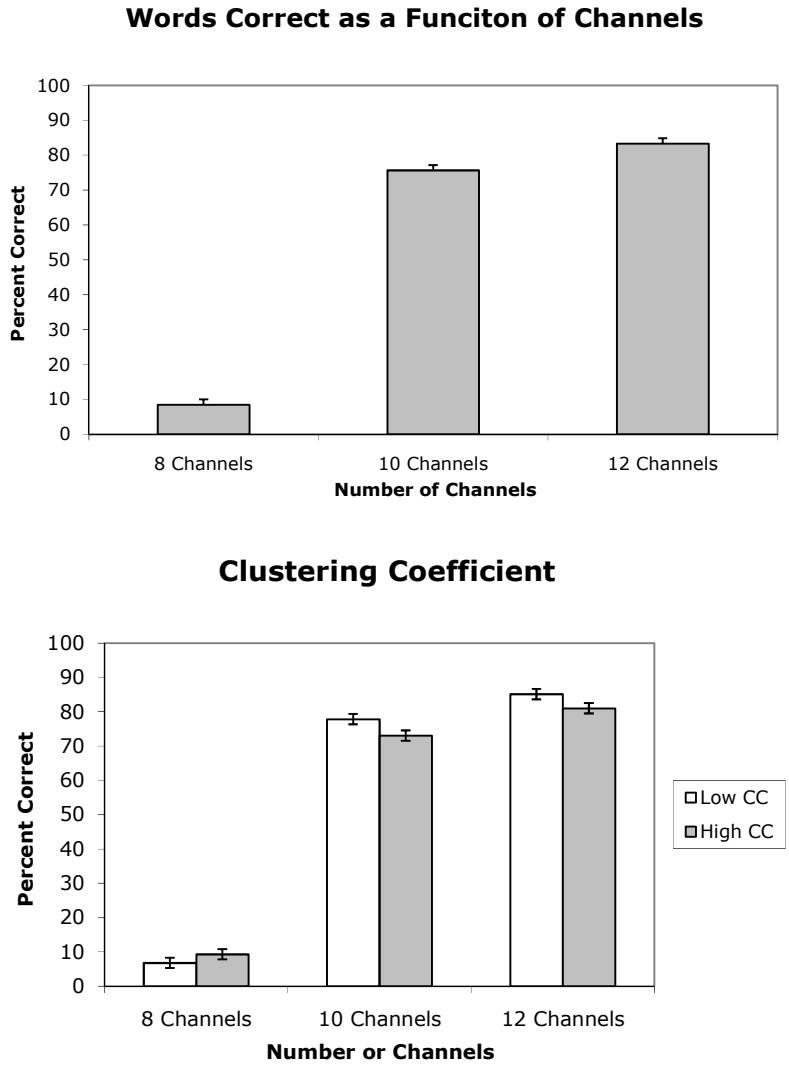


Figure 4. (a) The upper panel of the plot shows percent correct word identification as a function of number of channels. As the signal to noise ratio indicated by the number of channels increases, percent correct increases. Notice that the function is not linear but instead approximates a sigmoidal function. (b) The bottom panel shows percent correct word identification both as a function of CC and number of channels. At the low end of the performance function, listeners identify high CC words more accurately than low CC words. The opposite is true for the 10 and 12-channel conditions.

In the 10-channel condition, we observed an improvement in the percentage of words correctly identified. As predicted, listeners identified low CC words more accurately than high CC words. The mean percent correct for low CC words was 78.30, and for high CC words it was 74.2 percent ($F(1,20)$ = 21.94, p <.001). These results confirmed our results from Experiment 1. No significant difference was observed for density ($F(1,20)$= .55, p <1.0), and no interaction was observed ($F(1,20)$= .334, p <1.0). An items analysis also showed significant results for CC as predicted ($F(1,20)$ = 14.8, p <.001).

For the 12-channel condition, we observed an additional improvement in the percentage of words identified by listeners. As predicted, listeners identified low CC words more accurately, confirming both the results from Experiment 1 measuring reaction time, and the results from Experiment 3 in the 10-channel condition. The mean percent correct for low CC words was 85.12, and for high CC words, 81 percent ($F(1,20)$ = 18.3, p < .001). Again, no significant difference was observed for density ($F(1,20)$= .45, p <1.0), and no interaction was observed ($F(1,20)$ = 1.2, p <1.0). An items analysis showed a marginally significant difference for low and high CC ($F(1,20)$= 2.98, p <.10).

Analysis of Sub-Lexical segments

In another series of tests, we analyzed listener's responses in terms of percent correct of the onset, nucleus, and coda positions. In the first analysis, we examined the percent correct of the onset position as a function of CC and density. These results are shown below in Figure 5 for 8, 10, and 12 channels.

In the 8-channel condition shown in panel (a), the mean percent correct observed was 28.8 percent for high CC words, and was 37 percent for low CC words ($F(1,20)$ = 39.87, p < .001). The effects of lexical density were marginal but not significant. The mean percent correct observed was 31.5 percent for high-density words, and 34 was percent for low-density words ($F(1,20)$ = 3.294, p < .10). The CC x density interaction was not significant ($F(1,20)$ = .007, p < 1.0).

In the 10-channel condition shown in panel (b), we observed significant results for CC and density. We observed a mean percent correct of 90.83 for the onset in low CC words, and 88.87 in high CC words ($F(1,20)$ = 5.1, p <.05) (shown in Figure 5). No significant interaction was observed. The mean percent correct for the onset in low-density words was 88.3 and high density was 91.4 ($F(1,20)$ = 12.69, p <.01).

A similar trend was observed in the 12-channel data for CC shown in panel (c), although we did not observe significant effects for density, or the CC x density interaction. Listeners responded to the onset cluster correctly 94.12 percent of the time in low CC words, and 91.13 percent of the time for high CC words ($F(1,20)$ = 19.055, p <.001). The ANOVA for density and the interaction showed ($F(1,20)$ = 2.086, p = .164; $F(1,20)$ = .091, p <1.0), respectively.

A second set of analyses focused on the nucleus position of the word. The results for the 8-channel condition showed that listeners respond more accurately to the nucleus in high CC words relative to low CC words. The mean percent correct for nuclei in high CC words was 28.3, and for low CC words it was 21.3 percent correct ($F(1,20)$ = 25.308, p < .001). Analysis of nuclei also revealed significant results for density (mean HD = 27.9 and mean LD = 21.7 and $F(1,20)$ = 14.873, p <.001). The CC x density interaction was non-significant ($F(1,20)$ = .563, p = .462).

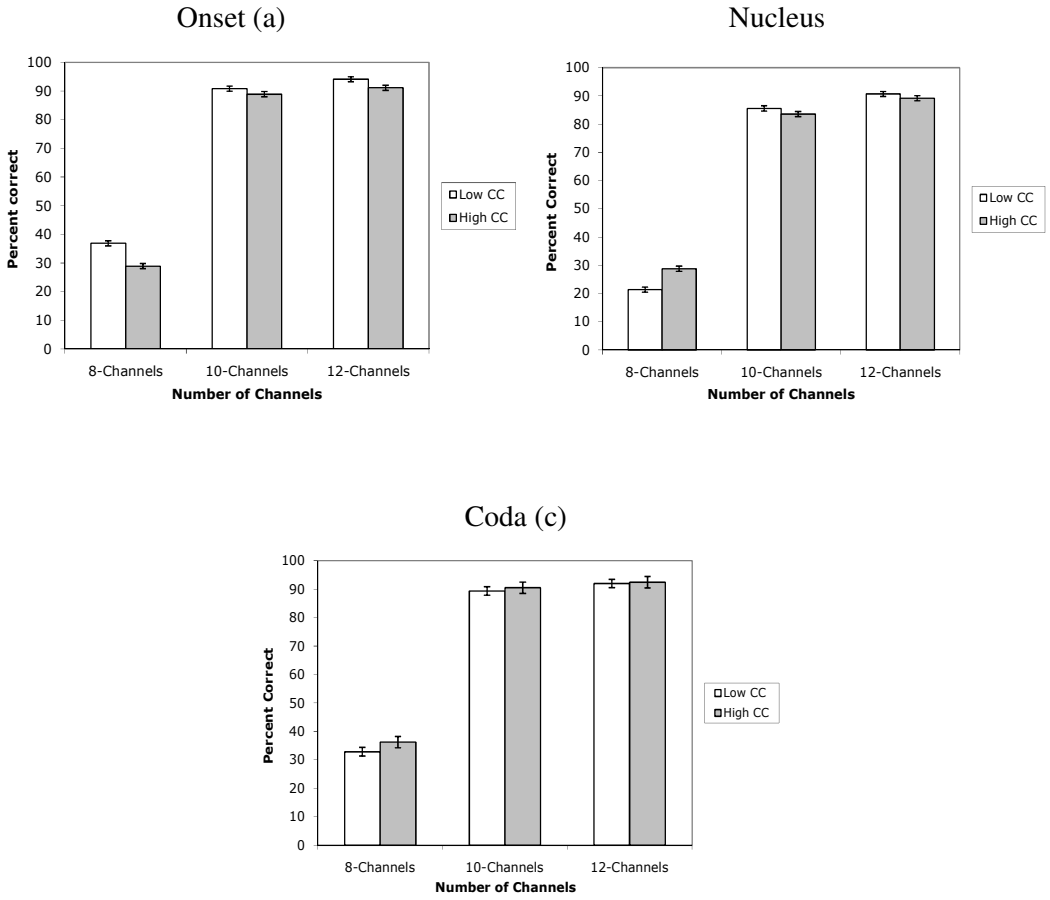


Figure 5. (a) Percent correct identification of the onset position as a function of CC and number of channels. (b) Percent correct identification of the nucleus position as a function of CC and number of channels. (c) Percent correct identification of the coda position as a function of CC and number of channels. Neighborhood density level, rather than CC has a more salient effect on identification in the coda position.

A significant result for CC was observed in the 10-channel condition. Listeners accurately identified the vowel in the nucleus position in low CC words 85.6 percent of the time and 83.6 percent of the time in high CC words ($F(1,20) = 8.157$, $p < .01$). No significant results were observed for neighborhood density ($F(1,20) = .09$, $p < 1.0$). Non-significant results were also observed for the CC x density interaction ($F(1,20) = 2.079$, $p = 165$).

Significant results were not observed for any of the independent variables in the 12-channel condition (mean HCC = 90.7 and mean LCC = 89.23). The mean percent correct for low-density words was 90, and for high-density words it was 89.94. The F value for the CC x density interaction was .005, and $p < 1.0$. Figure 5 (b) shows the percent correct for the nucleus position plotted as a function of CC and number of channels.

In a final analysis of the sub-lexical components of listener's responses, we examined the mean percent correct responses in the coda position. In the 8-channel condition, the data showed non-significant results for CC ($F(1,20) = 1.6$, $p = .22$). We observed significant effects for density with a mean percent

correct of 41.25 for high-density words, and 29.4 for low-density words ($F(1,20) = 40.675$, $p < .001$). Even though CC was not significant, the CC x density interaction was ($F(1,20) = 20.965$, $p < .001$).

In the 10-channel condition, significant results were observed for density with a mean of 88.27 percent correct for low-density words, and 91.55 percent correct for high-density words ($F(1,20) = 12.583$ and $p < .01$). Non-significant results were observed for CC ($F(1,20) = 2.326$, $p < .143$). We also observed a significant CC x density interaction ($F(1,20) = 8.325$, $p < .01$). Although significant effects were observed for neighborhood density in the coda position, they went in the opposite direction of what we predicted based on previous studies (Luce & Pisoni, 1998).

The twenty-one listeners in the 12-channel condition provided similar data as the 10-channel condition. The average percent correct for low-density words was 90.2, and 94.27 for high-density words ($F(1,20) = 4.537$, $p < .05$). No significant results were observed for CC, or the interaction. The percent correct identification across listeners for the coda position as a function of CC and number of channels is shown in panel (c) of Figure 5. Table 2 summarizes the proportion of correct response for high and low CC words in the onset, nucleus and coda positions as the number of channels increases.

Channel	Position	High CC	Low CC
8 C	Onset	.29	.37
	Nucleus	.28	.21
	Coda	.41	.29
10 C	Onset	.89	.91
	Nucleus	.84	.86
	Coda	.92	.88
12 C	Onset	.94	.91
	Nucleus	.90	.90
	Coda	.94	.90

Table 2. Proportion of correct responses across listeners for high and low CC words for number of channels in the onset, nucleus and coda positions.

Analysis of Incorrect Responses

Word Frequency of Incorrect Responses

Does the word frequency of incorrect responses differ across levels of CC and number of channels? Pollack, Rubenstein, and Decker (1960) reported that as the signal-to-noise ratio improved, the word frequency of listener's responses decreased independently of the stimulus word frequency. In a similar study of frequency bias in responses using a visual word recognition task, Goldiamond and Hawkins (1958) found that in the absence of any stimuli, subjects were more likely to respond with items they had been exposed to more frequently during training.

Not long after Pollack et al. (1960) published their results, Gerstman and Bricker (1960) discovered a serious methodological error in their study. Pollack et al. presented listeners with a list of 144 words three times at signal-to-noise ratios of 0, 5, 10, 15, 20, and 25, in that particular order. The observed decrease in the word frequency of incorrect responses was confounded with learning since the words were repeated multiple times across signal-to-noise ratios. In our study, the number of channels

was a between subject condition, where listeners were exposed to each word only once. We investigated whether we could replicate Pollack et al.'s results without the confounding of learning. Additionally, we analyzed incorrect responses to determine whether CC or density affect the word frequency of listener's responses.

To determine the word frequency of incorrect responses, we calculated the modal incorrect response to each stimulus in each of the three between subject conditions. The modal incorrect responses were then placed into the Washington university online speech and orthographic database (http://128.252.27.56/Neighborhood/Home.asp) where an items analysis computed the word frequency score for each word. A necessary condition for a word to be declared a modal incorrect response was that it be given as an incorrect response at least twice.

Figure 6 plots the log frequency of incorrect responses as a function of channels. As predicted by Pollack et al.'s data, the trend indicates that as the number of channels increases, the average word frequency tends to decrease. A pair-wise t-test between the 8 and 10 channel conditions showed a marginally significant trend, where the mean log frequency of incorrect responses for 8-channels was 2.732, and for 10-channels it was 2.45 ($t(17) = 2.01$, $p = .061$). While the average log frequency for the 12-channel condition was lower than the frequency for 10-channels (mean = 2.39), no significant difference was observed between those two conditions ($t(17) = .777$, $p < 1.0$). A t-test revealed that the only possible difference was between the 8 and 12-channel conditions (without correcting for multiple comparisons) ($t(17) = 5.0$, $p < .05$). In short, we observed effects similar to Pollack et al (1960) without confounding the level of stimulus degradation with word repetition.

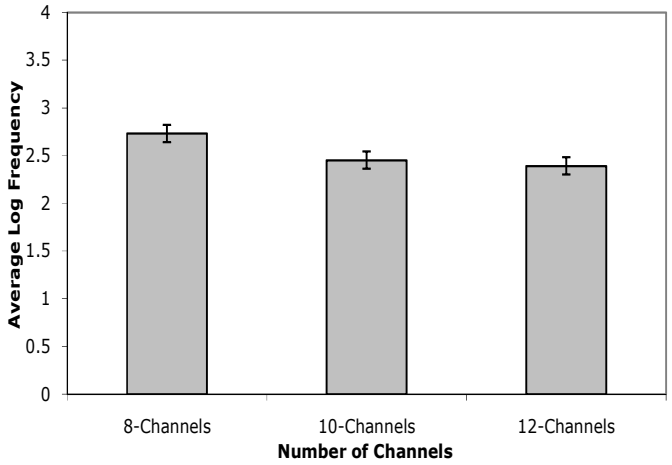


Figure 6. The plot shows the average log frequency of listener's incorrect responses as a function of the number of channels. The overall mean for each condition is collapsed across CC and neighborhood density.

Next, we analyzed the frequency of incorrect responses as a function of CC and density. In the 8-channel condition, the data showed higher word frequency of incorrect responses for high CC words relative to low CC words (mean HCC = 2.88 and mean LCC = 2.59; $F(1, 39) = 4.775$, $p < .05$). The effects of density and the CC x density interaction were non-significant ($F(1, 39) = .397$, $p < 1.0$; $F(1, 39) = .686$, $p < 1.0$). In Experiment 1 we observed faster reaction times, and in Experiment 3 we observed more accurate responses for low CC words than high CC words. This suggests that low CC words are more

easily recognized than high CC words. This explanation is consistent with Pollack et al.'s earlier result showing a decrease in word frequency of responses as signal-to-noise ratio increases. Only now, we are showing a similar pattern in the error responses for the independent variable CC rather than signal-to-noise ratio.

The effect of CC on the word frequency of incorrect responses was weaker in the 10 and 12-channel conditions—where no significant effects were observed (for CC and 10-channels, $F(1, 39) = .002$, $p < 1.0$; for density and 10-channels, $F(1, 39) = .680$, $p < 1.0$; for CC and 12-channels, $F(1, 39) = 1.72$; for density and 12-channels, $F(1, 39) = 4.35$, $p = .052$; and the interaction, $F(1, 39) = .50$, $p < 1.0$). Only density in the 12-channel condition was (marginally) significant. Figure 7 is a plot of the average log frequency as a function of number of channels, and levels of CC.

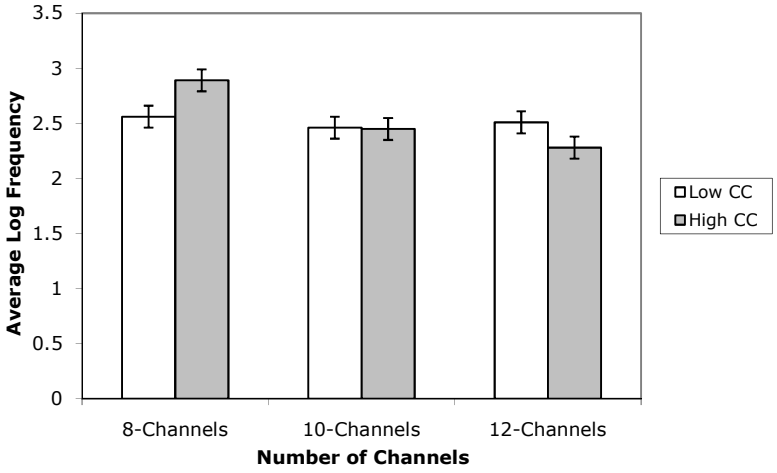


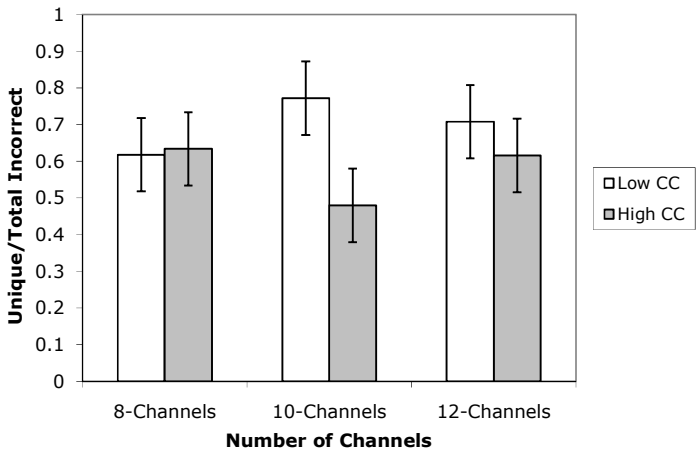
Figure 7. This figure shows the log frequency of listener's incorrect responses as a function of both CC and number of channels. CC is collapsed across neighborhood density.

Diversity of Incorrect Responses

In order to obtain a measure of uncertainty across experimental conditions, we also calculated the number of different incorrect responses given by listeners for each stimuli in the 8, 10, and 12 channel conditions. This analysis was carried out because we hypothesized that different levels of variability or entropy across CC and density might be related to response properties. That is, if listeners give many different responses relative to the total number of incorrect responses, it suggests that they were inconsistent in using incomplete information. A measure of *response entropy* was computed by dividing the number of unique incorrect responses by the total number of incorrect responses given to each stimulus word. We called the measure *response entropy* because it measures variability and uncertainty in the pattern of incorrect responses. The question in the following analyses is whether CC, density, or number of channels, affects the level of entropy in listener's incorrect responses.

Figure 8(a) shows the effect of CC on response entropy across channels, and Figure 8(b) shows analogous results for the effect of lexical density on response entropy.

8 (a)



8 (b)

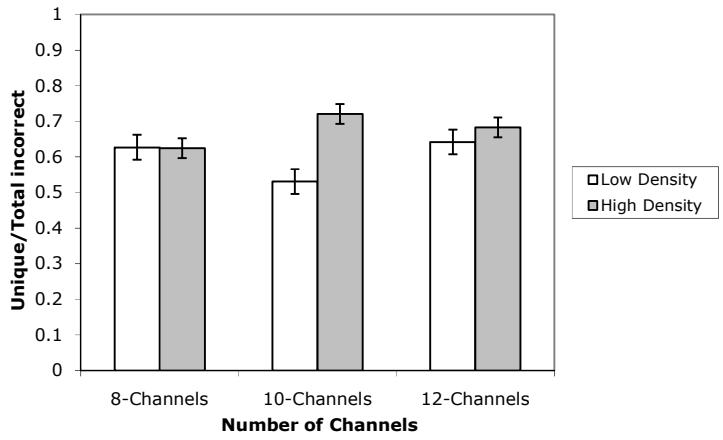


Figure 8. (a) The top panel of the figure shows entropy levels as a function of CC across noise channels. (b) The bottom panel shows entropy as a function of neighborhood density and noise channels.

The results from the ANOVA show that there was not an overall effect of number of channels ($F(1, 39) = 2.501$, $p = .130$), but there was a significant effect for CC ($F(1, 39) = 11.17$, $p < .01$) and neighborhood density ($F(1, 39) = .042$, $p < .05$). We also observed a significant interaction between CC and number of channels ($F(1, 39) = 17.662$, $p < .001$).

The data suggest that levels of response entropy were unaffected by lexical density or CC in the 8-channel condition as shown in Figure 8. The interaction between density and CC with respect to response entropy was marginally significant ($F(1, 39) = 3.457$, $p < .10$). The mean level of response entropy for low CC words was .618, and for high CC words it was .634 ($F(1, 39) = .509$, $p = .480$). The mean entropy for low-density words was .625, and for high-density words it was .627 ($F(1, 39) = .002$, $p < 1.0$).

Significant effects for response entropy were observed, however, across levels of CC and density in the 10-channel condition. A significant interaction between the two variables was observed as well. As expected, a higher level of response entropy was observed in high-density words compared to low-density words (mean HD = .531, and mean LD = .721; $F(1, 39) = 16.772$, $p < .001$). These data are consistent with the predictions of NAM (Luce & Pisoni, 1998) where high-density words are inhibited more by similar sounding phonological neighbors than low-density words. The higher level of inhibition and lexical competition in dense neighborhoods could potentially cause listeners to make a wider variety of errors. We observed higher levels of response entropy in Low CC words than in high CC words (mean LCC = .772, mean HCC = .480; $F(1, 39) = 41.591$, $p < .001$). This result was unexpected considering listeners respond more accurately to low CC words. We also observed a significant interaction between CC and density ($F(1, 39) = 16.13$, $p < .001$).

A trend similar to the 10-channel condition was observed under 12-channels. Listeners demonstrated higher levels of uncertainty for high-density words relative to low-density words (mean HD = .681 and mean LD = .570; $F(1, 39) = 7.747$, $p < .01$). Likewise, we observed the opposite trend for CC that we observed in the 10-channel condition (mean HCC = .502, mean LCC = .749; $F(1, 39) = 37.51$, $p < .001$).

Discussion

Analysis of Words Correct

In the 8-channel condition in Experiment 3, percent correct identification was low, suggesting that listeners were not able to reliably recognize words, but were guessing based on limited phonological information. The finding that listeners used guessing strategies rather than engaging in lexical access could explain why we did not observe the predicted results for CC: i.e., that low CC words would be recognized more accurately than high CC words. It is possible that if listeners had more exposure to the stimuli in the 8-channel condition, the pattern of results might be different and perhaps match the data from the 10 and 12 channel conditions where listening conditions improved. Observing the strategies used for recognizing spoken words after listeners adapt to highly degraded speech would be an interesting direction for future research.

The results from the 10 and 12-channel conditions confirmed the hypothesis that low CC words would be identified more accurately than high CC words. These results were based on predictions from Experiment 1 where listeners responded to low CC words more quickly because there was less competition from similar sounding words in the lexicon than there was for high CC words.

Analysis of Sub-Lexical Segments

The second major set of analyses concerned the response properties of the onset, nucleus, and coda. We examined whether accurate identification for each position differed as a function of CC or density. Another purpose of Experiment 3 was to investigate in which part of the word more errors were made.

We observed more accurate identification scores for onsets in low CC words than high CC words in 8, 10, and 12-channels. The data indicate that a lower degree of connectivity among the neighbors of a word allows listeners to more accurately identify the onset cluster. This reasoning is consistent with the results from Experiment 1 showing faster reaction times for low CC words, and the results from Experiment 3 showing more accurate identification for low CC words.

The data from the responses in the nucleus position differed from the onset position. Listeners correctly identified the nucleus in high CC words more accurately than the nucleus in low CC words in the 8-channel condition. Neighborhood density also significantly affected identification rates, where high-density words were more accurately identified than low-density words. The fact that density had the strongest effect in the coda position, regardless of the number of channels, suggests that when density is a factor in the recognition process, the effect is most salient for deletions, additions, or substitutions at the end of words.

General Discussion

The goal of this project was to investigate how the connectivity of a word's phonological neighbors in its subspace affect the reaction time and accuracy of spoken word recognition. This project extended the assumption of relational word recognition by embedding it within a graph theoretical framework to empirically test the proposal that the representation of words in memory can be modeled as a complex system. In order to obtain support for the hypothesis that the representation of words in memory might share properties with other complex systems, we carried out several behavioral studies to determine whether a word's CC affects spoken word recognition.

The theoretical motivation for the present set of experiments was based on the proposal that the mental lexicon can be viewed as a multidimensional space. Treisman (1979) described *partial identification theory* as a model for how the search of the lexical space is carried out by listeners. Partial identification theory assumes that only a subspace of the lexicon is searched, where the size of the subspace depends on the quality of the listening conditions. Recall that this is what distinguishes partial identification theory from Luce's universal forced choice model (1959). Also, when listening conditions are optimal, the subspace searched by the algorithm is very constrained, perhaps including the stimulus word and its immediate phonological neighbors. Under highly degraded listening conditions, the size of the subspace becomes much larger, and if conditions are degraded enough, the size of the subspace might well include most of the words in the lexicon. Under these listening conditions, the size of the subspace selected by partial identification theory would approximate Luce's universal forced choice model. Thus, one would predict that when listening conditions become severely degraded and listeners begin to use unconstrained guessing strategies, the word frequency of listener's incorrect responses would increase. In short, if listeners were searching a subspace consisting of a significantly large portion of the lexicon, their incorrect responses would generally be high frequency words because words with a higher frequency of occurrence in the language are more likely to be generated as responses to degraded and underspecified signals.

While the primary focus of these experiments was on reaction time and accuracy data, the incorrect responses from Experiment 3 provided information about structure in the lexicon. In the 8-channel condition in Experiment 3, listening conditions were highly degraded. The fact that listeners were biased to generate high frequency words as incorrect responses indicates that they were using a forced-choice decision rule over a very large subspace of the lexicon. The bias of listeners to generate high frequency error responses under degraded listening conditions replicated the previous results of Pollack et al. (1960).

The data from Experiment 3 also show that as the number of spectral channels increased from 8 to 10 to 12, the word frequency of listener's error responses decreased, suggesting that the subspace of the lexicon being searched was more highly constrained because more reliable stimulus information could be obtained from the signal. Therefore, the observation in Experiment 3 that the relation between frequency of the stimulus and frequency of the error response changed as listening conditions improved was

contrary to the predictions made by models of spoken word recognition that assume pure guessing strategies and the underlying assumption of acoustical or structural equivalence.[4] The results from the 10 and 12 channel conditions in Experiment 3 were consistent with Treisman's partial identification theory, which assumes that as listening conditions improve, the lexical neighborhood or subspace becomes more refined, reducing the bias to generate high frequency words as incorrect responses.

The results obtained in Experiment 1 using a same-different discrimination task suggest that increasing the level of CC among lexical neighbors slows the discrimination process down. Experiment 2, while methodologically flawed in the sense that listeners were not reliably recognizing words, showed that CC and neighborhood properties had different effects under degraded listening conditions, where listeners were using guessing strategies. That listeners were using guessing strategies was evident from the results obtained in the 8-channel condition in Experiment 3 in which the mean percent correct identification of words was under 10 percent. That is, it was likely that listeners were drawing upon a large number of potential lexical candidates in a large subspace of the lexicon based on partial phonological information. We also observed in the 8-channel condition in Experiment 3 that when listeners made errors, their incorrect responses were biased toward high frequency words, a prediction derived from previous studies (Pollack et al., 1960).

The results from the 10 and 12-channel conditions in Experiment 3 replicated the general pattern of results observed in Experiment 1. Putting the results from Experiments 1 and 3 together, it is important to begin considering various models that describe how the lexical effects of CC on the word recognition process operate. As discussed in the introduction, previous models of spoken word recognition have not addressed how the global properties of the mental lexicon affect spoken word recognition. Neither the "urn model" (Pollack et al., 1959; see also Oldfield, 1966) nor Logogen theory (Morton, 1979) explicitly define how words might be related to or connected to one another. Since current models in the field that assume relational word recognition like NAM, TRACE, and Cohort Theory have not described how global lexical variables operate and affect word recognition and retrieval, it is important to begin exploring how graph theoretical variables such as CC, which are integral to complex systems, affect word recognition in the context of a particular model. What type of model could predict these effects of CC, given the computational analyses carried out by Vitevitch (2004), Gruenenfelder and Pisoni (2006), as well as the behavioral results obtained in this current study, including the reaction time data generated in Experiment 1, and the results generated by the perceptual identification Experiment 3? The global structure and topology of the lexicon, including the interconnectivity of words in a lexical subspace is an important structural parameter affecting spoken word recognition.

Summary and Conclusions

Recent research on natural and artificial complex systems provided the theoretical motivation for this study. In the present set of experiments, we found effects of the graph theoretical variable CC on spoken word recognition. The motivation for studying this variable also came from several models that assume that spoken words are recognized relationally, and that the lexicon can be represented as a multidimensional acoustical space. Recent computational studies of the effects of graph theoretic variables on spoken word recognition also suggested the usefulness of this approach. More generally, our goal was to examine how the global structure and topology of words in the mental lexicon affect spoken

[4] See Treisman (1979) for a discussion on this topic. He argued that sophisticated guessing theory relies on the argument that the acoustical and structural properties of high and low frequency words are identical (Pollack et al., 1960). Since the phonological properties of high and low frequency words were believed to be identical, there was no need for the listener to focus on a particular "subspace" of the lexicon. In the introduction, there was a discussion regarding how Landauer and Streeter (1973) challenged this assumption.

word recognition. The three behavioral experiments analyzing correct and incorrect responses represent preliminary efforts to demonstrate the psychological reality and potential importance of CC as a new variable that affects spoken word recognition and performance.

Because we found that CC affects spoken word recognition in both same-different discrimination and perceptual identification tasks, it is important for future models of spoken word recognition to account for these new results showing that not all phonological neighborhoods of similar size are equal in their effects of the recognition process. Structural properties within a neighborhood including CC affect word recognition as well. Models motivated by the theoretical foundations of complex systems, like the general spreading activation model proposed here, should be developed to account for these findings.

References

Clark, H.H. (1973). The language-as-fixed fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335-359.

Crystal, D. (1980). *A first dictionary of linguistics and phonetics*. London: Andre Deutsch.

Eukel, B. (1980). A phonotactic basis for word frequency effects: Implications for automatic speech recognition. *Journal of the Acoustical Society of America, 68*, S33.

Gerstman L.J., & Bricker, P. (1960). Word frequency effects in learning unknown message sets. *Journal of the Acoustical Society of America, 32,* 1078-1079.

Goldiamond, I. & Hawkins, W.F. (1958). Vexierversuch: The log relationship between word frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology, 56,* 457-463.

Gruenenfelder, T.M. & Pisoni, D.B. (2006). Modeling the mental lexicon as a complex System: Some preliminary results using graph theoretic measures. Unpublished Manuscript.

Howes, D. (1957). On the relation between intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America, 29*, 296-305.

Kucera, F., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Landauer, T.K. & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for common and rare words. *Journal of Verbal Learning and Verbal Behavior, 12,* 119-131.

Levelt, W.J.M. (1978). A survey of studies of sentence perception. In W.J.M. Levelt & G.B. Flores d'Arcais (Eds.), *Studies in the perception of language*. New York: Wiley. Pp. 1–73.

Lively, S.E., Pisoni, D.B. & Goldinger, S.D. (1994). Spoken Word Recognition: Research & Theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, New York: Academic Press. Pp. 265-301.

Luce, R. D. (1959). *Individual Choice Behavior*. New York; Wiley.

Luce, P.A., Pisoni, D.B. (1998). Recognition of spoken words: The Neighborhood Activation Model. *Ear & Hearing, 19,* 1-36.

Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. A tutorial Review. In H. Bouma & D.G. Bouwhis (Eds.), *Attention and Performance X: Control of Language Processes.* Hillsdale, NJ: Erlbaum. Pp 125-150.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71-102.

Marslen-Wilson, W.D. (1990). Activation, competition, and frequency in lexical access. In G.T.M. Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic Computational Perspectives*. Cambridge, MA: MIT Press. Pp. 148-172

McClelland, J.L. & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

Morton, J. (1979). Word recognition. In J. Morton & J.C. Marshall (Eds.), *Structures and Processes*. Cambridge: MIT Press. Pp. 108-156.

Nusbaum, H.C., Pisoni, D.B. & Davis, C.K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10*, (Pp. 357-376), Bloomington, IN: Speech Research Laboratory.

Oldfield, R.C. (1966). Things, words and the brain. *Quarterly Journal of Experimental Psychology, 18,* 340-353.

Pollack, I., Rubenstein, H, & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31,* 273-279.

Pollack, I., Rubenstein, H. & Decker, L. (1960). Analysis of incorrect responses to an unknown message set. *Journal of the Acoustical Society of America, 32,* 454-457.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin, 125*, 410-436.

Savin, H.B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America, 35*, 200-206.

Shannon, R.V., Fan-Gang, Z., Kamath, V., Wygonski, J., Ekelid M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Speech and Hearing Lab Neighborhood Database. Retrieved from http://128.252.27.56/Neighborhood/Home.asp

Steyvers, M. & Tenenbaum, J.B. (2005). The large-scale structure of semantic networks: Statistical analysis and a model of semantic growth. *Cognitive Science, 29*, 41-78.

Trask, R.L. (1996). *A dictionary of phonetics and phonology*. Routledge: London.

Treisman, M. (1979). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning & Verbal Behavior*, *17,* 37-59.

Vitevitch, M.S. & Luce, P.A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9,* 325-329.

Vitevitch, M.S. & Luce, P.A. (1999). Probabilistic Phonotactics and Neighborhood Activation in spoken word Recognition. *Journal of Memory and Language, 40,* 374-408.

Vitevitch, M.S., & Luce, P.A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavioral Research Methods, Instruments & Computers, 36,* 481-487.

Vitevitch, M.S. Luce, P.A. Charles-Luce, J., & Kemmerer, D. (1996). Phonotactic and metrical influences on adult ratings of spoken nonsense words. Proceedings from the *Fourth International Conference on Spoken Language,* Volume 1, Issue, 3-6 Oct 1996 Pp. 82 – 85.

Vitevitch, M.S. (2004). Phonological neighbors in a small world: What can graph theory tell us about word learning? Unpublished Manuscript.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of small-world networks. *Nature*, 393, 440-442.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Implementing and Testing Theories of Linguistic Constituency I: English Syllable Structure[1]

Vsevolod Kapatsinski

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] I would like to thank the NIH for financial support through Training Grant DC-00012 and Research Grant DC-00111 to David Pisoni. Many thanks to Luis Hernandez for his help in creating the experimental program and to Adam Buchwald for helpful comments on an earlier draft of this paper.

Implementing and Testing Theories of Linguistic Constituency I: English Syllable Structure

Abstract. This paper proposes and tests an experimental method to evaluate models of linguistic constituency, including: 1) Connections within constituents are stronger than connections spanning constituent boundaries, 2) A constituent is more likely to be parsed out of the signal than a non-constituent (i.e., constituents are processing units), and 3) Both constituents and non-constituents are units, with constituents simply having higher frequency than non-constituents. The method, XOR learning, is designed to distinguish between associability of a whole and associability of its parts. Subjects learn to associate the whole with a different response than the response both of its parts have been associated with. We apply the method to the onset-rime organization of English CVC syllables with a lax vowel, showing that native English speakers can learn rime-affix associations but not body-affix associations. This difference in associability between bodies and rimes is observed in the absence of associability differences between onsets and codas, the only parts that bodies and rimes do not share. Competing theories of linguistic constituency are implemented in a Hebbian framework where parts of the syllable extracted from the signal become associated with affixes they co-occur with. Assuming automatic phonemic categorization, experimental results are explained only by models that assume that rimes and bodies differ in the level of activation they have during training. Applications of the experimental method and its variants to linguistic constituency in other domains are discussed.

Introduction

Theories of Constituency

This paper introduces a method to distinguish between different theories of linguistic constituency. Thus, the question we would like to address is: what are constituents? What does it mean to say that in an English syllable consisting of an onset, a nucleus, and a coda, the nucleus forms a constituent with the coda and not with the onset?

The traditional answer to this question in linguistic theory has been that the rime (nucleus+coda) is allocated a node in the tree structure while the body (onset+nucleus) is not (e.g., Fudge, 1987; Selkirk, 1982). A tree structure is a type of a network and, like in any network, it consists of nodes connected by links. By definition, then, a node is something that can be connected to/associated with something else. Thus, in the traditional view of linguistic constituency, constituents can be associated with other units, i.e., **constituents are associable, while non-constituents are not**. Thus, under this view, if the rime is a constituent while the body is not, rime-affix associations should be learnable while body-affix associations should not be.

In order to associate a unit X with another unit Y, the two units must be extracted from the signal. Thus, things that are associable must be extracted from the signal. In other words, something that is associable must be a **processing unit**. Under the traditional view of constituency, then, **constituents are processing units** (cf. Cutler et al., 2001; Mehler, 1981). That is, at the very least, if the rime is a constituent and the body is not, the rime should be more likely to be extracted from the acoustic signal than the body is.

An alternative to the tree-structural view of constituency is the dependency-based view, applied to syllabic constituency by Vennemann (1988) and Anderson and colleagues (e.g., Anderson & Ewen, 1987). Under this view, neither constituents not non-constituents are allocated nodes. Rather, **connections between parts of a constituent are stronger than connections that cross constituent boundaries**.

Under this view, to say that the rime is a constituent while the body is not means to say that the nucleus is connected to the coda more strongly than to the onset. The dependency-based view does not straightforwardly predict a difference in associability between constituents and non-constituents. Any such difference would be an epiphenomenon, deriving from differences in associability between parts that the constituent and the non-constituent do not share. Thus, in the case of the body, the rime would be expected to be more associable than the body if and only if the coda is more associable than the onset.

Finally, processing units may differ in how associable they are, depending on factors like frequency and the cumulative strength of associations they already have (e.g., Kamin, 1969; Moder, 1992). That is, nodes may differ in associability. The associative learning literature indicates that **frequent stimuli are harder to associate than infrequent stimuli** (the phenomenon known as pre-exposure, or desensitization effects, see Hall, 2003, for review). In the linguistic literature, Bybee and Brewer (1980) and Moder (1992) have argued that frequent words have weaker connections to similar words than infrequent ones. Thus, if a rime is more frequent than a body, the rime may be expected to be less associable than the body. Given this potential influence on associability, rimes and bodies may be equally likely to be parsed out of the signal (i.e., constituents and non-constituents may be equally **salient**) and still differ in associability. Equal salience is proposed within full-listing models in which **all possible segment strings (up to a particular length) are parsed out of the signal** (Skousen, 1989; see also Bod 1998 for an analogous approach to syntax in which all possible subtrees are parsed out).

In this paper, we will compare associability of rimes to associability of bodies and then implement the various theories of syllable structure in a common framework to see which can account for the experimental data.

XOR Learning

As discussed in the previous section, the tree-structural view of constituency predicts that rimes should be more associable than bodies in English while the dependency-based view claims that any such differences should be attributable to differences in associability between onsets and codas. Thus, to distinguish between the two alternatives, it would be helpful to have a way to train subjects on body-affix or rime-affix associations while controlling segment-affix associability. Such a way is provided by XOR learning.

A classic **XOR** (exclusive-or) distribution is one in which stimuli containing either A or B are classified as being of type X, while stimuli containing both A and B and stimuli containing neither A nor B are classified as being of type Y (A, not B → X; not A, B → X; A, B → Y; not A, not B → Y). This distribution has been most famously used by Minsky and Papert (1969) to argue against two-layer connectionist networks as models of cognition. Minsky and Papert showed that such purely distributed models cannot represent the XOR relation because if A is associated with X and B is associated with X, AB necessarily is, since there is no representation for the complex unit AB. Algorithms for learning in three-layer networks were later introduced in part to handle XOR. In such networks, some node(s) in the hidden layer, which intervenes between the input and output, are activated when AB is presented and not when A or B is presented in isolation, allowing the network to respond to AB in a different way than it responds to A or B.

In the present experiment, we used a modified version of the XOR distribution in which A is associated with X, B is associated with X, while C is associated with Y as is D but AB is associated with Y while CD is associated with X, as shown in figure 2. This distribution was chosen to avoid asking participants to form a category defined solely in negative terms (e.g., syllables in which the coda is not B and the nucleus is not A). In the distribution in Figure 2, all the dependencies involve specific segments or bigrams present in the stimuli.

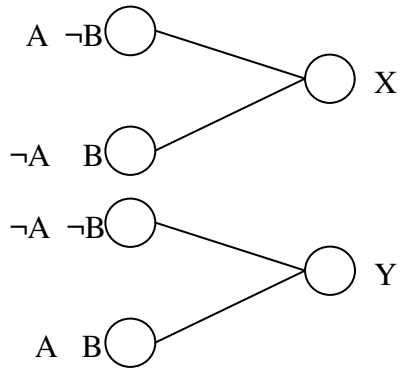


Figure 1. A classical XOR task where one response (X) is required when presented with either A or B but not both. If neither A nor B is presented or both A and B are presented, a different response, Y, is required. Here A might be an onset, and B might be a nucleus or A might be a nucleus with B being a coda.

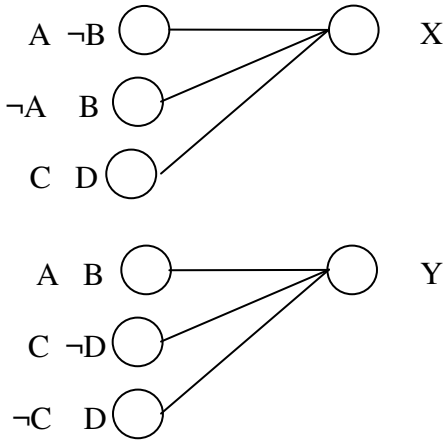


Figure 2. The task used in the present experiment. One response is associated with A in the presence of something other than B, with B in the presence of something other than A, and with CD. The other response is associated with C in the presence of something other than D, with D in the presence of something other than C, and with AB. Here A and C may be onsets with B and D being nuclei and AB and CD being bodies or A and C may be nuclei, B and D codas, and AB and CD rimes. AD and CB are never presented.

In the experiment presented here, X and Y are affixes. They can either precede or follow the stem. A, B, C, and D are individual segments. For some subjects, AB and CD are bodies, while for others they are rimes. We are interested, then, in how easily subjects learn AB-Y and CD-X associations when AB and CD are rimes compared to the case in which AB and CD are bodies.

The experiment is divided into two stages. In the first stage, AB and CD are never presented and subjects thus learn that A is paired with X, as is B, and that C and D are paired with Y. Thus, subjects learn segment-affix associations. By looking at subjects' accuracy with novel syllables containing familiar onsets or codas after the first stage, we can compare associability of onsets to associability of codas.

Since bodies and rimes share nuclei, all subjects learn nucleus-affix associations during stage I. Therefore, the results of stage I allow us to assess between-subject differences in learning rate. Thus, we can ensure that subjects who are assigned to learn rime-affix associations are not simply better learners than those that are assigned to learn bodies.

In the second stage, AB and CD tokens are introduced and subjects learn that AB is (surprisingly) associated with Y while CD is paired with X. Subjects are then asked to predict the affixes of unfamiliar syllables containing the now familiar AB and CD. The results of Stage II training indicate how easy it is to learn rime-affix vs. body-affix associations. Since we know how easy it is to learn coda affix vs. onset-affix associations from the results of Stage I, we can determine whether differences in associability between bodies and rimes can be explained by differences in associability between onsets and codas as predicted by dependency-based models of syllable structure.

Evidence for Syllabic Constituency

In order to investigate the relationship between constituency and associability, we need a case where constituency is uncontroversial. Such a case is provided by English CVC syllables with lax vowels. There are a number of reasons to believe that such syllables have an onset-rime structure (/C/ /V C/) and not body-coda structure (/C V/ /C/). That is, the vowel goes with the following consonant and not the preceding one. Furthermore, there is some evidence for the existence of the parts of syllabic constituents, the segments.

Strong evidence has been provided for the involvement of syllable structure in visual word recognition. In priming experiments reported by Ferrand et al. (1996) for French, Carreiras and Perea (2002) and Alvarez et al. (2004) for Spanish, and Ashby and Rayner (2004) for English the target word started with either a CVC or a CV syllable. The prime was presented visually and had either the form CV**** or CVC***. It was presented so fast that the subjects did not consciously notice its identity. All segments of the prime were present in the target. For instance, the primes may be *pa*** and *pas**** and the targets may be *pasivo* and *pastor* (Carreiras and Perea, 2002).[2]

If all that mattered for phonological priming were the number of segments or letters shared between the prime and the target or the duration of the shared part, we would expect CVC primes to produce more priming than CV primes for both types of targets. However, both studies showed a reliable interaction: while CVC primes produced more priming than CV primes for CVCCVC targets (*pas**** primed *pastor* more than *pa**** did), CV primes produced more priming than CVC primes for CVCVCV targets (*pa**** primed *pasivo* more than *pas**** did). These results follow directly from the syllable structure of the targets: *pas* shares a syllable with *pastor* but not with *pasivo*, while *pa* shares a syllable with *pasivo* but not with *pastor*. These studies provide convincing evidence that subjects are sensitive to the syllable structure of the word, although ambiguity remains regarding whether the sharing of syllables or syllabic constituents is at issue. While *pas* shares an onset and a rime with *pastor*, it only shares an onset with *pasivo*. Nonetheless, these data constrain possible models of syllable structure in that there must be nodes for either syllables or syllabic constituents, i.e., it is not sufficient to have only segment units.

Evidence for the onset-rime structure is provided by the fact that categorical co-occurrence restrictions involve VC's and not CV's in English (e.g., Fudge, 1987; Selkirk, 1982). In particular, lax vowels in English require a consonant to follow them while they do not require a preceding consonant.

[2] Alvarez et al. (2004) have addressed the concern that the effect is orthographic, rather than phonological in nature: the syllable priming effect did not diminish when the spelling of the first syllable was not shared by the prime and the target (*vi.rel-vi.rus* vs. *bi.rel-vi.rus*).

This also means that in a syllable with a lax vowel, the body is not a possible word while the rime is, which contributes to the constituency difference between the body and the rime in such syllables.

In addition, Kessler and Treiman (1997) and Lee (2006) have shown that there are statistical reasons for grouping vowels with codas in English: given the vowel, the coda is somewhat predictable while the vowel is not predictable given the onset. Treiman et al. (2000) found that wordlikeness judgments for nonsense CVC's are affected by the probabilistic constraints on vowel-coda co-occurrence.

Treiman (1983, 1986) and Derwing (1987) found that when English speakers are asked to use the beginning of the first (C)CVC(C) word and the end of the second (C)CVC(C) word to form a new word, they use the onset of the first word and the rime of the second and not the body of the first and the coda of the second. Treiman et al. (2000) found that the tendency is a little stronger with high-frequency rimes than with low-frequency rimes (although above 90% in both cases).

Treiman and Danis (1988) showed that when subjects are asked to recall a long list of words, they tend to make errors that are novel recombinations of previously presented onsets and rimes and not bodies and codas. In addition, Nelson and Nelson (1970), Vitz and Winkler (1973), Derwing and Nearey (1986), Bendrien (1992), and Yoon and Derwing (2001) found that CVC words sharing rimes are perceived to be more similar than words sharing bodies by English speakers in sound similarity judgment tasks. However, Geudens et al. (2005) have called the relevance of these results for testing syllabic constituency into question by showing that Dutch speakers judge syllables sharing rimes to be more similar than syllables sharing bodies yet show no preference for recombining onsets and rimes as opposed to bodies and codas in a serial recall task.

Lee (2006) has shown that the statistics favor the body in Korean, which explains why Korean speakers, unlike English speakers, tend to produce body-coda recombinations in serial recall when presented with syllables in which the nucleus co-occurs with the coda as strongly as it co-occurs with the onset. Furthermore, if English speakers are asked to memorize atypical syllables in which the nucleus co-occurs with the onset more than with the coda, they too tend to produce body-coda recombinations. Finally, Korean speakers presented with syllables in which the nucleus co-occurs with the coda more than with the onset, which are not typical in Korean, tend to produce onset-rime recombinations. These results constrain models of syllabic constituency in that the difference between bodies and rimes in English cannot be due to the fact that the beginning of the rime follows the beginning of the body within the syllable but rather must be due to the statistics of between-segment co-occurrence. In this paper, we will only use syllables in which statistics of co-occurrence favor the onset-rime division.

A number of studies have provided support for the psychological reality of segments. Vitz and Winkler (1973) found that sound similarity judgments for a pair of words had a correlation of 0.9 with number of mismatched segments. Kapatsinski (2006) observed that the mean likelihood of interrupting a word before replacing it in spontaneous speech production is very strongly correlated with log number of segments in the word (r^2=.991). Stemberger (1983) and Jaeger (2005), among others, find that most speech errors involve substitutions of single segments. Boothroyd and Nittrouer (1988), Nearey (1990, 2003), Benki (2003), and Felty (2007) found that accuracy of identification of nonsense syllables in noise is highly accurately approximated by a linear combination of average identification accuracies of the component segments. Hockema (2006) found that word segmentation based on segment transitions would be highly successful in English. Finally, some evidence for units smaller than syllabic constituents is provided by the fact that not all rimes are equally acceptable in English, e.g., */aʊp/.

Given the existence of evidence for both the onset-rime division of the types of syllables used in the present study and the status of segments as processing units, we can ask whether the rime is a processing unit as well or if differences in between-segment connection strength are sufficient to account for the structure of English syllables. Before getting into the main part of the paper, it is important to note that the present study concerns the nature of syllabic constituency in English syllables with lax vowels

that are, furthermore, morpheme-initial. There are a number of factors that make the rime a better constituent than the body in such cases which may not distinguish other types of syllabic constituents from other types of non-constituents. For instance, bodies that end in a lax vowel is not a possible word, while bodies that end in tense vowels are possible words. In addition, Davis (1989) has argued, based on speech error evidence, that the word-initial onset is particularly poorly integrated into the rest of the word. Thus, the present paper intends to show that in a case in which there is an extremely clear constituency difference between the body and the rime, the rime is much more associable than the body. However, this may not be the case with other types of rimes and other types of bodies. The nature of a proposed constituency difference is an empirical question that needs to be investigated separately in each case. There is no a priori reason to believe that the same model should be used to model constituency in such disparate domains as syntax and phonology or even in different contexts within phonology. This paper provides one way of selecting an appropriate model for a particular type of linguistic constituent.

The Experiment

Methods

The Paradigm. In our experiment, we wanted to dissociate constituent associability from associability of the component segments. Thus, we needed the subjects to learn that a whole is associated with a different response than either of its parts. In other, words, if AB is associated with Y, then A is associated with X and B is associated with X.

Native English speakers were randomly assigned to the four experimental groups shown in Table 1. As can be seen from the table all subjects were exposed to co-occurrences between the vowels /æ/ and /ʌ/ and the affixes /mɪn/ and /num/. However, /mɪn/ and /num/ came after the stem (were suffixes) for groups II and IV but they came before the stem (were prefixes) for groups I and III. In addition to being exposed to vowel-affix correlations, subjects in groups I and II were exposed to rime-affix and coda-affix correlations while subjects in groups III and IV were exposed to body-affix and onset-affix correlations.

Group	Associate	Part relations	Whole relations
I	Rimes & prefixes	num- CæC num-CV∫ mɪn-CʌC mɪn- CVg	mɪn- Cæ∫ num-Cʌg
II	Rimes & suffixes	CæC-num CV∫-num CʌC-mɪn CVg-mɪn	Cæ∫-mɪn Cʌg-num
III	Bodies & prefixes	num-CæC num-∫VC mɪn- CʌC mɪn- gVC	mɪn- ∫æC num-gʌC
IV	Bodies & suffixes	CæC-num ∫VC-num CʌC-mɪn gVC-mɪn	∫æC-mɪn gʌC-num

Table 1. The experimental design: what do subjects have to learn?

Two things to note about the design are that each subject has to learn two XOR distributions as in Figure 2, and that the temporal relation between the affix and the stem (prefixation vs. suffixation) is manipulated independently of whether the to-be-associated part of the stem is a constituent.

If the subjects were exposed to a single XOR distribution (e.g., num-CæC, num-CV∫, but mɪn-Cæ∫) they would not need to infer anything about associations of the parts (/æ/ and /∫/ in the example) since all regularities can be defined in terms that only involve /æ∫/: the presence of /æ∫/ indicates the presence of /num/ while its absence indicates the presence of /mɪn/.

The experimental design shown in Table 1 allows us to control for locality effects and possible differences in associability between suffixes and prefixes. Thus, if prefixes are more associable than suffixes, associations involving prefixes should be easier to learn than those involving suffixes, regardless of whether the rime of the stem or the body of the stem is involved. Alternatively, association between adjacent parts of the speech stream may be easier to learn than an association between non-adjacent parts. Such a result has been obtained in statistical learning of segment and syllable co-occurrences (Newport & Aslin, 2004; Bonatti et al., 2005). If this is the case, body-prefix and rime-suffix associations should be more learnable than body-suffix and rime-prefix associations.

The Sequence of Training and Testing Stages. Participants were first trained on vowel-affix and consonant-affix co-occurrence relations. During the 4 minute training session, they listened to stems containing the relevant consonants and vowels (but not to ones containing the relevant rimes or bodies) paired with affixes. They were instructed not to press any buttons. The stimuli were arranged so that every stem-affix combination ("**word**") was followed by another word whose stem differed from the stem of the first combination by one segment (either a consonant or a vowel), e.g., /bæ∫-mɪn/ followed by /bɪ∫-num/ or /bɪg-num/ followed by /bug-num/. The words in these minimal pairs differed in the affix they took on half of the trials. Thus, the difference between the words in a minimal pair was irrelevant for affix choice in half of the pairs (as in /bɪg-num/ followed by /bug-num/). This balancing ensured that the training does not place stimuli sharing the rime next to each other more often than it places stimuli sharing the body next to each other.

This training session was followed by a testing session in which subjects were presented with stems they have already heard but they heard Gaussian noise in place of the affix. The noise had the same amplitude contour and duration as the average of the two affixes (/mɪn/ and /num/). The subjects were instructed to guess which affix has been replaced with noise. Once they made their guess, the correct stem+affix combination was pronounced. In the subsequent generalization block, the subjects heard novel stems paired with noise. No feedback was given.

This initial stage of the experiment, "**training on parts**", leads to several critical comparisons. We can compare the associability of onsets and codas, allowing us to assess whether any differences found between bodies and rimes can be explained by differences between onsets and codas. Further, to ensure that subjects assigned to learn rime-affix dependencies are not just (by chance) better learners than those assigned to learn body-affix dependencies, we can compare individual differences in learning ability across subject groups by looking at how well the subjects are learning vowel associations. Third, we can compare the associability of prefixes and suffixes and assess the existence of locality effects, since the onset is adjacent to the prefix and the coda is adjacent to the suffix.

Finally, we can determine base rates of generalization from a set of rimes or bodies to a novel rime/body that consists of segments that have been presented and associated with the same response but have never been presented together. The last piece of information has dual significance: 1) we may expect that there will be less generalization to an unfamiliar rime than to an unfamiliar body if subjects are spontaneously storing rime-affix but not body-affix pairings during this stage, and 2) reaction to stimuli containing the crucial wholes (/∫æ/ and /gʌ/ or /æ∫/ and /ʌg/) prior to training on those wholes provides us with a baseline level of accuracy to which the accuracy level following training on wholes can be compared.

After the generalization block, subjects go through the training-feedback-generalization cycle again. The only difference is that now they receive training on both parts and wholes. Thus subjects in groups I and II learn rime-affix, vowel-affix and coda-affix associations while those in groups III and IV learn body-affix, vowel-affix and onset-affix associations.

Table 2 summarizes how much total training and testing subjects get for each type of potential constituent. Since previous research (Bonatti et al., 2005; Creel et al., 2006) suggests that vowels are less associable than consonants, subjects received more vowel training than consonant training.

Constituent	Training Trials		Feedback Trials		Total		Generalization Trials	
	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II
C	32	20	20	10	52	30	16	8
V	32	24	20	10	52	34	28	22
Body/rime	0	46	0	10	0	56	12	22
Total	64	92	40	30	104	122	56	52

Table 2. How much practice do they get?

Participants. The participants were introductory psychology students who received course credit in exchange for their participation. They were not rewarded for accuracy or speed. All participants reported being native English speakers with no history of speech, language, or hearing impairment. There were 17 participants in each of the four subject groups (rime-prefix, rime-suffix, body-prefix, and body-suffix).

Procedures. The stimuli were presented over headphones at a comfortable listening level. Subjects were seated at a testing station consisting of a computer on a desk surrounded by cubicle walls. Instructions appeared on the screen between the stages of the experiment. The instructions are presented in the appendix. The subjects could take as much time as they needed to read the instructions. Subjects were randomly assigned to one of the four subject groups and one of the four computers in the room.

Materials. Stems and affixes were recorded separately and concatenated using Praat. There are a few noteworthy things about the stimuli: 1) the segments that begin suffixes are the same segments that end prefixes (nasals), 2) prefixes and suffixes are acoustically identical and not prosodically integrated with the stem, 3) bodies are less frequent than rimes, while onsets and codas are similar in frequency, although onsets are somewhat more frequent than codas, and 4) stimuli used for testing do not include syllables used for training and include consonants and consonant clusters that are not used in training.

Controlling for Frequency. Convergent estimates of the frequencies of various rimes, bodies, onsets, vowels, and codas were obtained from the MRC Psycholinguistic Database (Coltheart et al., 1981; http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm), Kessler and Treiman (1997, http://brettkessler.com/SyllStructDistPhon/), and the Hoosier Mental Lexicon (Nusbaum et al., 1984; http://128.252.27.56/Neighborhood/SearchHome.asp). In order to minimize physical differences between onsets and codas, certain consonants were excluded from consideration: 1) voiceless stops due to the presence of aspiration in onsets of stressed syllables but not in codas, 2) nasals since they have much more nasalizing influence on the preceding vowel than on the following one, 3) /r/ because of fusion with the preceding vowel, 4) /l/ because of vast differences in pronunciation in onset and coda positions, 5) /w/ and /j/ because of restriction to word-initial position, 6) /d/ and /z/ because of possible morphological interpretation in coda position. Affricates were also eliminated because they may be more internally complex.

Table 3 presents the data across databases. Both of the consonants selected for the experiment are relatively balanced in how frequently they are used in the onset vs. in the coda. If anything, the consonants are slightly more frequent in the onset. The databases display a remarkably high agreement regarding the distributions of the consonants.

Database Consonant	**MRC**			**Kessler and Treiman (1997)**			**HML**		
	On	Cd	%On	On	Cd	%On	On	Cd	%On
b	216	90	71	154	62	71	150	75	67
g	102	97	**51**	88	67	**57**	80	77	**51**
v	61	85	42	45	54	45	39	65	38
f	144	97	60	92	68	58	109	70	61
s	183	169	73	126	116	52	121	113	52
ʃ	89	70	**56**	65	44	**60**	70	54	**52**

Table 3. Type frequencies of the consonants in codas vs. onsets of monomorphemic CVC words (consonants eventually used are in bold)

Table 4 shows the frequency distributions for rimes and bodies involving the consonants derived from the Hoosier Mental Lexicon. Since the databases display such a high agreement rate, only HML results are shown. As can be seen from the table, there is evidence that the chosen vowels are linked to the coda more strongly than they are linked to the onset. The rimes in the study are more frequent than the bodies. Thus, formal and usage-based criteria for constituency converge for the stimuli used in the present study.

	gV	**Vg**	**%rime**	**ʃV**	**Vʃ**	**%rime**
æ	9	20	69	**8**	**21**	**82**
ʌ	**5**	**17**	**77**	5	12	71

Table 4. Type frequencies of rimes and bodies of monomorphemic CVC words for the chosen consonants and vowels in the Hoosier Mental Lexicon (chosen rimes and bodies in bold)

Prefixes and Suffixes. The syllables /mɪn/ and /num/ were used as both the prefix and the suffix. The syllables were chosen so that none of the rules had an obvious phonetic motivation. In addition, we ensured that the set of consonants adjoining the stem was the same for prefixes and suffixes. These consonants (nasals) were chosen to be relatively perceptible, and relatively unlikely to interfere with the perception of the adjacent stem consonant. In addition, we made sure that they do not cause perceptual resyllabification.

The Generalization Stimuli. The generalization stimuli were of four types: 1) stimuli that contained consonant clusters in the unattended position (the position not involved in the generalizations always contained one consonant during training), e.g., /plæʃ/ for subjects learning rime-affix associations or /ʃælp/ for those learning body-affix associations, 2) stimuli that contained /l/ as the consonant in the unattended position (the consonant did not occur during training), e.g., /læʃ/ for rime-trained subjects or /ʃæl/ for body-trained ones, and 3) stimuli in which the two consonants of the stem were identical, e.g, /zæz/, /gʌg/, /ʃɪʃ/. No differences between generalization stimulus types were found, hence the results reported later are averaged across generalization stimulus types.

Recording the Stimuli. All syllables involved in the study were produced by a single male native American-English speaker who was unaware of the purpose of the study. In addition to the stimuli used in the study, the speaker also produced a large number of distractors that did not involve the target rimes and bodies. The affixes were produced only once by the speaker. The speaker did not know that these syllables had any special status. Each syllable was produced in isolation in response to a visual prompt appearing on a monitor for a fixed amount of time.

Results

Bodies vs. Rimes

After Training on Wholes. As shown in Figure 3 and Table 5, subjects who were exposed to rime-affix correlations pressed the appropriate button in response to novel stimuli containing a relevant rime in about 70% of the cases, which is significantly above chance (50%) according to a one-sample t-test (t=5.955, df=33, p<.0005). That is, the subjects succeeded in learning to respond with /mɪn/ when presented with a syllable ending in /æʃ/ and to respond with /num/ when presented with a syllable ending with /ʌg/. By contrast, subjects who were exposed to body-affix correlations did not learn those correlations, responding at chance levels. That is, when presented with a syllable that began with /ʃæ/ or /gʌ/ the subjects responded with either /mɪn/ or /num/ with equal probability.

Analysis of variance with constituency, affix location, and correct response (mɪn vs. num) as independent variables and accuracy as a dependent variable showed the rime vs. body difference to be statistically significant (F(1,66)=123.431, p<0.0005) with a large effect size (Cohen's d=1.17) while there is no significant difference in associability between prefixes and suffixes (F(1,66)=.431, p=.517), no significant effect of correct response and consequently rime identity (mɪn vs. num, æʃ/ʃæ vs. gʌ/ʌg (F(1,66)=1.026, p=.315), and no significant interactions. Planned by-subjects and by-items t-tests confirmed the significance of the effect of constituency (by subjects: t=5.401, df=66, p<.0001, by items: t=13.445, df=42, p<.0001).

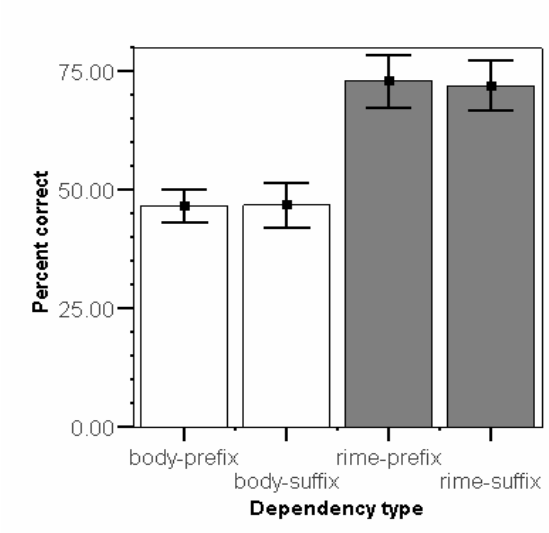


Figure 3. Rime associations are easier to learn than body associations (bars show means, error bars show by-subject standard errors) [3]

[3] By-item standard errors are smaller (+/-1.4%).

	rime-prefix	**rime**-suffix	**body**-prefix	**body**-suffix
%correct	73	72	47	47
Average across affixes	**73**		**47**	

Table 5. Subjects' generalization accuracy (% correct) on rime-affix vs. body-affix correlations after training.

These results suggest that rimes are more associable than bodies. At this point we can reject any explanation of these results that is based on differences in associability between prefixes and suffixes as well as an explanation based on locality effects. The rime is more associable than the body regardless of whether the subjects are learning prefix associations or suffix associations and thus also regardless of whether the affix is adjacent to the stem-internal segment sequence that it co-occurs with. In addition, lack of significant interactions with correct response indicates that individual bodies and rimes functioned similarly.

Prior to Training on Wholes. We need to determine whether the results obtained might be due to differences in the likelihood of generalizing consonant and vowel associations learned during stage I to novel rimes and bodies. This section provides the results from generalization trials following training on parts (prior to training on wholes).

Prior to training on stimuli that contain the relevant bodies and rimes, the subjects are trained to acquire associations of the consonants and vowels that compose those rimes and bodies. Since the whole is associated with a different response than both of its parts in XOR learning, training on parts alone should make stimuli containing the whole become associated with responses appropriate for stimuli that contain only one of the parts. If this happens, accuracy on wholes should be below chance prior to training on wholes. Table 6 shows the results. The rime-suffix condition is significantly different from the other conditions (the interaction of constituent and affix is significant: $F(1,64)=5.51$, $p=.019$).

%correct	**rime**-prefix	**rime**-suffix	**body**-prefix	**body**-suffix
Generalization 1	29	43	28	26

Table 6. Subjects' generalization accuracy (% correct) on rime-affix vs. body-affix correlations prior to training

Given this result, subjects who are about to learn rime-suffix dependencies need to improve by fewer percentage points than other subjects, including those learning body-affix dependencies, to reach the same level of performance following Stage II. In the remainder of this section we address this concern by showing that a subject's accuracy level following Stage I is in fact a very poor predictor of the same subject's accuracy level following Stage II indicating that differences in accuracy between groups following Stage I do not give rise to differences in accuracy following Stage II.

Table 7 shows bivariate Pearson correlations between each subject's performance on vowels, consonants and wholes during the various stages of the experiment. Within a training stage, accuracy on wholes has a strong negative correlation with accuracy on segments.[4] However, accuracy following Stage I (C1, V1, or W1) does not correlate with accuracy on wholes following Stage II (W2). This indicates that differences in the accuracy level prior to training on wholes cannot account for differences in accuracy

[4] This is expected because whenever a whole is perceived, its parts are perceived as well. Thus, when one hears something like /Cæʃ-mɪn/ one re-associates /æ/ and /ʃ/ with /mɪn/.

level following training on wholes. Thus, despite the fact that subjects are more accurate on rime-suffix associations than on other whole-affix associations prior to training, this cannot account for the difference between how well subjects perform on rime-suffix vs. body-prefix dependencies following Stage II training.

		W1	C1	V1	C2	V2
W1	r		-.623	-.428		
	sig.		.000	.000		
W2	r	.111	-.016	-.011	-.526	-.504
	sig.	.366	.897	.930	.000	.000

Table 7. Relations between each subjects' performance on various stimuli in different stages of the experiment (W = wholes, i.e., bodies or rimes, V = vowels, C = consonants, stage 1 precedes training on wholes, stage 2 follows).

Familiar vs. Novel Syllables. One possible interpretation of the better performance of subjects exposed to rime-affix dependencies compared to subjects exposed to body-affix dependencies is that all subjects are actually learning syllable-affix dependencies but that such pairings are easier to learn when syllables paired with the same affix are similar. Previous research has found that English speakers judge syllables sharing rimes to be more similar than syllables sharing bodies (Nelson & Nelson, 1970; Vitz & Winkler, 1973; Derwing & Nearey, 1986; Bendrien, 1992; Yoon & Derwing, 2001). Therefore, subjects could perform better when exposed to rime-affix dependencies than when exposed to body-affix dependencies if they are learning syllable-affix pairings in either case and there is no difference in associability between bodies and rimes (cf. Geudens et al., 2005). However, if subjects are learning syllable-affix pairings, they should perform better when presented with familiar syllables than when presented with novel syllables.

Table 8 compares subjects' performance on novel syllables (not presented during training) and familiar syllables. Separate ANOVA's were conducted for rime-affix and body-affix subject groups. There is no main effect of syllable familiarity ($F(1,66)=.002$, $p=.98$). Examination of Table 8 shows that the only subjects for whom there is a significant difference between familiar and novel syllables in the expected duration are subjects acquiring body-prefix associations. For rimes, the effect is in the other direction ($F(1,33)=5.433$, $p=.02$): subjects perform slightly better on novel syllables. Therefore, we can reject the hypothesis that our subjects are learning syllable-affix associations instead of rime-affix associations.

Constituent	**Rime**				**Body**			
Affix	Prefix		Suffix		Prefix		Suffix	
Stimuli are	Familiar	Novel	Familiar	Novel	Familiar	Novel	Familiar	Novel
%correct	71	74	66	71	60	49	49	49

Table 8. Testing with familiar vs. novel syllables

Consonants and Vowels

In this section we compare associability of onsets to associability of codas. This is necessary to show that the rime/body difference is not due to an onset/coda difference. We also examine whether the

subjects are really forming rime-affix associations and not just re-associating consonants and/or vowels with an inappropriate response by examining consonant and vowel associations after training on wholes.

Prior to Training on Wholes. Table 9 shows how accurate subjects were on consonant-affix and vowel-affix associations prior to training on wholes. The results were analyzed using an ANOVA with constituent, affix, segment type (consonant vs. vowel), and correct response as independent variables. There is a significant difference between consonants and vowels ($F(1,66)=48.108$, $p<.0005$): accuracy is lower for vowel-affix associations than for consonant-affix associations. In addition, there is a significant affix type by segment type interaction ($F(1,64)=16.908$, $p<.0005$) and a marginally significant segment type by correct response interaction ($F(1,64)=3.702$, $p=.054$). The difference between consonants and vowels is larger when suffix occurrence is being predicted than when prefix occurrence is, as shown in Table 9.

Association Type	**Local**		**Non-local**		**Local**		**Non-local**	
Associated	Coda	Vowel	Coda	Vowel	Onset	Vowel	Onset	Vowel
%correct	61	56	71	58	73	55	67	60

Table 9. Accuracy on consonant-affix and vowel-affix relations prior to training on wholes

Table 9 shows that consonants were more associable than vowels. This result is especially striking given the fact that the subjects received more instruction with vowels than they did with consonants: while there were 30 training trials involving consonant-affix relations, there were 44 involving vowel-affix relations (Table 2). The lower associability of vowels compared to consonants is in line with findings of Bonatti et al. (2005), who found that non-adjacent statistical dependencies between vowels are harder to learn than non-adjacent dependencies between consonants but not with Newport and Aslin (2004), who failed to find a difference.

In addition, there is no significant effect of syllabic position on consonant associability. It is not the case that codas are significantly more or less associable than onsets. Finally, subject groups do not differ significantly on how well they acquire vowel associations. This result indicates that the differences that are observed between subjects acquiring rime associations and those acquiring body associations are not simply due to how good the subjects assigned to those groups are at associative learning of the type tested in this experiment.

After Training on Wholes. The reason to examine how well subjects perform on segments after being trained on wholes is that it could be the case that training on wholes simply causes subjects to re-associate the segments with the response that is appropriate for the whole. In this section, we show that this hypothesis is ruled out because accuracy on segments does not fall below chance, indicating that segments do not get associated with the response that is appropriate for the rime.

After body-affix or rime-affix associations are introduced, generalization accuracy on vowels and consonants displays main effects of constituent ($F(1,66)=10.421$, rime vs. body, $p=.001$), and segment type ($F(1,66)=23.851$, consonant vs. vowel, $p<.0005$).

Table 10 shows how accurate subjects were on consonant and vowel associations after having been introduced to body or rime associations as a function of which constituent was introduced, where the affix was located relative to the root, what the segment was, and how much training was given. The table shows a main effect of constituent: subjects are somewhat more accurate on segment associations when they are exposed to body, rather than rime associations. This could be due to a competition between the rime and the segments it contains. Since the body never gets associated with a response, it does not

compete with the segments it contains. In addition, the main effect of segment type is present: subjects are more accurate on consonants than on vowels.

Association	**Local**		**Non-local**		**Local**		**Non-local**	
Associated	Coda	Vowel	Coda	Vowel	Onset	Vowel	Onset	Vowel
%correct	57	44	52	44	61	53	53	48
	48		46		49		55	

Table 10. Accuracy on consonant-affix and vowel-affix relations after body/rime training: effects of affix location, segment type, and constituent/consonant location.

Comparing Learning Rates

Table 11 displays accuracy as a function of amount of training on a unit type. The number of training trials is the number of training trials preceding Generalization 1 for consonants and vowels and Generalization 2 for rimes and bodies. The baseline is chance for consonants and vowels and accuracy prior to training on wholes for rimes and bodies. Table 11 shows that the speed of acquiring the association (change in accuracy by number of training trials) is much smaller for vowels than it is for consonants and much larger for rimes than for either consonants or vowels.

	Rime	**Body**	**Consonant**	**Vowel**
Training only	0.80	0.50	0.50	0.19
Training & feedback	0.66	0.41	0.32	0.12

Table 11. Learning rates for rimes, bodies, consonants and vowels (per trials of exposure). The top row of numbers use the number of training trials in the denominator while the bottom row uses the sum of training and feedback trials.

The Model

The Framework

Representing Syllable Structure. In the present framework, the syllable is represented as a matrix of resting activation levels, as shown in Table 12. Each cell in the table corresponds to a node or a link. The cells corresponding to nodes are the ones for which the row label and the column label are identical. The columns correspond to the nodes to which the links point while the rows are the nodes from which the links originate. Cells corresponding to links of strength equal to zero are left empty. Note that if a node has a resting activation level of zero, the strengths of links heading to and radiating from the node are assumed to have a strength of zero. Thus nodes that have a strength of zero simply don't exist (to the rest of the network). Since the body does not exist in the structure shown in Table 12, cells corresponding to links pointing to and from the body are merged and shaded, indicating that they must be equal to zero.

Table 12 shows a version of the traditional structure of the syllable in which the rime is a node while the body is not, and each part is connected only to the whole that immediately dominates it. Thus, for instance, the nucleus is not connected to the syllable because it is dominated by the rime. However, one should note that this representation is more specific than the traditional tree diagram in that we specify whether a given connection is excitatory or inhibitory as well as exact strengths of both nodes and connections. Furthermore, all links in the present framework are directed. Thus, for instance, in Table 12,

wholes inhibit parts while parts excite wholes, and the amount of inhibition is smaller than the amount of excitation.

	Onset	**Nucleus**	**Coda**	**Body**	**Rime**	**Syllable**
Onset	$R_O = 1.0$					$R_{O>S} = 0.3$
Nucleus		$R_N = 0.7$			$R_{N>R} = 0.3$	
Coda			$R_C = 1.0$		$R_{C>R} = 0.3$	
Body						
Rime		$R_{R>N} = -0.2$	$R_{R>C} = -0.2$		$R_R = 0.5$	$R_{R>S} = 0.3$
Syllable	$R_{S>O} = -0.2$				$R_{S>R} = -0.2$	$R_S = 1.0$

Table 12. An example of a syllable structure.

The syllable structure in Table 12 can also be represented as in Figure 4.

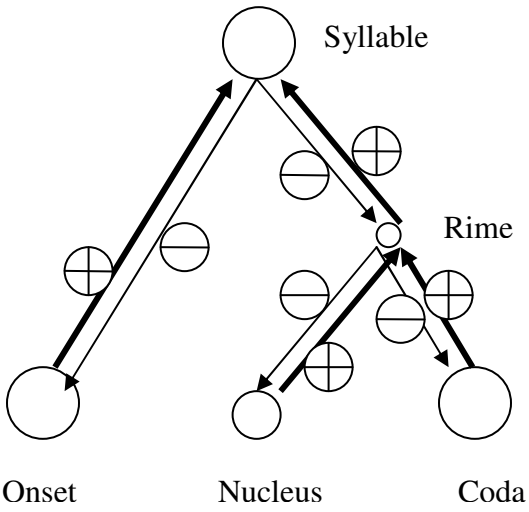


Figure 4. A graphical representation of the syllable structure in Table 12: The coda, the onset and the syllable are the most salient nodes, followed by nucleus, followed by rime. There is no body node. There are excitatory bottom-up connections and weaker inhibitory top-down connections.

The resting activation values displayed in the syllable matrix are used to determine how much activation will be received by memory traces representing each part of a syllable presented to the model. The amount of activation received by a chunk of a particular type is increased when 1) the resting activation level of the node representing the type is high, 2) there are many strong excitatory links pointing to the node, and 3) the inhibitory links pointing to the node are weak and few in number. In addition, there is a free parameter in the model, which determines whether a node loses (some) activation that spreads from it into the radiating links. If this parameter is positive, the amount of activation stored in the node after a syllable containing its referent is presented is reduced when the cumulative strength of links radiating from the node, i.e., the sum of absolute values of their activation levels, is high.

The resting activation levels are used to derive the proportion of input activation that is going to be stored in representations corresponding to each part of a syllable (and the syllable as a whole) after the syllable has been presented.

For each node, the formula in (1) is applied where X is the current node, $\neg X_1$ through $\neg X_{num\_earlier}$ are the nodes that represent chunks that end before X ends in the syllable or that end at the same time and are part of X, and k is the free parameter that determines how much activation that spreads out of the node is lost to the node. When the end of the syllable is reached, the formula in (2) is applied to each node where remaining nodes are the ones that come after X in the syllable or of which X is a part. The updating sequence and corresponding *num_earlier* and *remaining* numbers are shown in (3). Note that multiplying the final result by R_X ensures that nodes that don't exist do not end up getting any activation and thus are not associable.[5] This multiplication represents the likelihood of parsing the chunk out of the signal during testing. If the likelihood is zero, then how often the chunk has co-occurred with a given affix during training is not relevant.

(1) $$A_X = R_X - \frac{k * \sum_{i=1}^{5} | R_{X>\neg X_i} |}{\sum_{i=1}^{5} | R_{X>\neg X_i} | + R_X} + \sum_{i=0}^{num\_earlier} \frac{A_{\neg X_i} * R_{\neg X_i > X}}{\sum_{j=1}^{5} | R_{\neg X_i > \neg X_j} | + | R_{\neg X_i > X} |}$$

(2) $$A_X = R_X * (A_X + \sum_{i=0}^{remaining} \frac{A_{\neg X_i} * R_{\neg X_i > X}}{\sum_{j=1}^{5} | R_{\neg X_i > \neg X_j} | + | R_{\neg X_i > X} |})$$

(3) Formulas are applied in the following order:
onset → nucleus → body → coda → rime → syllable

where

num_earlier =	0	1	2	3	4	5
remaining =	5	4	3	2	1	0

Training. The model is presented with a list of syllables paired with affixes. For each onset, nucleus, coda, body, rime and syllable encountered, the model calculates the number of times it is encountered with each of the affixes, e.g., it might remember that the onset /g/ co-occurred 18 times with /num/ and 0 times with /min/ during stage 1.

After going through Stage I training, the model saves the co-occurrence statistics that it has gathered. Numbers of chunk-affix co-occurrences gathered during Stage II are simply added to the numbers gathered in Stage I. Thus, if the onset /g/ co-occurs 24 times with /mɪn/ and 10 times with /num/ during Stage II, the model will know that overall /g/ co-occurred with /num/ 28 times and with /mɪn/ 24 times.

Each time the model encounters a chunk-affix pair, it will increment the counter for that chunk-affix pair by the activation value A for that chunk. Thus, in our example, if A_{onset} is 0.9, the onset /g/'s /num/ **score** will be 18*0.9=16.2 following Stage I. Thus, activation level of a chunk depends on where it is in the syllable, and nodes with low activation levels are less associable than nodes with high activation levels where activation level is a product of the resting activation level of the node, how much activation

[5] When R_X is zero, the summed strength of all links radiating from it is set to 1 to avoid dividing by zero. The choice of 1 is arbitrary since the only number divided by the strength of links radiating from a non-existent node is zero thus the result of division is always zero.

and inhibition it receives from other nodes, and, if k > 0, how much activation it sends away to excite or inhibit other nodes.

There is a free parameter in the model that allows the model to assign different weights to each stage of training. This is done by dividing all scores from a given stage by the parameter after training on the immediately following stage. The parameter has no effect on performance on the test immediately following the training stage for which it is set. Thus, if the parameter is set to 2 during Stage I, all co-occurrence numbers from stage I will be divided by 2 before being added to co-occurrence numbers from Stage II. Thus, in our example above, on Test II, which follows Stage II, the model will think that /g/ co-occurred with /num/ 19 times, even though in reality it co-occurred with it 28 times because half of the co-occurrences that happened during Stage I will be forgotten or discounted. Thus, /g/'s /num/ score following stage II will be 18*0.9*0.5+10*0.9=17.1.

Thus, the overall score for a given chunk-affix pair is given by the formula in (4) where S is the score for a given chunk-affix pair, D_i is the decay parameter associated with stage I, which is always equal to 1 for the training stage immediately preceding the test, and $P(chunk, affix)_i$ is the number of times the chunk co-occurred with the affix during stage i.

$$(4)\quad S_{(chunk,affix)} = \sum_{i=1}^{\#\_of\_preceding\_stages} (P(chunk, affix)_i * A_{Chunk} * D_i)$$

Testing. During testing the model is presented with novel syllables that are not paired with an affix. For each syllable, the model extracts all the chunks that are present in it. It then recalls the scores representing how often each of the chunks present in the syllable has co-occurred with each of the suffixes during training and how salient these pairings have been overall, depending on the activation level of the chunk and the distribution of chunk-affix co-occurrences across stages of training. To predict which affix should go with the presented syllable, we sum the min-scores of all the chunks in the syllable and subtract the sum of num-scores for each chunk in the syllable from the result. If the result is positive, /num/ is the predicted affix, while /mɪn/ is predicted if the result is negative.

Thus, the prediction value for a given syllable is given in equation (5) where each syllable consists of six chunks: onset, nucleus, coda, body, rime, and syllable. One can think of the prediction value as the model's confidence in its choice or the likelihood that the model would go along with its choice when random noise is injected into its performance.

$$(5)\quad \Pr_{Syllable} = \sum_{i=1}^{6} (S_{(chunk_i,num)} - S_{(chunk_i,mIn)})$$

Comparing Models

Any model of syllable structure should explain several basic results. Namely, it should predict that: rimes are more associable than bodies (Table 5), rimes (and possibly bodies) are more associable than segments in terms of speed of acquiring association (Table 11), it is possible to associate a rime with suffix X while its parts are not associated with any suffix, being at chance (Table 5 and Table 10), the coda is somewhat less associable than onset but the difference in associability between coda and onset is much smaller than that between rime and body (Table 5 vs. Table 9), training on parts of a rime does not generalize to the rime as much as training on parts of the body generalizes to the body (Table 6), "body associations" do not show full generalization to novel syllables while rime associations do (Table 8), the onset-rime structure is not universal, hence cannot be explained by the fact that the rime follows the body

temporally (Lee, 2006), and either syllable nodes or rime nodes are necessary to account for the fact that priming with two shared segments is more effective than priming with three shared segments when the two segments form a syllable in the target word (Fernald et al., 1996; Carreiras & Perea, 2002).

The fact that there is cross-linguistic variability in syllable structure (Yoon & Derwing, 2001) and that frequency of between-segment co-occurrence is a crucial determinant of whether the nucleus is more closely tied to the coda or to the onset within a single language (Lee, 2006) indicate that the rime/body difference is not about which comes first temporally or its segmental shape (CV vs. VC). Therefore, we will not consider explanations for the asymmetry that propose that connections from the preceding to the following are stronger than connections going in the reverse direction. If our model is to capture variation in syllabic constituency, explanations must focus on consequences of constituency, not precedence.

This leaves the following space of possibilities: 1) parts of constituents may have a lower resting activation level than parts of non-constituents, 2) part→whole connections may be stronger within constituents than across constituent boundaries, 3) part-part connections may be stronger within constituents than across constituent boundaries, and 4) constituents may have higher resting activation level than non-constituents. However, option 1 is ruled out by the finding that a difference in associability between wholes does not entail a difference in associability between parts: while rimes are more associable than bodies regardless of whether local or non-local associations are at issue, codas are less associable than onsets only for local associations.

For each of the three remaining possibilities, the following questions should be addressed: a) are there excitatory part→whole connections, b) are there whole→part connections, c) are the whole→part connections inhibitory or excitatory, and d) is activation that spreads from a node into the radiating links lost to the source node.

The following possibilities will be left outside the scope of this paper: 1) there is more inhibition coming into a non-constituent than into a constituent, 2) there are more links radiating from a non-constituent than from a constituent, and 3) inhibitory connections leading from a constituent to its parts are stronger than those leading from a non-constituent to its parts. The reason these models will not be considered is that the author can think of no plausible psychological interpretations for them.

We will start out with a maximally simple model in which the only nodes that exist are segment nodes. Then we will make additions to this model to create asymmetric structures and examine the resulting differences in performance, i.e., prediction value, relative to the original model. Every pair of models compared will differ only in how the rime or connectivity within the rime are handled, so we can compare the models by looking at their performance on tests given to subjects trained on rime-affix associations. In this way, we will determine which modifications increase associability of the constituent and change the amount of generalization to novel syllables and the extent to which they rely on increases in part associability to increase associability of the whole.

The Distributed Baseline

We start out with a model that is unable to capture the rime/body asymmetry with the intent of examining ways of elaborating it to produce the observed results. For the purposes of comparing the models, we have eliminated the asymmetry in amount of training between vowels and consonants during stage I so that vowel and consonant scores after stage I are identical if vowels and consonants are equally associable under the model considered.

The syllable structure is shown in Figure 5. In this model, the only nodes are onset, nucleus, and coda. There are no nodes corresponding to larger structures. There are no links within the syllable.

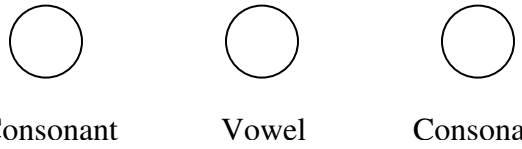


Figure 5. A network representation of a CVC syllable under the distributed baseline model.

Table 12 shows the results of training the model using the same stimuli presented to human subjects. The row labeled 'C' shows the model's accuracy on stimuli that share a consonant with the training stimuli, the consonant being associated with one of the affixes. The row labeled 'V' shows accuracy on stimuli that contain a vowel that has been paired with a particular affix during training.

Stimulus type	**Stage I**		**Stage II**
	Familiar syllables	**Novel syllables**	**Novel syllables**
C	18	18	5
V	18	18	7
CV or VC		-36	-12

Table 12. The simple distributed model's performance on testing trials following Stage I and II of training. Positive scores indicate that the association is correct while negative scores indicate that it is incorrect. The greater the absolute value of a score, the stronger the association. This is a version of the model with no discounting of Stage I training.

While the model does not achieve accurate performance on wholes without assuming that the training from Stage I is discounted, the score for a whole changes faster than the score for any individual part. How fast then does this purely distributed model predict the learning of whole-affix associations to be relative to the speed of learning part-affix associations?

The speed of learning associations of the whole (**V for 'velocity'**) is given by accuracy on the wholes after stage II minus accuracy on the wholes after stage I divided by the number of whole-affix pairings during stage II training, or

(6) $V_W = (\mathrm{Pr}_{W_2} - \mathrm{Pr}_{W_1}) / P_{w_2}$

The speed of learning associations of a consonant is given by accuracy on the part after stage I divided by the number of consonant-affix pairings during Stage I. Similarly, accuracy on vowels is accuracy on vowels after Stage I divided by number of vowel-affix pairings during Stage I.

(7) $V_C = \mathrm{Pr}_{C_1} / P_{C_1}$
(8) $V_V = \mathrm{Pr}_{V_1} / P_{V_1}$

In the purely distributed model the accuracy score for the whole is always equal to the sum of accuracy scores for the parts multiplied by –1. Thus,

(9) $\mathrm{Pr}_{W_i} = -(\mathrm{Pr}_{C_i} + \mathrm{Pr}_{V_i})$

Every time a whole is perceived both of the parts are perceived as well. Thus, whenever a whole-affix pairing is presented in stage II, the parts of the whole are paired with the affix with which they are

not associated during stage I. Thus to derive the accuracy on a part after stage II we need to add the number of times that part was paired with the same affix as in Stage I and subtract the number of times it was paired with a different affix, the latter being equal to the number of times the whole was presented. The equation for the consonants is shown in (10).

$$(10)\quad \mathrm{Pr}_{C_2} = \frac{\mathrm{Pr}_{C_1}}{d} + V_C * P_{C_2} - V_C * P_{W_2}$$

Now we can express the speed of acquiring whole associations as a function of the speed of acquiring part associations and number of times parts and wholes are presented during each stage of the experiment.[6]

$$(11)\quad V_W = (\mathrm{Pr}_{W_2} - \mathrm{Pr}_{W_1}) / P_{W_2}$$

$$\mathrm{Pr}_{W_1} = -(V_C * P_{C_1} + V_V * P_{V_1})$$

$$\mathrm{Pr}_{W_2} = -(V_C (\frac{P_{C_1}}{d} + P_{C_2} - P_{W_2}) + V_V (\frac{P_{V_1}}{d} + P_{V_2} - P_{W_2}))$$

Therefore,

$$(12) V_W = (V_C (P_{C_1} - \frac{P_{C_1}}{d} - P_{C_2} + P_{W_2}) + V_V (P_{V_1} - \frac{P_{V_1}}{d} - P_{V_2} + P_{W_2})) / P_{W_2}$$

If d=1, as in our current model,

$$(13)\quad V_W = (V_C (P_{W_2} - P_{C_2}) + V_V (P_{W_2} - P_{V_2})) / P_{W_2}$$

In the present experiment,

$$(14)\quad \begin{aligned} P_{C_2} &= 10 \\ P_{V_2} &= 12 \\ P_{W_2} &= 23 \end{aligned}$$

Therefore,

$$(15)\quad V_W = (13V_C + 11V_V) / 23 = 0.57V_C + 0.48V_V$$

Thus, under this simplest model, rimes and bodies are not predicted to be more associable in our experiment than both consonants and vowels unless

$$(16)\quad \begin{aligned} 0.43V_C &< 0.48V_V \\ 0.52V_V &< 0.57V_C \end{aligned}$$

[6] Pr = prediction, p = frequency of occurrence, v = learning rate.

That is,

(17) $0.9V_C < V_V < 1.1V_C$

This prediction contrasts with the experimental findings where rimes are more associable than single segments while V_c is far greater than V_v. The model is also unable to account for subjects' ability to learn to associate a rime with response X while its parts are not associated with response X. Finally, the model does not produce incomplete generalization, which we have observed with body-affix dependencies in the experiment.

We will now examine possible reasons for observing incomplete generalization, ways of making constituents more associable than their parts, and making constituents more associable than non-constituents of the same size.

Incomplete Generalization

There are two ways to produce incomplete generalization of constituent-affix associations: storage of partially overlapping constituents and storage of a larger unit, the syllable. Thus, we observe incomplete generalization of body-affix associations in the model shown in Figure 6 and in the one shown in Figure 7.

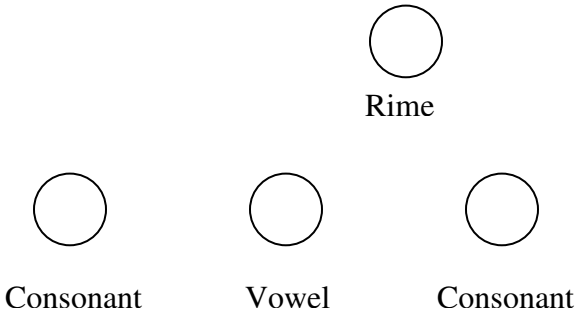


Figure 6. A model that produces incomplete generalization of body-affix associations due to the existence of a rime node.

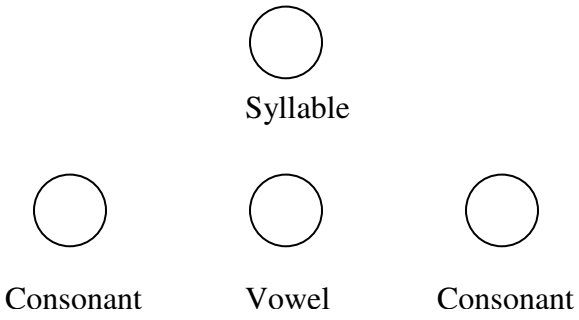


Figure 7. A model that produces incomplete generalization of body-affix associations due to the existence of a syllable node

Table 13 shows that scores on feedback trials, which consist of familiar syllables, increase relative to the distributed baseline while scores on generalization trials, which consist of novel syllables, stay the same when either rime nodes or syllable nodes are introduced.

	Distributed Baseline	Store Rimes	Store Syllables
Trained syllables	-12	-11	-9
Novel syllables	-12	-12	-12

Table 13. Storage of syllables and/or rimes produces incomplete generalization of body associations: scores on body stimuli after Stage II

Thus, both models produce incomplete generalization of body associations and, interestingly, incomplete generalization of body associations is consistent with the existence of rime nodes. The observed absence of incomplete generalization for rime associations is expected if body nodes do not exist. Alternatively, it may indicate that syllable associations are not formed if the subjects are performing the task relatively accurately since overall accuracy is higher after training on rimes.

Increasing Constituent Associability

Constituent associability can be increased by increasing the activation level that the constituent node has during training. This can be accomplished by either increasing the resting activation of the constituent node (Figure 8), by increasing connection strength between parts of a constituent and the constituent node (Figure 9) and by increasing connection strength between parts of a constituent if those nodes will send some of their activation to the constituent node (Figure 10). In all the models here, nodes do not lose the activation that spreads from them to other nodes, i.e., k=0.

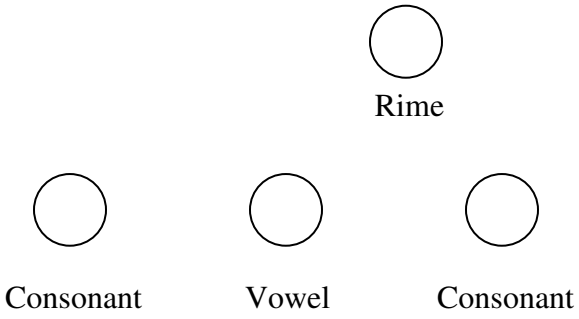


Figure 8. The parseability/salience model: the rime has a higher resting activation level than the body.

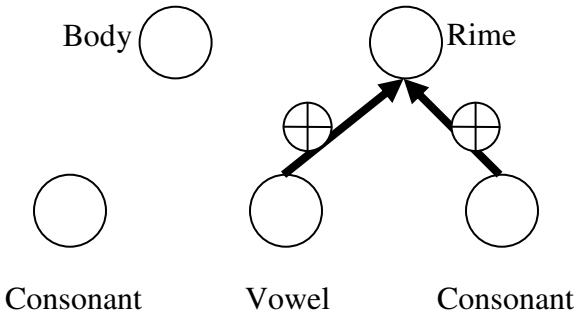


Figure 9. The part→whole connectivity model: segment→rime are stronger than segment→body connections where both body and rime exist.

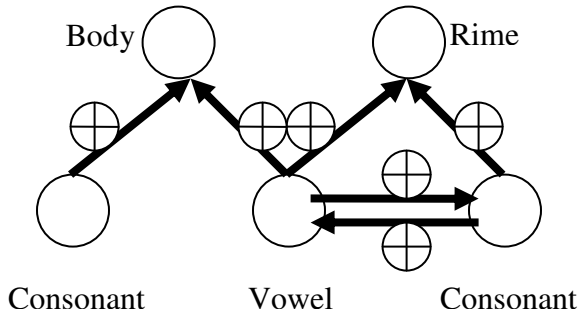


Figure 10. The part→part connectivity model: the onset←→nucleus connection is weaker than the nucleus←→coda connection where both body and rime exist, and both segment→body and segment→rime connections exist and are of equal strength.

It is important to note that the representation of the body in Figure 8 is the same as the representation of the rime in Figure 7, representation of the body in Figure 9 is the same as the representation of the rime in Figure 8, and representation of the body in Figure 10 is the same as the representation of the rime in Figure 9. Thus the way a model in figure number N-1 performs on rimes is equivalent to the way a model in figure number N performs on bodies.

Therefore, if we want to see whether one of the above models produces the rime-body asymmetry, we can simply compare its performance on rimes to the preceding model's performance on rimes. Table 14 shows performance of the distributed baseline model, the parseability/salience model, the part→whole connectivity model, and the part→part connectivity model on C, V, and rime stimuli after Stage I and after Stage II. Only generalization trials are shown.

	Distributed		**Parseability-based**		**Part→whole**		**Part→part**	
	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II	Stage I	Stage II
Coda	18	5	22	14	23	17	25	17
Nucleus	18	7	20	9	21	10	24	12
Rime	-36	-12	-36	11	-36	19	-41	16

Table 14. Accuracy scores for generalization trials for the distributed baseline model, the parseability/salience model, the part→whole connectivity model, and the part→part connectivity model.

Table 14 shows that associability of the whole rises if the resting activation level of the whole is increased (compare the parseabilty-based model to the distributed model). Therefore, a possible account for why rimes are more associable than bodies is that rimes are more salient, or are more likely to be parsed out of the acoustic signal, than bodies.

Stronger part→whole connections also make a whole more associable, as shown by the fact that the rime is more associable under the part→whole model than under the parseability-based model. Therefore, a possible account of the high associability of rimes relative to bodies is that the rime node receives more activation from the segment nodes than the body node does.

Finally, part-part connections also influence associability of the whole if the parts are connected to the whole and thus can spread activation to it. However, strengthening part-part connections actually decreases the predicted accuracy on wholes after Stage II, as shown by the comparison of the part→whole model to the part→part model. This prediction is incorrect as is the prediction that subjects should be less accurate on rimes (whose parts are strongly interconnected) than on bodies (whose part-part connections are weak) following Stage I. In actuality, accuracy on rimes is higher than accuracy on bodies following both Stage I and Stage II. This finding indicates that the difference in associability between bodies and rimes cannot be accounted for between-segment connection strength.

It is important to note that to make a whole more associable than either of its parts, the activation level of the whole need not exceed the activation level of the parts. Associability of the whole in our model when the whole is allocated a separate node is equal to

$$(18)\quad V_W = (13V_C + 11V_V + 23A_W)/23 = 0.57V_C + 0.48V_V + A_W$$

Thus, for V_w to be greater than V_c and V_v,

$$(19)\quad \begin{aligned} 0.57V_C + 0.48V_V + A_W &> V_C \\ 0.57V_C + 0.48V_V + A_W &> V_V \end{aligned}$$

Since V_C and V_V are approximately equal to, respectively, A_C and A_V if the generalization test on parts does not involve wholes that have been extensively presented during training, the relations in (20) need to hold for the whole to be more associable than the parts. These do in fact hold in our experiment.

$$(20)\quad \begin{aligned} A_w &> 0.43A_C - 0.48A_V \\ A_w &> 0.52A_v - 0.57A_C \end{aligned}$$

Interestingly, increasing associability of the whole also increases associability of the parts. Thus, predicted accuracy on consonants and vowels is higher under the parseability-based model than under the distributed model, and higher still under the part→whole model. This is because rimes presented during training on parts may be repeated during testing. This leads to the incorrect prediction that units that are part of a larger constituent should be more associable than parts that are not. If anything, the opposite is true in the experimental data: the coda is somewhat less associable than the onset if local associations are considered (coda-suffix associations are weaker than onset-prefix associations). In the next section, we consider ways of eliminating this problem by introducing top-down inhibition or loss of sent-off activation, which can be considered alternative ways of implementing between-level competition.

Decreasing Associability of Parts of Constituents

Figure 11 presents a way to modify the parseability-based model to account for the finding that parts of constituents are somewhat less associable than units that are not part of a constituent. The mechanism involves inhibitory connections from the whole to the part. If constituents have a higher resting activation level than non-constituents, the amount of inhibitory activation sent to the parts will be greater if the whole is a constituent than if it is not. An alternative way to reduce associability of constituent parts is to make parts lose the activation that they send off to the constituent node.

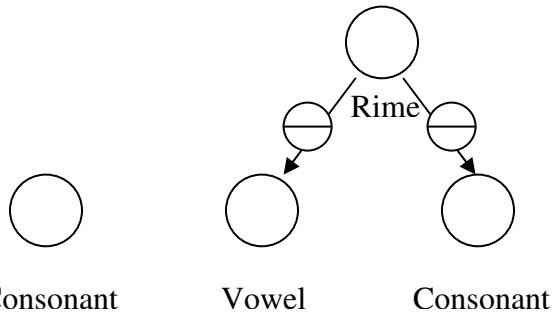


Figure 11. The parseability-based model with top-down inhibition.

Table 15 shows the predicted accuracy scores for the parseability-based model with and without top-down inhibition and the part→whole connectivity-based model with and without loss of sent-off activation. We do not consider the part→part connectivity-based model because the extra activation sent off by parts of a constituent is simply given to the other part of the constituent in the model. If the sent-off activation is lost to the sending node, there is no net increase in activation flowing into the constituent node relative to a model that does not have part-part connections, i.e., where the strength of the within-constituent part-part connections is zero.

	Parseability-based				**Part→whole**			
	with		without		with		without	
Stage	I	II	I	II	I	II	I	II
C	19	13	22	14	20	16	23	17
V	17	8	20	9	17	9	21	10
W	-31	13	-36	11	-30	19	-36	19

Table 15. Accuracy scores for generalization trials for the models with vs. without between-level competition

Table 15 shows that between-level competition effectively eliminates the problem of increased associability of parts following from greater associability of the whole. Models that include between-level competition and differences in activation level between constituents and non-constituents due to either resting activation level of the node representing the whole or the strength of part→whole connection correctly account for the experimental data. In these models, constituents are more associable than non-constituents but constituent parts are somewhat less associable than units that are not part of a larger constituent. An alternative possibility is to claim that the onset is inherently more associable than the coda due to temporal precedence. The two alternative hypotheses make different predictions regarding relative associability of onset and coda in syllables in which the body is a constituent while the rime is not.

General Discussion

Explanations for Differences in Constituency

A number of factors conspire to ensure that the rime is more cohesive than the body in the present stimuli, including 1) overall statistics of segment co-occurrence within English syllables, 2) statistics of segment co-occurrence within the particular syllables used in the present study, 3) potential lexicality, a.k.a., possible word status, and 4) actual lexicality. Thus, nuclei tend to co-occur with codas more than with onsets in English, and the rimes (/æʃ/ and /ʌg/) used in the present study are more frequent than the bodies (/ʃæ/ and /gʌ/).

In addition, /æʃ/ and /ʌg/ are well-formed words of English whereas /ʃæ/ and /gʌ/ are neither words nor well-formed in that they cannot stand on their own. It may be the case that for a difference in associability to be observed between a pair of segment strings, the segment strings must differ in potential or actual lexicality. For instance, Norris et al. (1997) and Cutler et al. (2001) argue that effects of syllable structure in monitoring tasks where subjects tend to detect syllables more easily than segment strings that cross syllable boundaries are due to the fact that syllables are possible words in the language and segment strings crossing constituent boundaries are often not. Thus, it may be the case that two segment strings that differ in linguistic constituency but do not differ in lexicality may not differ in associability either. Consequently, we would not be able to claim that such strings differ in parseability or salience and the difference in constituency would be more appropriately modeled by a difference in between-segment connection strength. Thus, it is important to manipulate lexicality, potential lexicality, and between-segment co-occurrence in future research.

Another possibility is that, while constituents are more parseable than non-constituents of the type investigated here, they are not parsed out in the course of normal language processing but can be parsed out if needed. A possible alternative to the present static models of constituent structure in which constituents are processing units (e.g., Mehler, 1981) is a dynamic distributed model that would dynamically create constituent nodes if they are needed for a learning task. Equipped with the assumption that a node is easier to create if the chunk it represents consists of two strongly connected parts than if it contains of weakly interconnected parts, such a model could potentially explain the results.

This model predicts that a constituent node will only be formed when it is needed for a configural learning task, i.e., a task that requires the subjects to associate the whole with a different response than either of its parts is associated with. Therefore, if subjects were exposed to a set of syllables sharing rimes paired with a particular affix, they would not be expected to form rime-affix associations, forming segment-affix associations instead. As a result, at least if syllable familiarity effect is not obtained, there should not be any rime familiarity effect. Thus, generalization to novel Cæʃ stimuli is no easier from a list of Cæʃ-mɪn training items than from a list of CVʃ-mɪn and CæC-mɪn training items that do not include Cæʃ-mɪn stimuli under this model.

Parseability/Salience vs. Part→Whole Connectivity

The difference in resting activation levels between rimes and bodies can have two manifestations: rimes should be more likely to be parsed out of the signal than bodies and, a body that has been parsed out of the signal should be less salient than a rime that has been parsed out.

The idea that a salient unit is more associable than a less salient unit has a long history in associative learning (e.g., Bush & Mosteller, 1951; Rescorla & Wagner, 1972). For instance, Rescorla and Wagner (1972) propose that associability of a stimulus is determined largely by how surprising it is (as well as how surprising its co-occurrence with the other stimulus is). It appears to be highly plausible that stimuli one attends to should be easier to associate than stimuli one does not attend to. While the high frequency of a stimulus makes its processing faster, it also makes its occurrence more predictable, thus reducing attention to the stimulus, making it less surprising.

However, saying that frequency reduces salience of a stimulus assumes that the stimulus is parsed out of the environment (cf. Kamin, 1969). One may be more likely to parse something out of the environment if its parts frequently co-occur, i.e., if it has high frequency (see Lee, 2006, for evidence regarding syllabic constituency). Because one can only associate something that one has perceived/parsed out, frequency is expected to also have positive effects on associability in the part of the frequency continuum that is below the part where it has negative effects.

The inverted U-shaped effect of predictability on processing speed and accuracy has been found in multiple linguistic domains. Several recent studies report U-shaped frequency effects on speed of word

recognition and production (Balota et al., 2004; Bien et al., 2005; Tabak et al., 2005) based on large-scale multiple regression analyses of existing collections of experimental data. Kapatsinski and Radicke (2007) find a U-shaped predictability effect in auditory particle recognition: *up* is detected more easily in medium-frequency verb-particle combinations than in low-frequency and high-frequency ones.

If constituency increases parseability, rimes are more likely to fall onto the part of the parseability continuum where increases in frequency are detrimental for associability than bodies are. Whether or not the model actually predicts a negative correlation between frequency and associability for rimes depends on whether the rimes are expected to reach a ceiling on the parseability continuum at a certain non-maximum point on the frequency continuum.

By contrast, the model that assumes that constituency influences part→whole connectivity predicts that part-whole co-occurrence should have a monotonic positive correlation with associability, under the standard Hebbian assumption that high frequency of co-occurrence between parts and wholes strengthens connections.

A Specificity-based Alternative

All of the models discussed above assume that the differences between constituents and non-constituents stem from how easy it is to extract constituents vs. non-constituents during an encounter with a linguistic signal that contains it. A major alternative explanation is that the differences in associability come from how easy it is to detect that two words share the same rime as opposed to how easy it is to detect that they share the same body.

If listeners do not automatically categorize incoming speech into phonemic categories automatically (e.g., Port & Leary, 2005), and variations in the coda have a greater impact on vowel quality than variations in the onset, the equivalence of different tokens of the same rime may be easier to detect than the equivalence of different tokens of the same body (cf. Geudens et al., 2005). As a result, acquiring rime associations would be easier than acquiring body associations. Thus, if phonemic perception is not assumed, the results seem consistent with a full-listing approach in which there are no differences between bodies and rimes, except for how acoustically similar the tokens of a given chunk type are to each other. The specificity-based approach may be able to account for the results of Lee (2006) as well. Given that there are more codas than onsets in Korean, a given rime is more acoustically variable than a given body, leading Koreans to treat the body as a constituent. Assuming that categorization is easier when the category has many exemplars, one would also predict that high-frequency units should be more associable.

In addition, generalization of rime associations to novel syllables would be easier than generalization of body associations simply because the generalization stimuli would be (subphonemically) more similar to the training stimuli in the rime condition relative to the body condition (cf. Cutler et al., 2001). However, this hypothesis necessarily predicts incomplete generalization in both conditions, while we have observed full generalization of rime dependencies and partial generalization of body dependencies.

In order to obtain this pattern of results, one would have to propose that the phonemic-level category to which new tokens of a body are compared is influenced by the tokens of that body previously encountered in the experiment more than the phonemic-level category to which new tokens of a rime are compared is. That is, the representation of the rime on the phonemic level would be more stable and less dependent on recent experience.

Categories are more stable when they have a large number of members. Thus, a prediction this model makes is that complete generalization is more likely for dependencies involving high-frequency units. Rimes are more frequent than bodies in the present stimuli. However, if the difference in

generalization is found when frequency is controlled, the categorization-based model would have to adopt the traditional model's assumption that, at least at the phonemic level, rimes are more likely to be parsed out of the signal than bodies are. In that case, the only difference between this model and the traditional model would be in what is stored at the subphonemic level with essentially the same explanation for constituency effects.

The Garner interference paradigm (Garner, 1974; Garner & Felfoldy, 1970) provides a possible further source of evidence on this hypothesis. In this paradigm, the subject is asked to classify stimuli along one dimension while some other dimension is either held constant, varies randomly, or is correlated with the attended dimension. If the subject cannot separate the dimensions, i.e., cannot attend to the dimension that is relevant for categorization without attending to the other, random variation on the 'unattended' dimension may make classification along the relevant dimension harder than if all stimuli have the same value on the 'unattended' dimension.

In our case, the subphonemic perception hypothesis predicts that since the vowel is supposed to be influenced by the coda more than by the onset, random variation in the coda should make classifying stimuli based on the vowel harder than random variation in the onset. Furthermore, given that onsets and codas do not differ significantly in associability, ease of vowel classification should account for most of the variation in accuracy and reaction time that the rime-body distinction accounts for.

Consonants vs. Vowels

Vowels were found to be less associable than consonants. These results are consistent with previous findings of Bonatti et al. (2005) who found that French speakers can learn statistical dependencies between non-adjacent consonants better than they can learn dependencies between non-adjacent vowels. However, it is not clear what is responsible for these results.

One possibility is that there is more variation in vowels than in consonants (Creel et al., 2006). This could make vowel associations harder to acquire since the difference between the speaker's and the listener's vowel categories may be greater than the difference between their consonant categories. In addition, the speaker's vowel productions may vary more than his consonant productions. One possible way to control for this factor is to have the stimuli synthesized anew for each listener recalibrating the synthesizer to match the listener's vowels.

Another possibility is that consonants are more associable because there are more irrelevant consonant types than vowel types: the consonants occurring in training are: /b/, /d/, g/, /θ/, /ð/, /s/, /z/, /f/, /v/, /ʃ/, /ʒ/, /dʒ/, and /tʃ/. By contrast, the only vowels that occur are /i/, /ɪ/, /o/, /u/, /æ/, and /ʌ/. As a result, the subjects might have a harder time noticing which vowels are predictive (/æ/ and /ʌ/) since all vowels are relatively frequent in the experiment. Conversely, given that a vowel occurs in a greater variety of contexts that a consonant does, it may be harder to notice that all tokens of the vowel belong to the same phoneme. Equalizing the numbers of consonants and vowels used as distractors in the experiment may eliminate the consonant-vowel asymmetry. However, Creel et al. (2006) investigated this issue by teaching native English speakers artificial lexicons consisting of VCVC or CVCV words and found that words sharing consonants were more confusable for English speakers than words sharing vowels even if the number of vowels in the artificial language exceeded the number of consonants.

Related Methods

XOR learning is a subtype of configural learning. That is, it is on e of multiple paradigms in which associations of the whole are not predictable from associations of the parts. Other configural learning paradigms involve the classical XOR distribution outlined in the introduction as well as negative patterning and biconditional discrimination.

In negative patterning, subjects need to learn that presentation of either stimulus A or stimulus B on its own is followed by a reward while the presentation of A and B together is not (Woodbury, 1943), as shown in Figure 12. The argument is that if subjects succeed at the task, they must be treating AB as not just A and B next to each other. In other words, there is a node corresponding to AB (Wagner, 1971; Pearce, 1987, 1994).

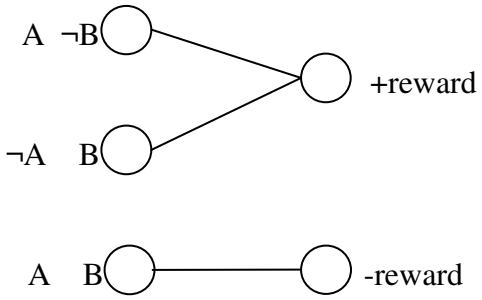


Figure 12. Hypothetical representation of stimulus associations in a negative patterning task: when A and B are presented together, +reward and -reward are activated equally strongly, cancelling each other.

In biconditional discrimination, introduced by Saavedra (1975) and illustrated in Figure 13, four stimuli (A, B, C, and D) are arranged into four compounds (AB, CD, AC, BD), two of which are associated with response X and two with response Y. Note that the compounds are arranged in such a way that any individual cue (A, B, C, and D) is just as likely to be paired with X as with Y. The discrimination is easy to encode if AB, CD, AC, and BC have a node corresponding to each of them as shown in Figure 5.

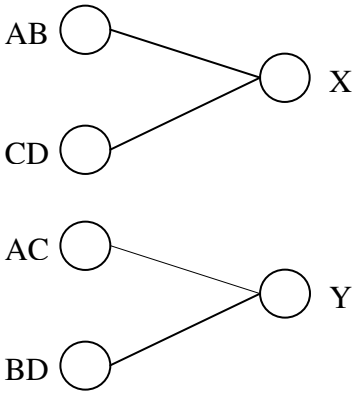


Figure 13. Hypothetical representation of stimulus associations in biconditional discrimination: AB and CD are associated with response X while AC and BD are associated with response Y.

The advantage of biconditional discrimination compared to the present task is that in the present paradigm there needs to be a large amount of variability in each position within the stimulus (e.g., the vowel /æ/ needs to be paired with a wide variety of codas to ensure that subjects learn /æ/-affix rather than rime-affix associations). It is sometimes impossible to create sufficient variability. For instance, if one wanted to examine the nature of constituency in article-noun strings in English, there is only a limited number of articles that could be paired with a noun. In biconditional discrimination, the parts of a whole whose existence is being tested need not be associated with anything. Therefore, one could compare, for instance, the learning of 'the cat-Y, a dog-Y, a cat - X, the dog -X' to the learning of 'cat the-Y, dog a-Y,

cat a - X, dog the - X' by exposing subjects to sentences like 'He gave the cat a blanket'. The disadvantage of biconditional discrimination is that it does not provide a way to examine associability of wholes and associability of parts within a single experiment.

The issue of psychological reality of complex units can also be addressed through typological research by comparing the frequencies of patterns whose acquisition requires configural learning to distributions that can be acquired elementally. If two units are likely to be chunked together, then learning distributions that require such chunking should not be much harder than learning distributions that do not require the units to be chunked. For instance, Pertsova (2007) has looked at syncretism distributions in personal pronouns across a sample of world languages, focusing specifically on cases in which the first and second person, singular and plural are represented by only two pronouns. Given this restriction, only the arrangements in Table 18 are possible. Pertsova found that distributions that require associating each pronoun with a combination of a person feature and a number feature ("person-number pronouns" in the table) were less frequent than distributions in which each pronoun could be associated with either just a person feature or just a number feature. In the present framework, these results suggest that person and number features are unlikely to be chunked together into a single complex feature unit since if they were, configural learning would be just as easy as elemental learning.

	Number Pronouns		**Person Pronouns**		**Person-number Pronouns**	
	Singular	Plural	Singular	Plural	Singular	Plural
1st Person	X	Y	X	X	X	Y
2nd Person	X	Y	Y	Y	Y	X

Table 16. Possible syncretism patterns in the domain of 1st and 2nd person personal pronouns, given that there are only two pronouns (X and Y) in the domain. Based on Pertsova (2007).

There are other ways to approach the issue in addition to configural learning methods. Wilson (2006) describes what he calls the "poverty of the stimulus method" (PSM), which involves exposing learners to input that is consistent with more than one possible generalization. By looking at which generalization subjects choose, one can determine which one is more natural.

It appears possible to apply this method to the study of constituency in the following way. If rimes are units and bodies are not, one can expect that given a choice between an onset-affix association and a body-affix one, the subjects would make the onset-affix association, while given the choice between a coda-affix association and a rime-affix association, the subjects would choose the rime-affix association (cf. Shanks et al., 1998, for a non-linguistic example). Such an experiment would also be helpful to test for possible biases in favor of generalizations made on the basis of larger units.

There are at least two possible reasons for such a bias: 1) generalizations based on larger units are more likely to be appropriately constrained (e.g., Williams et al., 1994) and 2) syllabic constituents may be easier to extract from syllables than segments because syllables are more similar to syllabic constituents than to segments (McNeill & Lindig, 1973). For instance, McNeill and Lindig showed that when subjects are presented with a sequence of words, it is easier for them to detect syllables than segments in the words. However, given that listeners usually do not hear subsyllabic units in isolation, syllabic constituents would be expected to be more detectable during normal speech perception than segmental units in general, making distributed models of syllabic constituency doubtful.

In order to test this hypothesis, all training tokens might involve the same rime or body while testing tokens would include novel syllables with either novel or familiar rimes or bodies. If subjects make a rime-based association, they should display imperfect generalization to novel rimes. By contrast, perfect generalization to novel bodies might be expected.

Another potential way to test associability is to expose subjects to training in which the presence of a C and a V as either the rime or the body is associated with a certain outcome, e.g., both *ka-* and *-ak* would be associated with mɪn. Then one can test generalization to novel syllables containing either *ka-* or *-ak*. We expect more generalization to syllables with the known rime than to those with a known body.

Another prediction of constituency is that generalization should be more likely when the generalization stimuli share the relevant constituent with the training stimuli. That is, VC rime associations should be hard to transfer to testing items that contain the same VC sequence that is not the rime in the testing item. Thus, there should be incomplete generalization of -æʃ-mɪn associations to CæʃV testing items.

Finally, a possible way to examine differences between tree-based and dependency-based models of constituency is to look for pairs of stimuli that have the same structure under one view and different structures under the other view. If such a pair is found, one can look for the presence of structural priming between the stimuli. If structural priming is observed, the view of constituency that assigns the same structure to the two stimuli is supported (see Snider, 2007, for an example from syntax).

Conclusion

This paper introduces a domain-general method for implementing and testing models of constituency. We have applied this method to a test case in which multiple factors conspire to make constituency particularly clear and uncontroversial, the case of English syllables containing a lax vowel. In this particular case, the constituency difference corresponds to a difference in associability. That is, a constituent is easier to associate with an affix than a non-constituent of the same length regardless of whether the affix is a suffix or a prefix. We have also shown that the effect is not explained by similarity relations between syllables because subjects were found to be as accurate with novel syllables as with familiar ones. Finally, subjects assigned to different groups were found to be equally good learners and there were no significant differences in associability between the parts that were not shared between a constituent and a non-constituent. We have examined a number of models of constituency to determine which of them could account for the findings. Under the assumption of phonemic speech perception the data were best accounted for by localist models that assume an activation level difference between constituents and non-constituents, implemented either as a difference in parseability, or in part→whole connection strength. Thus, rimes of English syllables with lax vowels are more salient than bodies to English speakers. It is an open question whether all proposed linguistic constituents are more associable than comparable non-constituents and whether all types of constituency are profitably modeled by a localist, tree-structural representation. XOR learning, and configural learning more generally, provides a method to address this question.

References

Alvarez, C.J., Carreiras, M., & Perea, M. (2004). Are syllables phonological units in visual word recognition? *Language and Cognitive Processes*, *19*, 427-52.

Anderson, J.M., & Ewen, C.J. (1987). *Principles of Dependency Phonology*. Cambridge, UK: Cambridge University Press.

Ashby, J., & Rayner, K. (2004). Representing syllable information during silent reading: Evidence from eye movements. *Language and Cognitive Processes*, *19*, 391-426.

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H., & Yap, M.J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283-316.

Bendrien, T.A. (1992). Sound similarity judgments in English CVC's. BA Honors Thesis: U of Alberta.

Benki, J. (2003). Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *Journal of the Acoustical Society of America*, *113*, 1689-1705.

Bien, H., Levelt, W.M.J., & Baayen, R.H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences*, *102*, 17876-81.

Bonatti, L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*, 451-9.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America*, *84*, 101-14.

Bush, R.R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313–23.

Bybee, J., & Brewer, M.A. (1980). Explanation in morphophonemics: Changes in Provençal and Spanish preterite forms. *Lingua*, *52*, 201-42.

Carreiras, M., & Perea, M. (2002). Masked priming effects with syllabic neighbors in a lexical decision task. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 1228-42.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

Creel, S.C., Aslin, R.N., & Tanenhaus, M.K. (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language*, *54*, 1-19.

Cutler, A., McQueen, J.M., Norris, D., & Somejuan, A. (2001). The roll of the silly ball. In E. Dupoux (ed.), *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler* (pp.181-94). Cambridge, MA: MIT Press.

Davis, S. (1989). On a non-argument for the rhyme. *Journal of Linguistics*, *25,* 211-5.

Derwing, B.L. (1987). A cross-linguistic experimental investigation of syllable structure. Part I: Background and methodology. *Proceedings of the Third Annual Meeting of the Pacific Linguistics Conference*, 93-102.

Derwing, B.L., & Nearey, T.M. (1986). Experimental phonology at the University of Alberta. In J. Ohala (ed.), *Experimental phonology* (pp.187-209). San Diego: Academic Press.

Felty, R. 2007. Context effects in spoken word recognition of English and German by native and non-native listeners, Unpublished Doctoral Dissertation, University of Michigan.

Ferrand, L., Segui, J., & Grainger, J. (1996). Masked priming of words and picture naming: The role of syllabic units. *Journal of Memory and Language*, *35*, 708-23.

Fudge, E. (1987). Branching structure within the syllable. *Journal of Linguistics, 23*, 359-377.

Garner, W.R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Garner, W.R., & Felfoldy, G.L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology¸1*, 225-41.

Geudens, A., Sandra, D., & Martensen, H. (2005). Rhyming words and onset-rime constituents: An inquiry into structural breaking points and emergent boundaries in the syllable. *Journal of Experimental Child Psychology*, *92*, 366-87.

Hall, G. (2003). Learned changes in the sensitivity of stimulus representations: Associative and nonassociative mechanisms. *Quarterly Journal of Experimental Psychology*, *56B*, 43-55.

Hockema, S.A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development*, *2*, 119-46.

Jaeger, J.J. (2005). *Kids' slips: What young children's slips of the tongue reveal about language development*. Mahwah, NJ: Erlbaum.

Kamin, L.J. (1969). Predictability, surprise, attention and conditioning. In B.A. Campbell & R.M. Church (eds.), *Punishment and aversive behavior*. (pp. 279-296). New York: Appleton-Century-Crofts.

Kapatsinski, V.M. (2006). Frequency and cohesion: Evidence from repair. Paper presented at High Desert Linguistics Society VII. Albuquerque, NM.

Kapatsinski, V.M., & Radicke, J. (2007). Frequency and the emergence of prefabs: Evidence from monitoring. This volume.

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language*, *37*, 295-311.
Lee, Y. (2006). Sub-syllabic constituency in Korean and English, Unpublished Doctoral Dissertation, Northwestern University.
McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior*, *12*, 419-30.
Mehler, J. (1981). The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *295*, 333-52.
Minsky, M.L., & Papert, S.A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
Moder, C.L. (1992). Productivity and categorization in morphological classes, Unpublished Doctoral Dissertation, SUNY Buffalo.
Nearey, T. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*, 347-73.
Nearey, T. (2003). On the factorability of phonological units in speech perception. In R. Ogden, J. Local, & R. Temple (eds.), *Laboratory Phonology 6*. (pp. 197-221). Cambridge, UK: Cambridge University Press.
Nelson, D.L., & Nelson, L.D. (1970). Rated acoustic (articulatory) similarity for word pairs varying in number and ordinal position of common letters. *Psychonomic Science*, *19*, 81-2.
Newport, E.L., & Aslin, R.N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127-62.
Norris, D., McQueen, J.M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*, 191-243.
Nusbaum, H.G., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10* (pp. 357-76). Bloomington, IN: Speech Research Laboratory, Indiana University.
Pearce, J.M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*, 61-73.
Pearce, J.M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587-607.
Pertsova, K. (2007). Towards learning form-meaning correspondences of inflectional morphemes. Paper presented at LSA Annual Meeting, Anaheim, CA.
Port, R.F., & Leary, A.P. (2005). Against formal phonology. *Language*, *81*, 927-64.
Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (eds.), *Classical conditioning II: Current research and theory* (pp.64-99). New York: Appleton Century Crofts.
Saavedra, M.A. (1975). Pavlovian conditioning in the rabbit. *Learning and Motivation*, *6*, 314-26.
Selkirk, E.O. (1982). The syllable. In H. Van der Hulst & N. Smith (eds.), *The structure of phonological representations (part II)* (pp.337-83). Dordrecht: Foris.
Skousen, R. (1989). *Analogical modeling of language*. Dordrecht: Kluwer.
Snider, N. (2007). Evidence from priming for hierarchical representation in syntactic structure. Paper presented at LSA Annual Meeting, Anaheim, CA. Available online at http://www.stanford.edu/~snider/pubs/snider,lsa,2007.pdf.
Stemberger, J. (1983). *Speech errors and theoretical phonology: A review*. Bloomington: Indiana University Linguistics Club.
Tabak, W., Schreuder, R., & Baayen, R.H. (2005). Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. In M. Reis & S. Kepser (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp.529-55). Berlin: Mouton de Gruyter.
Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. C*ognition, 15*, 49-74.
Treiman, R. (1986). The division between onsets and rimes in English syllables. *Journal of Memory and Language*, *25*, 476-491.

Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 145-52.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M.B. Broe & J.B. Pierrehumbert (eds.), *Laboratory Phonology V: Acquisition and the Lexicon* (pp.269-82). Cambridge, UK: Cambridge University Press.

Vennemann, T. (1988). The rule dependence of syllable structure. In C. Duncan-Rose & T. Vennemann (eds.), *On language: Rhetorica, phonologica, syntactica: A festschrift for Robert P. Stockwell from his friends and colleagues* (pp.257-83). London: Routledge.

Vitz, P.C., & Winkler, B.S. (1973). Predicting the judged "similarity of sound" of English words. *Journal of Verbal Learning and Verbal Behavior*, *12*, 373-88.

Wagner, A.R. (1971). Elementary associations. In H.H. Kendler & J.T. Spence (eds.), *Essays in Neobehaviorism: A memorial volume to Kenneth W. Spence* (pp.187-213). New York: Appleton-Century-Crofts.

Williams, D.A., Sagness, K.E., & McPhee, J.E.. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 694-709.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*, 945-82.

Woodbury, C.B. (1943). The learning of stimulus patterns by dogs. *Journal of Comparative Psychology*, *35*, 29-40.

Yoon, Y.B., & Derwing, B.L. (2001). A language without a rhyme: Syllable structure experiments in Korean. *Canadian Journal of Linguistics*, *46*, 187-237.

Appendix: Instructions

Learning Toy Languages
Instructions

There are three parts to this experiment.

Within each part, you will first listen to words paired with either 'min' or 'noom'. Whether you hear 'min' or 'noom' depends on what the other word sounds like. This will last for about three minutes.

Then you will be presented with some more word pairs in which the min/noom part has been replaced by noise. You will be asked to guess whether the noise replaces 'min' or 'noom' by pressing a button on the button box. Once you have made your guess, the correct answer will be revealed to you. PLEASE, TRY TO RESPOND AS ACCURATELY AS POSSIBLE. This will last for about five minutes.

Finally, you will be presented with some more words paired with noise. You will again be asked to guess whether the noise replaces 'min' or 'noom. This time, though, you will not get feedback. PLEASE, TRY TO RESPOND AS ACCURATELY AS POSSIBLE. This will take about four minutes.

This entire three-part sequence will be repeated three times. The whole experiment will last for about 40 minutes.

The following messages appeared on the screen:

Prior to each training session (each subject saw either 'preceding' or 'following'):

Whether you hear 'min' or 'noom' depends on what the preceding/following word is like.
Press any button to begin

Prior to each feedback session:

You will now be given feedback.
Once you make your guess the correct answer is going to be pronounced.

Prior to each testing session:

You will now be tested on what you know
Press any button to begin

End of experiment:

The experiment is over.
Thank you for your participation.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Cross-modal Repetition Priming in Spoken Word Recognition[1]

Adam Buchwald[2], Stephen J. Winters[3], and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by grant number DC00012. The authors would also like to thank Melissa Troyer for her assistance running participants for this study. We also acknowledge Tessa Bent and Susannah Levi for helpful comments on earlier versions of this paper.
[2] Currently at New York University, Department of Speech-Language Pathology and Audiology.
[3] Currently at University of Calgary, Department of Linguistics.

Cross-modal Repetition Priming in Spoken Word Recognition

Abstract. Multimodal speech perception has become a topic of considerable interest to speech researchers. Previous research has demonstrated that perceivers use information from the visual modality to inform the process of spoken word recognition. In this paper, we used a cross-modal repetition priming paradigm to explore questions about multimodal speech perception. First, we report that participants identified spoken words mixed with noise more accurately when the words were preceded by a dynamic video clip of the word being produced than when the words were preceded by a static image. Second, analyses of the responses indicate that both correct and incorrect responses are constrained by dynamic visual information. These complementary results indicate that perceivers integrate speech information from two different sensory modalities even when the signals are presented asynchronously. Third, we addressed the nature of multimodal integration, and found that the cross-modal repetition priming was maintained even when visual and auditory signals come from different sources. We discuss implications of these results for theories of multimodal speech perception.

Introduction

Multimodal speech perception and the cognitive processes by which individuals integrate auditory and visual speech information with linguistic knowledge have become major areas of research in the field of speech perception (Bernstein, 2005; Calvert, Spence, & Stein, 2004; Kim, Davis, & Krins, 2004; Massaro & Cohen, 1995; Massaro & Stork, 1998; Massaro, 1998; Rosenblum, 2005). As Sumby and Pollack (1954) reported more than 50 years ago, normal-hearing listeners reliably make use of information from the visual speech signal to increase intelligibility of auditory speech over a wide range of signal-to-noise ratios. In addition, McGurk and MacDonald (1976) reported a perceptual illusion in which incongruent information from visual (e.g., [ba]) and auditory (e.g., [ga]) speech signals led to misperceptions (e.g., perceived: [da]) of the speech sounds. More recently, it has been found that presenting visual speech prior to auditory speech facilitates processing of the latter signal in a lexical decision task (Kim et al., 2004). These phenomena clearly suggest that perceptual integration – or binding – of the information from these two modalities is an integral part of speech perception, and that characterizing the nature of multimodal representations of speech is critical to a full understanding of speech perception (Summerfield, 1979).

This paper contributes to our understanding of multimodal speech perception by exploring the conditions under which information from auditory and visual signals is integrated. First, we provide critical evidence indicating that neither temporal synchrony of the auditory and visual signals nor identity in the source of the two signals is a necessary condition for this type of audiovisual integration to be observed, and that binding of these sources of information in the processes involved in spoken word recognition can occur in the absence of both of these conditions. In particular, we employed a repetition priming task in which we found that visual-only speech signals facilitated open-set recognition of subsequent noise-degraded audio-only speech signals, and that this effect persists even when the visual and auditory signals are clearly produced by different speakers (i.e., when there is a talker gender mismatch). Second, detailed analyses of the participants' responses reveal critical differences in responses to the auditory signals when participants first see a dynamic video clip compared to when they first see a static visual image, even when the signals cannot be reliably identified (i.e., when the open-set identification is incorrect). We show that these differences in responses are under stimulus control; that

is, the additional visual speech information constrains the responses in a manner consistent with the phonetic information present in the signal.

Audiovisual Integration in Speech Perception

The topic of multimodal speech perception and audiovisual integration have received attention from researchers addressing a wide variety of problems including second language acquisition (Davis & Kim, 2001; Kim & Davis, 2003; Davis & Kim, 2004), neurological processes and impairment (Skipper, Nusbaum, & Small, 2005; Hamilton, Shenton, & Coslett, 2006), speech production (Yehia, Rubin, & Vatikiotis-Bateson, 1998) and voice identity (Lachs, 2002; Kamichi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a, 2004b) as well as issues directly related to spoken word recognition (Dodd, Oerlemens, & Robinson, 1989; Kim et al., 2004; Mattys, Bernstein, & Auer Jr., 2002). These studies – combined with the pioneering work of Sumby and Pollack (1954) – all reveal that audiovisual integration is a fundamental part of speech perception which is seen in a variety of tasks and under a variety of conditions. In this paper, we focus on two possible conditions that are expected to promote audiovisual integration: temporal synchrony and source identity of the auditory and visual signals.

Previous research has shown effects of asynchronously presented visual speech on tasks involving auditory speech perception. Dodd, Oerlemans, and Robinson (1989) observed lexical repetition priming effects with visual-only primes and auditory-only targets. Using a semantic categorization task, Dodd et al. reported facilitation when participants were presented with visual-only primes of a speaker saying a word followed by an auditory target compared to a condition with no prime. This finding suggests that the visual prime and the auditory target activate common semantic representations in memory. It is worth noting that Dodd et al. used visual speech stimuli which were readily identified on their own (by at least 80% of participants in a screening task); thus, it remains possible that the facilitation in the semantic categorization task arises from separate, accurate identification of the stimuli from each modality rather than from audiovisual integration.

More recently, Kim et al. (2004) had participants perform a lexical decision task on spoken words that were preceded by visual-only speech signals. Participants' reaction times in lexical decision on trials with a consistent visual speech prime were compared to trials with inconsistent visual speech primes; they reported facilitation (i.e., faster reaction times) in the responses for trials with consistent visual speech primes, and concluded that speech perception is amodal because the priming effect suggests that visual and auditory signals activate common representations. However, it remains possible that the difference between the baseline and experimental conditions in Kim et al.'s study was due to response inhibition in the presence of inconsistent stimulus information as opposed to response facilitation in the presence of consistent information. Nevertheless, both of these explanations suggest that there is some type of common representation activated by both the visual and auditory signals which affects processes used to perform the lexical decision task. Moreover and critical to the present investigation, Kim et al. presented the two stimuli asynchronously.

Thus, these two lines of evidence suggest that speech information from auditory and visual modalities need not be presented synchronously to observe effects of perceptual binding in multimodal speech perception (see van Wassenhove, Grant, & Poeppel, 2006 for a recent discussion of temporal synchrony in audiovisual integration).

There also exists evidence that source identity is not required for audiovisual integration. Green, Kuhl, Meltzoff and Stevens (1991) reported a study in which participants readily perceived the McGurk illusion discussed above even when there was a gender mismatch between the face producing the visual

speech signal and the voice producing the auditory speech signal. Green et al. suggest that this finding supports the hypothesis that audiovisual integration occurs over abstract representations of speech and not over the detailed signals present in the environment. This unintuitive result contradicts any view of speech perception that does not allow for cognitive operations over abstract representations of speech, a strong view which has sometimes been attributed to event-based perception in general (Gibson, 1966), and Direct Realism in particular (Fowler, 1986).

This section has highlighted two findings regarding audiovisual integration in multimodal speech perception: 1) audiovisual integration has been observed in the absence of temporal synchrony (as in Kim et al., 2004, with both auditory and visual signals coming from the same speech event) and 2) audiovisual integration has been observed in the absence of source identity (but with temporal synchrony, as in Green et al., 1991). One goal of the present investigation is to determine whether audiovisual integration is observed in the absence of both temporal synchrony and source identity; that is, do observers integrate auditory and visual speech signals that are both separated in time and clearly come from different speech events in the world?

The other main component of the present investigation is to build on these previous works by investigating the nature of the audiovisual integration. To achieve this, we examined differences in correct and incorrect responses due to asynchronously presented visual speech in open-set spoken word identification.

Experiment 1

Experiment 1 sought to replicate the cross-modal priming in speech perception findings of Kim et al. (2004) using a task that allows us to explore the nature of the priming effect more directly. We employed a spoken word recognition task with auditory targets preceded by either static primes or dynamic video clips of the same speaker producing the same word. One critical goal of this experiment – beyond that reported by Kim et al. (2004) – was to try to gain a deeper understanding of the nature of audiovisual integration in a cross-modal task. We investigated not only whether the dynamic video clip prime would increase overall spoken word recognition accuracy, but also whether the responses on trials with dynamic video clips primes are more constrained than on those with static primes, and whether the nature of these constraints is predictable by (and can shed light on) the nature of visual and multimodal speech perception.

Participants

Forty Indiana University undergraduate students, ages 18-23, participated in Experiment 1. All participants were native speakers of English with no speech or hearing disorders. Participants received either course credit or monetary compensation for their participation in this study.

Materials

All stimulus materials were drawn from the Hoosier multi-talker audio-visual (AV) database (Sheffert, Lachs, & Hernandez, 1997). Monosyllabic, CVC words produced by one female speaker and one male speaker in the database were selected for this study. The stimulus set for each participant contained 96 different word tokens. In each condition, half of the stimuli were “Easy” words – high frequency words from lexically sparse phonological neighborhoods (e.g., “fool”), while the other half were “Hard” words – low frequency lexical items from lexically dense phonological neighborhoods (e.g., “hag”; see Luce and Pisoni, 1998).

Auditory Stimuli. In each condition, we used envelope-shaped noise (Horii, House, & Hughes, 1971) to reduce performance on the spoken word recognition task. The experimental stimuli were created by processing the audio files through a MATLAB program that randomly changed the sign bit of the amplitude level of 30% of the spectral samples in the acoustic waveform. Reducing auditory-only word recognition performance to below-ceiling levels is a necessary prerequisite to detect the effects of cross-modal repetition priming in the spoken word recognition task. Pilot data indicated that this level of noise degradation reduced auditory-only open-set recognition to about 50% correct.

Visual Stimuli. Two kinds of visual primes were used: Static and Dynamic. Dynamic primes consisted of the original, unedited video clips associated with each target word. Previous research has shown that the overall identification accuracy on these stimuli presented in visual-only condition was 14%, with less than 1% of the individual tokens accurately identified more than 90% of the time (Lachs & Hernandez, 1998). Thus, the specific words used in the study were not consistently identifiable in a visual-only condition. The video track of the Static primes consisted of a still shot of the speaker whose duration was identical to that of its counterpart in the Dynamic prime condition. The same still image was used in the Static condition for each target word. This image was taken from a resting state of each speaker.

Procedure

Participants were tested in groups of four or fewer in a quiet room with individual testing booths. During testing, each participant listened to the auditory signals over Beyer Dynamic DT-100 headphones at a comfortable listening level while sitting in front of a Power Mac G4. A customized SuperCard (v4.1.1) stack presented the stimuli to each participant. Participants were instructed to watch the computer monitor and then type the English word that they heard over the headphones using the computer keyboard.

On each trial (see Figure 1), participants first saw either a Static or a Dynamic visual prime. 500 milliseconds after the presentation of the visual prime, participants heard the degraded auditory target word over the headphones. A prompt then appeared on the screen asking the participant to type the word they heard. Presentation of the next stimulus was participant-controlled.

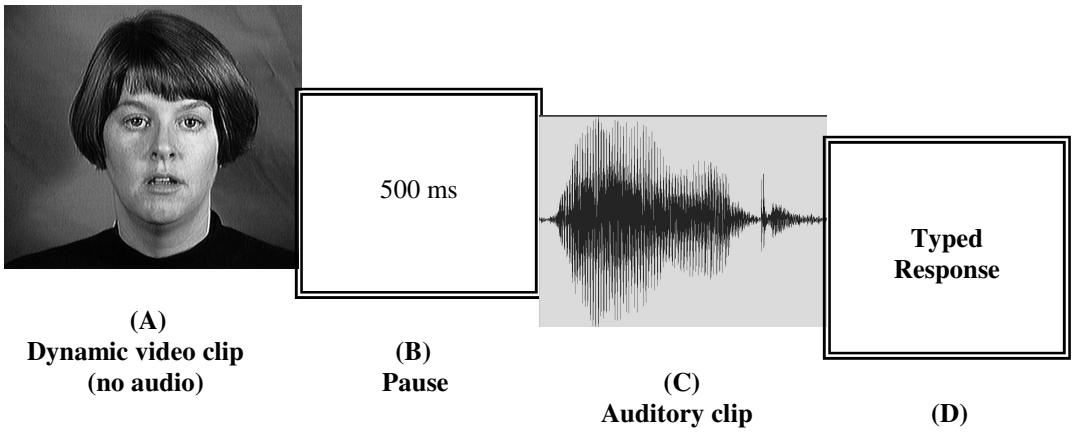


Figure 1. Schematic of trial in cross-modal repetition priming experiment. In Experiment 1, the video clip (A) and auditory clip (C) come from same token of a single speaker. In Experiment 2, (A) and (C) come from the same speaker or from different speakers producing the identical lexical item.

Participants were either presented with all female talker stimuli (both targets and primes) or all male talker stimuli. Words were presented to participants in random order, with Dynamic and Static primes randomly interleaved over the course of the experiment. Each participant responded to 48 words in each priming condition, half of which were lexically "Easy" targets and half of which were lexically "Hard" targets.

Results: Experiment 1

Word Identification Accuracy. For analyses reported in this section, the dependent variable was word recognition accuracy. The results revealed that the participants benefit from the presentation of a dynamic video identity prime when compared to a static prime. Overall, participants in Experiments 1 exhibited a 14% accuracy gain on trials in which a dynamic video prime preceded the degraded audio signal (67%) compared to trials in which a static face prime preceded the auditory target word (53%). The word recognition accuracy data for the female and male talkers were analyzed with separate 2x2 Prime Type (Dynamic/Static) vs. Target Type (Easy/Hard) repeated measures Analyses of Variance (ANOVAs).

The ANOVAs revealed a significant main effect of Prime Type for both the female speaker (Dynamic = 66.8%, Static = 49.1%; $F(1,19) = 166.8$, $p < 0.001$) and the male speaker (Dynamic = 67.2%, Static = 56.7%; $F(1,19) = 166.7$, $p < 0.001$), as well as significant main effects of Target type for both speakers (Female speaker: Easy = 67.1%, Hard = 49.9%; $F(1, 19) = 121.2$, $p < .001$; Male speaker: Easy = 69.9%, Hard = 52.5%, $F(1, 19) = 196.7$, $p < .001$). Thus, better performance was obtained on trials with Dynamic primes compared to trials with Static primes, and on trials with Easy targets compared with trials with Hard targets. The interaction between Prime Type and Target type was not significant for either speaker.

Response Analysis. To enrich our understanding of the information observers perceive and encode in the Dynamic prime condition when compared to the Static prime condition, we performed several analyses comparing the responses participants made on Dynamic trials to those made on Static trials. For the purposes of increasing power over the analyses, the data from participants who observed the Female speaker and those who observed the Male speaker were combined for all analyses reported in this section.

Collapsing over all the data, there were 1920 responses for each trial type. A total of 465 unique responses were given for Dynamic trials, whereas 610 unique responses were given for Static trials. A chi-square analysis revealed that there were significantly more unique responses to Static trials than to Dynamic trials [$\chi^2(1) = 46.40$, $p < 0.01$]. This finding strengthens the whole-word results reported above and indicates that the information present in the Dynamic prime acts as a constraint on the participants' responses to the auditory word presented in noise. Additionally, there were significantly fewer unique responses on trials with Easy targets (476) compared to trials with Hard targets [599; $\chi^2(1) = 19.23$, $p < .01$]. The difference in the number of unique responses for Easy words with the two prime types (Dynamic: 190; Static: 286), and Hard words with the two prime types (Dynamic: 275; Static: 324) approached but did not reach significance [$\chi^2(1) = 3.68$, $p < .06$].

Additional analyses were geared towards exploring the nature of the constraints on the response selection process. Initially, each response was coded for the number of correct segments of the CVC word (i.e., 0-3 segments correct), and the average number of correct segments for each participant was then computed for each condition. A repeated measures ANOVA with Prime Type (Dynamic vs. Static) and Target Type (Easy vs. Hard) as independent variables and overall segmental accuracy as the

dependent variable revealed a main effect for Prime Type [$F(1,39) = 75.81$, $p < .001$], with higher segmental accuracy for targets with Dynamic primes (mean = 2.49, SD = 0.20) than for targets with Static primes (mean = 2.21, SD = 0.19). The ANOVA also revealed a main effect of Target Type ($F(1,39) = 69.46$, $p < .001$), with significantly higher segmental accuracy for Easy targets (mean = 2.44, SD = .28) than for Hard targets (mean = 2.25, SD = .24). There was also a significant interaction between Target Type and Prime Type [$F(1,39) = 12.57$, $p < .001$]. The locus of the interaction indicated that the effect of Target Type was larger for the Dynamic primes (Easy: 2.62; Hard: 2.34) than for the Static primes (Easy: 2.26; Hard: 2.16). Overall, these results support the claim that the responses on trials with Dynamic primes were more constrained (i.e., closer to the target) than the trials with Static primes.

To further address whether incorrect responses are also more constrained when preceded by Dynamic primes, we limited the analysis described above to responses in which the participant gave the wrong whole word response (thus giving a possible range of 0-2 segments correct). Using the number of correct segments in incorrect responses as the dependent variable, we performed 2x2 Prime Type (Dynamic/Static) vs. Target Type (Easy/Hard) repeated measures ANOVA. This analysis again revealed a significant main effect of Prime type [$F(1,39) = 15.35$, $p < .001$]; the number of correct segments on trials with Dynamic primes (mean = 1.47, SD = 0.18) was significantly greater than the number of correct segments on trials with Static primes (mean = 1.34, SD = 0.14). This result indicates that the information present in the dynamic video signal constrains all of the participants' responses, leading to greater accuracy even for incorrect responses.

Additionally, a main effect of Target type was obtained [$F(1,39) = 23.56$, $p < .001$]; however, when the analysis was limited to incorrect responses, the responses to trials with Hard targets had significantly higher overall segmental accuracy (mean = 1.49, SD = .21) than responses on trials with Easy targets (mean = 1.29, SD = .30). This result may at first appear surprising; however, it reflects a significantly higher proportion of incorrect responses with 2 segments correct on trials with Hard targets (558/927, 60.2%) than Easy targets (257/621, 41.4%; $\chi^2(1) = 52.02$, $p < .001$). Given the working definition of lexical neighbors as words sharing N-1 segments of an N-segment word (which was used to generate the Easy/Hard targets for this experiment; see Luce and Pisoni, 1998), participants' incorrect responses that contain two correct segments are, by definition, lexical neighbors of the target word. Thus, incorrect responses to Hard targets (words from lexically dense neighborhoods) were more likely to be neighbors of the target than incorrect responses on trials with Easy targets (words from lexically sparse neighborhoods). The interaction between Prime type and Target type was significant [$F(1,39) = 6.78$, $p < .05$], with the effect of Target type attenuated for Dynamic prime trials (Easy: 1.43; Hard: 1.53) compared to Static prime trials (Easy: 1.16; Hard: 1.46). Thus, the effect of Target type was stronger in the condition when there was no dynamic visual information about the target, further suggesting that this additional optical information provides a constraint on participants' open-set word identification responses.

The above results reveal that incorrect responses on trials with Dynamic primes are closer to the target than incorrect responses on trials with Static primes. To gain a more detailed understanding of how the dynamic video information constrains responses on the word recognition task, we analyzed the likelihood of correct responses for each syllable position of the CVC words as a function of Prime type.[4] These data, presented in Table 1, show that the accuracy is greater for words in the Dynamic condition

[4] For the remaining analyses, we report data collapsed over Easy/Hard targets, as the data within each of these Target types matches the overall pattern of the data. We return to a discussion of the effects of Target type in Experiment 2 and in the General Discussion.

compared to the Static condition for each of the syllable positions, revealing that the dynamic information helped constrain responses for all three syllabic positions of the CVC words.

	Dynamic % SD	***Static*** % SD	***Analysis***
Onset	80.3 (8.3)	67.6 (7.4)	$t(39) = 9.18$, $p < .001$
Nucleus	87.1 (6.1)	78.4 (7.3)	$t(39) = 6.38$, $p < .001$
Coda	81.3 (7.9)	74.8 (8.3)	$t(39) = 4.39$, $p < .001$

Table 1. Response accuracy for each of the three syllable positions in the CVC stimuli as a function of prime type (Experiment 1).

To determine whether there was a difference in the accuracy benefit for any of the three positions, we computed a difference score (Dynamic – Static) for each syllable position. Planned comparisons indicated that the cross-modal priming effect was significantly greater for onset position than it was for either nucleus position [$t(39) = 2.51$, $p < .05$] or for coda position [$t(39) = 3.58$, $p < .01$], but there was no difference between accuracy on the nucleus position and coda position [$t(39) = 1.49$, *ns*].

The data analyzed in this section thus far suggest that there was a global benefit from the dynamic visual information which constrained all components of the participants' responses, and that this effect was particularly robust for onset position. However, it remains possible that the information in the dynamic video clip constrained responses by limiting specific components of the set of competing hypotheses about the target word. To address this possibility, we examined the participants' identification of particular phonological properties of the target stimulus. Specifically, we examined the likelihood that participants would correctly identify the place features, manner features, and voicing features of the onset and coda consonants in the target word. These analyses were performed by collapsing the data obtained from all forty subjects, and comparing the accuracy on these individual dimensions for target words with Dynamic primes and target words with Static primes. The results of these analyses are presented in Table 2.

		Dynamic % correct	***Static*** % correct	***Analysis***
Place	Onset	86	76	$\chi^2(1) = 52.66$, $p < 0.001$
	Coda	90	85	$\chi^2(1) = 19.45$, $p < 0.001$
Manner	Onset	88	79	$\chi^2(1) = 58.89$, $p < 0.001$
	Coda	89	86	$\chi^2(1) = 7.28$, $p < 0.01$
Voice	Onset	98	97	$\chi^2(1) = 2.28$, *ns*
	Coda	97	95	$\chi^2(1) = 2.25$, *ns*

Table 2. Response accuracy in reporting features for onset and coda consonants as a function of prime type (Experiment 1).

The data in Table 2 indicate that the dynamic video clip primes created a robust increase in accuracy with respect to place and manner of articulation for both onset and coda consonants. This result suggests that the participants were able to use the optical information available in the dynamic video clip to limit the set of possible responses, and that this information was useful in specifying both place and

manner of articulation. With respect to voicing, we limited our analysis to those trials in which the target and response were obstruents and thus the voice feature would have to be specified as part of the response. Not surprisingly, there was no significant effect of prime type on accuracy of the voice feature, a finding that is consistent with the hypothesis that voicing is poorly specified in visual-only speech signals (Summerfield, 1979).

Discussion

The data presented above help to sharpen our understanding of the information specified in different sensory modalities used in speech perception. In particular, we presented evidence that a visual-only speech signal facilitates identification of asynchronously presented auditory speech when the latter is presented in noise. This result complements and builds on previous results in the literature indicating that speech perception is not limited to the auditory modality (e.g., Sumby & Pollack, 1954; Massaro, 1998; Bernstein, 2005). We will return to a discussion of these broad issues in the General Discussion.

More specifically, Experiment 1 provided critical evidence suggesting that observers are able to integrate information presented in two modalities, even when the signals are separated in time; thus, temporal synchrony is not a necessary condition for audiovisual integration to be observed. This finding converges with the results reported by Kim et al. (2004), who found similar asynchronous cross-modal priming in a lexical decision task (also see Dodd et al., 1989). In an attempt to build on their earlier results, we explored an additional factor which may be a necessary feature for audiovisual integration in repetition priming: source identity of the two input modalities. In our view, the strong version of the Direct Realism approach maintains no role for abstract internal representations or cognitive operations in perception (Gibson, 1966), and should hold that identity of the source of the two streams of information is a necessary condition for audiovisual integration in speech perception. However, if we find that repetition priming due to audiovisual integration persists even when the visual and auditory signals come from different speech events that are temporally asynchronous, we must conclude that there is some additional component to audiovisual integration which relies on activation of common abstract (i.e., removed from the signal; non-episodic) representations in addition to the perception of the specific event itself.

Experiment 2

The second experiment sought to extend the findings of Experiment 1 by presenting participants with trials in which the visual signal and the auditory signal were produced by distinct talkers. To achieve this goal, we used two speech sources that would clearly be perceived as different speakers: a male face/voice and a female face/voice. This experiment represents an extension and elaboration of Green et al.'s (1991) intriguing finding that observers will perceive the typical McGurk illusion even with a gender mismatch between the face and voice. Thus, there exists some prior experimental evidence that source identity is not a necessary condition for audiovisual integration. The present study seeks to extend this finding by investigating whether audiovisual integration – indexed by repetition priming – may be achieved with neither source identity nor temporal synchrony.

Participants

Twenty-six Indiana University undergraduate students, ages 18-23, participated in Experiment 2. All participants were native speakers of English with no speech or hearing disorders. Participants received either course credit or monetary compensation for their participation in this study. None of the participants from Experiment 2 had participated in Experiment 1.

Materials

As with Experiment 1, all stimulus materials were drawn from the Hoosier multi-talker audio-visual (AV) database (Sheffert et al., 1997). Monosyllabic, CVC words produced by the same female speaker and male speaker as in Experiment 1 were selected for this study. In Experiment 2, 240 different word tokens were used. As with Experiment 1, half of the stimuli were "Easy" words – high frequency words in sparse phonological neighborhoods (e.g., "fool"), while the other half were "Hard" words – low frequency lexical items in high density neighborhoods (e.g., "hag"; Luce and Pisoni, 1998).

Procedure

The testing situation was identical to that used in Experiment 1. Each participant was presented with eight different trial types, with all permutations of prime type (Dynamic vs. Static), prime gender (Female vs. Male), and target gender (Female vs. Male). The experimental trials were analyzed as two groups: Matched (Female prime and Female target; Male prime and Male target) and Mismatched (Female prime and Male target; Male prime and Female target).

Results

Word Identification Accuracy. Data from Experiment 2 were analyzed with a 2x2x2 Prime type (Dynamic/Static) vs. Target type (Easy/Hard) vs. AV-matching (Matched/Mismatched) ANOVA. Consistent with the results reported from Experiment 1, the results indicated a significant main effect of Prime type; words from Dynamic prime trials were identified more accurately (mean = 65.6%) than words from Static prime trials (mean = 54.4%; $F(1, 25)=108.3$, $p < .001$). A significant main effect of Target type was also observed, with Easy targets recognized more accurately (mean = 65.1%) than Hard Targets (mean = 54.9%, $F(1, 25) = 85.6$, $p < .001$). No significant main effect was found for AV Matching ($F(1, 25) = 0.9$, *ns*), reflecting the lack of a difference in overall accuracy on AV-Matched and AV-Mismatched trials, regardless of Prime or Target type. Critical planned comparisons examined effects of Prime type separately for AV-Matched and AV-Mismatched trials. These comparisons revealed a significant effect of Prime type for both Matched (Dynamic: mean = 66.6%; Static: mean = 55.1 %; $t(25) = 3.27$, $p < .01$) and Mismatched (Dynamic: mean = 64.4%; Static: mean = 54.4%; $t(25) = 4.01$, $p < .001$), revealing that the spoken word recognition priming effect observed in the single-speaker condition does not crucially rely on the signals in the two stimulus presentation modalities coming from an identical source. No significant interactions were obtained in the two-speaker conditions (all Fs < 1.6).

Response Analysis. We performed the same analyses on the set of responses in Experiment 2 as we did in Experiment 1. Collapsing over all the data, there were 3120 responses to targets with Dynamic primes and 3120 responses to targets with Static primes. A total of 1010 unique responses were given for Dynamic trials, whereas 1180 unique responses were given for Static trials. A chi-square analysis revealed that there were significantly more unique responses to Static trials than to Dynamic trials [$\chi^2(1) = 19.70$, $p < 0.01$]. This finding strengthens the results reported above, indicating that the information present in the Dynamic prime acts as a constraint on the participants' responses to the auditory word presented in noise. When we examined the number of unique responses on the 1620 Matched Dynamic trials (659 unique responses) and the 1620 Mismatched Dynamic trials (700 unique responses), there was no significant difference between these two groups [$\chi^2(1) = 1.98$, *ns*], indicating that there was no difference in the constraint on responses for these conditions reflected by the number of unique responses for Matched trials and for Mismatched trials. Overall, there were more unique responses to Hard targets (893 unique responses) than to Easy targets [769 unique responses; $\chi^2(1) = 18.59$, $p < .01$]. No significant

differences were found in the proportion of unique responses to Easy and Hard words for any of the priming conditions (Dynamic Matched: Easy – 309, Hard – 350; Dynamic Mismatched: Easy – 324, Hard – 376; Static: Easy – 524, Hard – 638).

Following the analyses used in Experiment 1, each response was coded for the number of correct segments of the CVC word (i.e., 0-3 segments correct), and the average number of correct segments for each participant was computed for each condition. A repeated measures ANOVA revealed a significant main effect of Prime type [$F(1,25) = 69.26$, $p < .001$] on the number of correct segments, with responses on trials with Dynamic primes (mean = 2.46, SD = 0.15) having significantly more correct segments than responses on trials with Static primes (mean = 2.27, SD = 0.14). The difference between the Matched (mean = 2.48, SD = .16) and the Mismatched (mean = 2.43, SD = .17) groups was not significant, though there was a trend towards better performance on Matched trials ($t(25) = 2.03$, $p < .06$). When compared to the static trials, performance on Dynamic trials was significantly better for both Matched ($t(25) = 8.41$, $p < .001$) and Mismatched ($t(25) = 6.72$, $p < .001$) trials. These data provide further support for the claim that the responses are constrained by the presence of the optical information available in the dynamic video clip. This ANOVA also revealed a significant main effect of Target type [$F(1,25) = 34.71$, $p < .001$], with responses on trials with Easy targets having more segments correct (mean = 2.41, SD = .17) than responses on trials with Hard targets (mean = 2.30, SD = .19). The interaction between Prime type and Target type was not significant.

When the analysis was limited to responses in which the participant gave the wrong whole word response, a repeated measures ANOVA revealed a significant main effect of Prime type [$F(1,25) = 14.10$, $p < .05$], with the number of correct segments on trials with Dynamic primes (mean = 1.43, SD = .28) significantly greater than the number of correct segments on trials with Static primes (mean = 1.36, SD = .18). There was no significant difference between performance on Matched (mean = 1.44, SD = .21) and Mismatched (mean = 1.42, SD = .16) trials ($t(25) = 0.69$, *ns*). When compared to the number of segments correct in incorrect responses for the Static condition, there were significantly more segments correct for Dynamic trials in both the Matched ($t(25) = 2.34$, $p < .05$) and Mismatched ($t(25) = 2.10$, $p < .05$) AV conditions. This latter result confirms again that the information present in the dynamic video signal constrains all of the participants' responses leading to greater accuracy even for incorrect responses, and that this effect is not attenuated by having a gender mismatch between the source of the dynamic video prime and the auditory target.

The ANOVA also revealed a significant main effect of Target type [$F(1,25) = 26.39$, $p < .05$], with the number of segments correct in incorrect responses higher for Hard targets (mean = 1.48, SD = .14) than for Easy targets (mean = 1.31, SD = .19). As in Experiment 1, this reflects a greater number of neighbors given as responses for Hard targets (807/1406, 57.4%) than for Easy targets (521/1089, 47.8%; $\chi^2 = 22.12$, $p < .05$). The interaction between Prime type and Target type was not significant for Experiment 2.

Following the analyses in Experiment 1, we analyzed the likelihood of correct responses for each syllable position of the CVC words as a function of Prime type (collapsing over Target types). These data are presented in Table 3, with Matched and Mismatched conditions listed separately as well as combined. These data indicated that the overall accuracy is increased for words in the Dynamic condition compared to the Static condition for each of the syllable positions, revealing that the dynamic information helped constrain responses for all three segments of the CVC words.

		Dynamic % SD	*Static* % SD	*Analysis*
Onset	**Total**	**79.9 (6.7)**	**70.4 (5.0)**	***t*(25) = 8.35, *p* < .001**
	Matched	81.3 (6.3)		$t(25) = 9.76$, $p < .001$
	Mismatched	78.3 (8.4)		$t(25) = 5.40$, $p < .001$
Nucleus	**Total**	**84.6 (6.1)**	**78.6 (5.5)**	***t*(25) = 5.82, *p* < .001**
	Matched	*85.4 (6.4)*		$t(25) = 6.01$, $p < .001$
	Mismatched	*83.7 (6.0)*		$t(25) = 4.25$, $p < .001$
Coda	**Total**	**81.7 (4.9)**	**76.8 (4.6)**	***t*(25) = 4.86, *p* < .001**
	Matched	81.8 (6.1)		$t(25) = 4.19$, $p < .001$
	Mismatched	81.5 (5.3)		$t(25) = 4.02$, $p < .001$

Table 3. Response accuracy for each of the three syllable positions in the CVC stimuli as a function of prime type (Experiment 2). Matched and Mismatched Dynamic trials are compared to overall data from Static trials.

To determine whether there was a difference in the accuracy benefit for any of the three positions, we computed a difference score for each syllable position. Overall planned comparisons indicated that the cross-modal priming effect was significantly greater for onset position than it was for either nucleus position [$t(25) = 3.07$, $p < .01$] or for coda position [$t(25) = 3.46$, $p < .01$], but there was no difference between accuracy on the nucleus position and coda position [$t(25) = 0.88$, *ns*]. Comparisons limited to Matched and Mismatched dynamic trials exhibit the same pattern, with onset position having significantly greater priming benefit than nucleus or coda, and with no significant difference observed between nucleus and coda.

The analyses of the data from Experiment 2 presented thus far suggest that there was a global benefit from the dynamic information which constrained all components of the participants' responses. Further, these effects were observed even when there was neither temporal synchrony nor source identity of the auditory and dynamic video speech signals. As discussed above, it is critical to investigate whether the priming benefit reflects a general benefit from the information present in the video clip, or whether the responses are constrained by the stimulus by limiting specific components of the set of competing hypotheses about the target word.

Following the analyses in Experiment 1, we examined the likelihood that participants would correctly identify particular phonological properties of the target stimulus. In particular, we examined the likelihood that participants would correctly identify the place features, manner features, and voicing features of the onset and coda consonants in the target word. The results are presented in Table 4.

The data in Table 4 indicate that the dynamic video clip primes promote a robust increase in accuracy with respect to place and manner of articulation for both onset and coda consonants. Crucially, these effects hold for both Matched and Mismatched primes; that is, the responses were significantly more accurate for both place and manner features even when the prime and target were presented asynchronously and when they came from a different source. The performance on Matched and Mismatched trials did not differ significantly for any comparisons in Table 4 other than Onset place, where the identification of place for Matched trials was significantly better than identification of place for Mismatched trials ($\chi^2 = 8.68$, $p < .05$). With respect to voicing, following the analyses in Experiment 1, we limited our analysis to those trials in which the target and response were obstruents and thus the voice feature would have to be specified as part of the response. As with Experiment 1, there was no

significant effect of prime type on accuracy of the voice feature. This result is again consistent with the claim that voicing is not well-specified as part of the visual-only speech signal and that the other attributes of responses are under stimulus control.

Feature	Position	Prime Type	*Dynamic* %	*Static* %	*Analysis*
Place	Onset	**Total**	**86**	**78**	**$\chi^2(1) = 52.3, p < 0.001$**
		Matched	87		$\chi^2(1) = 54.6, p < 0.001$
		Mismatched	84		$\chi^2(1) = 17.2, p < 0.001$
	Coda	**Total**	**89**	**85**	**$\chi^2(1) = 26.0, p < 0.001$**
		Matched	89		$\chi^2(1) = 13.8, p < 0.001$
		Mismatched	89		$\chi^2(1) = 19.4, p < 0.001$
Manner	Onset	**Total**	**88**	**83**	**$\chi^2(1) = 29.7, p < 0.01$**
		Matched	89		$\chi^2(1) = 26.7, p < 0.01$
		Mismatched	87		$\chi^2(1) = 12.4, p < 0.01$
	Coda	**Total**	**90**	**88**	**$\chi^2(1) = 8.18, p < 0.05$**
		Matched	90		$\chi^2(1) = 6.48, p < 0.05$
		Mismatched	90		$\chi^2(1) = 4.10, p < 0.05$
Voice	Onset	**Total**	**98**	**97**	**$\chi^2(1) = 3.08$, *ns***
		Matched	98		$\chi^2(1) = 0.72$, *ns*
		Mismatched	98		$\chi^2(1) = 2.47$, *ns*
	Coda	**Total**	**95**	**94**	**$\chi^2(1) = 1.80$, *ns***
		Matched	95		$\chi^2(1) = 3.55$, *ns*
		Mismatched	95		$\chi^2(1) = 1.96$, *ns*

Table 4. Accuracy in identifying the place, manner, and voice for onset and coda consonants. Statistical analyses compare the performance on static trials to performance on total Dynamic trials, as well as to Matched and Mismatched trials separately.

General Discussion

The experimental work reported in this paper reflects a novel application of the conventional repetition priming paradigm. Here, we used this paradigm to investigate central issues pertaining to the nature of multimodal speech perception. Participants were required to identify spoken words presented in envelope-shaped noise that were preceded by dynamic or static visual-only primes. In Experiment 1, the results indicated that participants were more accurate at identifying spoken words when the auditory stimulus was preceded by a dynamic visual stimulus of the same word compared to a static image of the speaker's face. Furthermore, detailed analyses of the participants' responses indicated that the dynamic video information constrained the responses to the auditory target even on trials where spoken word recognition was not successful. In Experiment 2, the same priming benefit was observed even when it was readily apparent that the auditory and visual signals came from different speakers.

These results raise several important issues regarding the nature of multimodal speech perception. First, we have demonstrated that cross-modal repetition priming in speech perception requires neither temporal synchrony nor source identity; the repetition priming effect was observed even when the commonality that exists between the dynamic video clip prime and auditory target was only at

the level of the lexical identity of the token being produced, and not identity of the token or specific "episode" that is being perceived. This result is consistent with a view of multimodal speech perception in which integration of auditory and visual information is part of the cognitive process(es) involved in speech perception (Bernstein, 2005; Hamilton et al., 2006; Kim et al., 2004; Massaro & Stork, 1998).[5] According to this account, language users store and maintain in memory abstract, internal representations of the external auditory world, such as a representation of the speech sound /p/. The results of the cross-modal repetition priming experiments reported here suggest that these representations may be activated directly by an acoustic waveform containing particular sounds, and they may also be activated (either directly or indirectly) by dynamic visual displays of a speaker creating the articulatory gesture that produces the same speech sound (e.g., a labial closure).

The results reported here also reveal that the nature of the benefit observers received from the dynamic video prime was under tight stimulus control. In particular, the participants' responses were constrained in several important ways. First, more correct responses to auditory targets were observed on trials with dynamic video clip primes. Second, across responses from all participants, there was a smaller range of responses provided on trials with dynamic primes compared to static primes. Third, the presentation of the dynamic primes increased identification of segments in all three of the syllable positions of the CVC targets, with onsets benefiting more than the nucleus and coda. Fourth, the responses on trials with dynamic primes were more likely to exhibit accurate identification for two kinds of sub-segmental information: place of articulation and manner of articulation of both onset and coda consonants. In contrast, dynamic primes did not significantly increase the likelihood of accurately reporting the correct voicing status of the target obstruents, revealing that the components of the speech signal that are not available in the visual speech stream did not receive a benefit from the dynamic visual display.

Audiovisual Integration and Cross-Modal Identity Matching

Another line of research in the multimodal speech perception literature has revealed that perceivers are able to match a video of a speaker's face to the appropriate corresponding voice when visual and auditory stimuli are presented separately (Lachs, 2002; Lachs & Pisoni, 2004a, 2004b). The cross-modal matching task can be performed successfully even when the linguistic content of the two signals differs (Kamichi et al., 2003), suggesting that the perceptual cues used for cross-modal identity matching are independent of the idiosyncrasies of a particular utterance.

Lachs and Pisoni (2004a, 2004b) suggested that their participants' success in cross-modal identity matching – in which the correctly matched stimuli came from the same utterance – may be rooted in event-based perception (Gibson, 1966). Lachs and Pisoni's auditory and visual stimuli provided information about the same physical event in the world, and they argued that "integration" of the two modalities of information came from the real-world event itself, which shaped and constrained the pattern of sensory stimulation impinging on the eyes and ears. Within the direct realist event-based theoretical framework (Fowler, 1986), the locus of audiovisual integration is in the real world, and acoustic and optical speech signals are integrated seamlessly because they are two sources specifying information about the same distal event (also see Fowler, 2004).

[5] This type of theoretical approach posits that sensory information from the world is encoded in modality-specific representations, and that these modality-specific representations are either: a) linked directly to one another (Massaro & Stork, 1998); or b) linked to a separate "multimodal" representation that integrates information from the different sources (Skipper et al., 2005; Hamilton et al., 2006). However, the difference between these proposals cannot be addressed by the research reported here.

In our view, the results of Experiment 2 – which provided clear and consistent evidence indicating that the effects of priming on both overall accuracy and in a detailed error analyses are maintained even in a condition where there was a mismatch between the speakers – suggest that an event-based perception account would need to additionally permit a level of abstraction in the multimodal speech perception process. Experiment 2 presented listeners with visual-only primes and auditory-only targets which were lexically identical (e.g., both stimuli are "cat"), but clearly produced by two different speakers (one male, one female). Thus, the prime and target stimuli came from two different perceptual events in the world. If a strong version of the event-based perspective on audiovisual integration outlined above were correct, the repetition priming effect should be absent in this condition. When the speakers differ – as in our experimental manipulation – the two sensory input modalities no longer provide the perceiver with sensory information about the same event in the world. However, if perceptual identity is defined with respect to the articulatory gestures that create the visual and auditory percept (e.g., [p] defined as voiceless labial stop, regardless of who produces it), then there is no reason to predict that the cross-modal priming effect would be absent when there is a lack of identity in the source of the two stimuli. However, it is worth noting that accounting for the data presented above requires that the identity between the two signals is processed at some abstract level of representation (e.g., identity of underlying segments without identity in the actual events producing the segments).

Open-set Identification and Lexical Access

One additional finding which emerged from this study provides further insight into the nature of lexical competition in the process of lexical access regardless of input modality. For both experiments reported here, when we looked at the incorrect whole word responses (i.e., failures of lexical access), we observed more correct segments on trials with "hard" target words (low frequency words from dense lexical neighborhoods) than on trials with "easy" target words (high frequency words from sparse lexical neighborhoods). This finding was largely attributable to a larger number of incorrect responses with two segments correct on trials with hard targets than on trials with easy targets. The definition of lexical neighbor used in this paper, based on Luce and Pisoni (1998), was a word that shares all but one segment with the target word. Thus, it was more likely that incorrect responses for "hard" targets were neighbors of the target word (i.e., sharing two of the three segments) than it was that incorrect responses for "easy" targets were neighbors of the target word. While this result follows from the Neighborhood Activation Model (NAM) of Luce and Pisoni (1998) in a straightforward manner, it is a novel empirical demonstration of a critical component of NAM.

NAM holds that the strength and number of competitors directly influences the ease with which lexical items are accessed (Luce & Pisoni, 1998). Previous attempts to understand the role of neighborhood density in lexical access have typically focused on the increase in accuracy and processing time (Luce & Pisoni, 1998; Vitevitch & Luce, 1998, 1999; Vitevitch, Luce, Pisoni, & Auer Jr., 1999) for words with strong competitors (i.e., "hard" words) compared to words with weaker competitors (i.e., "easy" words). However, previous accounts have not included detailed response analyses of the type presented in this paper. The results reported here provide further support for the fundamental claim underlying NAM by demonstrating that when lexical access fails, the response is more likely to be a lexically similar neighbor/competitor for "hard" words than it is for "easy" words.

Conclusion

We reported results from a cross-modal priming study in which identification of spoken words mixed with noise was facilitated by the earlier presentation of a dynamic video clip of the utterance compared to a static image of a speaker. The present set of findings indicate that neither temporal

synchrony in the presentation of the two signals nor identity in the source of the two signals is a necessary precondition for audiovisual integration in multimodal speech perception, suggesting that the set of neural and cognitive processes involved in multimodal speech perception includes activation of abstract representations of speech. The cross-modal repetition priming paradigm can be used in the future to provide critical new information pertaining to the nature of multimodal representations of speech by exploring the nature of the stimuli that produce this effect. We expect that these lines of research will converge to address additional issues related to multimodal perception of linguistic information, such as the time-course of audiovisual integration in speech perception processes or the neural mechanisms underlying repetition priming (see Grill-Spector, Henson, & Martin, 2006 for a recent review) and multimodal perception (e.g., see Ghazanfar & Schroeder, 2006). In addition, these lines of research are relevant to understanding the relation of the two input modalities in clinical populations such as hearing-impaired listeners who have experienced a period of auditory deprivation that may encourage reorganization and remodeling of the typical developmental processes (Bergeson & Pisoni, 2004).

References

Bergeson, T. R., & Pisoni, D. B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Bernstein, L. E. (2005). Phonetic processing by the speech perceiving brain. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 79-98). Malden, MA: Blackwell.

Calvert, G. A., Spence, C., & Stein, B. E. (Eds.). (2004). *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Davis, C., & Kim, J. (2001). Repeating and Remembering Foreign Language Words: Implications for Language Teaching Systems. *Artificial Intelligence Review, 16*, 37-47.

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *Quarterly Journal of Experimental Psychology, 57A*(6), 1103-1121.

Dodd, B., Oerlemens, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Language, 22*, 59-77.

Fowler, C. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3-28.

Fowler, C. (2004). Speech as a supramodal or amodal phenomenon. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Science, 10*, 278-285.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 38*, 269-276.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Science, 10*(1), 14-23.

Hamilton, R. H., Shenton, J. T., & Coslett, H. B. (2006). An acquired deficit of audiovisual speech processing. *Brain and Language, 98*, 66-73.

Horii, Y., House, A. S., & Hughes, G. W. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America, 49*, 1849-1856.

Kamichi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': Matching identity across modality. *Current Biology, 13*, 1709-1714.

Kim, J., & Davis, C. (2003). Task effects in masked cross-script translation and phonological priming. *Journal of Memory and Language, 49*, 484-499.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition, 93*(1), B39-B47.

Lachs, L. (Ed.). (2002). *Vocal tract kinematics and crossmodal speech information*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Lachs, L., & Hernandez, L. R. (1998). Update: The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.

Lachs, L., & Pisoni, D. B. (2004a). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 30*(2), 378-396.

Lachs, L., & Pisoni, D. B. (2004b). Crossmodal source identification in speech perception. *Ecological Psychology, 16*(3), 159-187.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1-36.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science, 4*, 104-109.

Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist, 86*, 236-244.

Mattys, S. L., Bernstein, L. E., & Auer Jr., E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics, 64*(4), 667-679.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 51-78). Malden, MA: Blackwell.

Sheffert, S., Lachs, L., & Hernandez, L. R. (1997). The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage, 25*, 76-89.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.

Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36*, 314-331.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2006). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in spoken word perception. *Psychological Science, 9*, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language, 40*, 374-408.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer Jr., E. T. (1999). Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language, 68*, 306-311.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*, 23-43.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Frequency and the Emergence of Prefabs: Evidence from Monitoring[1]

Vsevolod Kapatsinski and Joshua Radicke

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] We would like to thank the NIH for financial support through Training Grant DC-00012 and Research Grant DC-00111 to David Pisoni. We are very grateful to Luis Hernandez for his help in creating the experimental program. Many thanks also go to Joan Bybee, Chris Conway, Jill Morford, David Pisoni and Rena Torres-Cacoullos for helpful comments.

Frequency and the Emergence of Prefabs: Evidence from Monitoring

Abstract. Native English speakers were instructed to detect instances of /ʌp/ in spoken sentences by pressing a button as soon as they hear /ʌp/ regardless of whether it is inside another word. We observe that detection of the particle *up* is slower when the frequency of the verb+*up* collocation is low or extremely high than when it is medium. In addition, /ʌp/ is more difficult to detect in high-frequency words than medium-frequency or low-frequency words. Thus word frequency has a monotonic effect on detectability of word parts while the effect of phrase frequency is U-shaped. These results support the hypotheses that lexical units compete with their parts during speech perception and that words and ultra-high-frequency phrases are stored in the lexicon.

Introduction

There is much evidence that language users are sensitive to co-occurrence statistics between words in both perception and production. Just in perception, MacDonald (1993) observes that a noun can bias the interpretation of the following word that is ambiguous between a noun reading and a verb reading. McDonald and Shillcock.(2004) and Underwood et al. (2004), using eye-tracking, find that words that are probable given the preceding word or words are fixated for a shorter time than words that are not probable. Bod (2001) finds that subjects are faster in deciding that a three-word subject-object-verb sentence is grammatical when the sentence is frequent (*I like it*) than when it is not (*I keep it*). Reali and Christiansen (2007) present self-paced reading data that shows center-embedded relative clauses to be read faster when the embedded clause consists of a frequent pronoun-verb combination (*I liked*) than when it consists of an infrequent one (*I phoned*). Thus the frequency with which words co-occur (or some other co-occurrence statistic) must be stored in memory. The question we address is what effect frequent co-occurrence has on the memory representation of a pair of words.

One hypothesis, which we shall call **the distributed account**, is that co-occurrence simply increases the strength of an associative connection between the co-occurring words. Another hypothesis, **the localist account**, is that the co-occurring words fuse into a larger unit, the prefab, which has its own separate representation in memory (e.g., Bybee, 2001: 60-62; 2002; Solan et al., 2005; Wray, 2002). This does not mean that the representations for the component words are lost as a result of the fusion. They may well be retained and even used during the production and perception of the frequent phrase. However, under the localist account, the prefab has its own node in the lexicon. That is, the prefab is a lexical unit, just like the words and morphemes that it contains. As Wray (2002: 265) puts it, a formulaic sequence is morpheme-equivalent.

Both theories can account for the finding that high-frequency phrases are processed more easily. In a high-frequency phrase, the end is somewhat predictable given the beginning and will therefore be easier to perceive. Sensitivity to predictability does not necessarily imply that the predictor and the predicted fuse into a unit. Rather, co-occurrence may simply make the co-occurring words able to prime each other.

However, in order to predict that high-frequency phrases are processed more easily than low-frequency phrases, the distributed account must predict that the more predictable a word, the easier it is to process and detect (due to contextual priming). In particular, the final word of a frequent phrase should

be perceived more easily than the final word of a less frequent phrase because the final word of a frequent phrase is predictable given the rest of the phrase and is primed by it.

This is not necessarily the case under a localist account in which prefabs are processed more easily (in part) because they are stored in the lexicon. The predictions of the localist account depend on how the processing of lexical units is hypothesized to interact with the processing of the units' parts. If one assumes that recognition of the whole helps with recognition of the parts (as, for instance, in the Interactive Activation Model of McClelland and Rumelhart 1981), then the localist account makes the same prediction as the distributed one (Healy 1994). If, on the other hand, recognition of the lexical unit interferes with processing of the unit's parts (Healy 1976), parts of high-frequency lexical units (i.e., prefabs) are predicted to be more difficult to detect than parts of low-frequency lexical units.

The idea of between-level competition during lexical access has been proposed independently by Healy (1976), Hay (2003) and Sosa and MacFarlane (2002). Corcoran (1966) and Healy (1976) observed more letter detection errors on the ultra-high-frequency word 'the' than on other words, e.g., the low-frequency word 'thy'. Furthermore, frequency has an effect even when grammatical class is controlled: letters are more difficult to detect in high-frequency nouns than in low-frequency nouns (Healy, 1976; Minkoff & Raney, 2000). Healy proposed the Unitization Hypothesis to account for the result:

> We can […] identify […] syllables, words, or even phrases, without having to complete letter identification. The identification of these higher-order units is facilitated by familiarity […] Once a larger unit is identified, the processing of its component letter units is terminated, even if the letters have not yet reached the point of identification. Instead, processing and attention are directed to the next location in the text. Because letter identification is not always completed for highly familiar words […] many letter-detection errors are made on these words. (Healy, 1994: 333)

A limitation of the work using orthographic stimuli is that the results could be due to the fact that readers are less likely to fixate low-frequency words than high-frequency words during reading (Corcoran, 1966; Inhoff & Rayner, 1986). High-frequency words can be perceived parafoveally, where visual acuity is lower, which may impair the reader's ability to identify individual letters within words. Consistently with this interpretation, Hadley and Healy (1991) found that letter detection is no harder in *the* than in other words when subjects can view only five letters at once while reading text and thus are forced to fixate every word.

In the auditory modality, Sosa and MacFarlane (2002) found that detecting the word *of* in spoken sentences taken from the Switchboard Corpus was more difficult when *of* occurred in an ultra-high-frequency phrase such as *kind of* or *sort of* than when it occurred in a lower-frequency phrase, such as *couple of* or *think of*. No difference between medium-frequency and low-frequency collocations was found. Sosa and MacFarlane argue that extremely frequent phrases (prefabs) are stored in the lexicon and thus detecting *of* in them entails the extra step of morphological decomposition.

A limitation of Sosa and MacFarlane's study is that *of* undergoes much articulatory reduction in high-frequency collocations, such as *kind of* or *sort of*, often appearing without the consonant. This introduces a dilemma for investigating detectability of *of* in such phrases: if a reduced token of *of* is used, it is acoustically non-salient and difficult to perceive as well as being difficult to perceive as an instance of *of*. If a non-reduced token is used, then one is presenting the subject with an instantiation of *of* that is not typical for the context in which it appears. In either case, reaction times may be slowed down for reasons other than the collocation being stored as a single unit.

Thus, in the present study we asked subjects to monitor spoken sentences for a stimulus that does not show much articulatory reduction, the particle *up*. As Sosa and MacFarlane did with *of*, we examine the influence of the frequency of the prefab in which *up* occurs on how easy *up* is to detect. Based on Sosa and MacFarlane's results, we would expect *up* to be more difficult to detect when it occurs in a high-frequency verb+*up* combination like *sign up* than in a less frequent one like *pin up* or *run up*. Using *up* should allow us to test the idea that "it is frequency of use itself that determines the units of storage […] The fact that the phrase is not (yet) reduced does not mean that it is not stored in memory as a unit" (Bybee, 2001: 161). If high-frequency verb+*up* combinations are stored as lexical units, we would find evidence in support of the idea that item-specific phonological behavior is not a necessary precondition for storage.

Despite the fact that Sosa and MacFarlane did not find differences between low-frequency and medium-frequency phrases, there are reasons to suspect that *up* should be harder to detect in low-frequency phrases than in medium-frequency ones. Morton and Long (1976) and Dell and Newman (1980) found that phoneme detection was faster in words that were relatively predictable given the part of the sentence that preceded them relative to words that were not predictable[2], e.g., *book* vs. *bill* following *He sat reading a*; and *beer* vs. *brandy* following *He had a drink of* (from Morton & Long, 1976). While at first glance this result appears to conflict with the results of Sosa and MacFarlane (2002), predictability of *beer* in *He had a drink of beer* is much lower than the predictability of *of* in *This was done kind of badly*. Conversely, *of* is still relatively predictable in the lowest-frequency collocations used by Sosa and MacFarlane (2002), e.g., *sense of*, *piece of*, *each of*. Thus, existing evidence points to a U-shaped effect of phrase frequency on detectability of the phrase's parts: parts of a low-frequency phrase should be harder to detect than parts of a medium-frequency phrase which should be easier to detect than parts of an ultra-high-frequency phrase.

One type of model that predicts a U-shaped effect of phrase frequency on part detectability is one that assumes that a collocation is likely to be stored in the lexicon only if its frequency is above a certain threshold. This type of model has been advocated by Alegre and Gordon (1999) who did not find whole-word frequency effects for regularly inflected English words with a frequency below 6 per million while finding frequency effects throughout the frequency range for monomorphemic controls. If, like regularly inflected words in Alegre and Gordon's model, phrases are stored in the lexicon only if they are frequent enough and, other things being equal, predictability improves detectability, we should find facilitatory effects of predictability in phrases whose frequencies are insufficient for the phrase to become a stored prefab. One version of the theory is depicted in Figure 1.

However, a U-shaped relationship between phrase frequency and word detectability is also expected in a model that assumes that the ease of detecting a word is a function of how easy it is to parse the word from the acoustic signal (parseability) and how surprising, and therefore salient, the occurrence of the word is.[3] If the more predictable a word, the easier it is to parse from the signal, words in high-frequency phrases should be easier to detect than words in low-frequency phrases. However, at the same time, the occurrence of a word is not surprising if it is predictable and thus is less likely to attract attention, which could in turn lead to lower detectability. If, as phrase frequency increases, parseability rises faster than salience falls and parseability reaches ceiling (i.e., *up* is always parsed out) before salience reaches floor (i.e., the occurrence of *up* is not paid any attention at all), a U-shaped relationship between phrase frequency and word detectability is expected. Before parseability reaches the ceiling, detectability increases with increases in phrase frequency. After the ceiling is reached, salience is the

[2] Cloze probabilities were used in the studies.
[3] This is a generalization of Corcoran's (1966) idea that predictable words are skipped over to the domain of auditory perception.

only factor influencing detectability, hence further increases in phrase frequency should decrease word detectability.

In order to distinguish between the two theories, we need to look at what happens when parseability is not at ceiling and when wholes at the low end of the frequency continuum are also likely to be stored. This can be accomplished by looking at stimuli in which the to-be-detected stimulus, /ʌp/, is not a word but instead occurs inside a word, e.g., *puppy*. In these cases, *up* is less likely to be parsed from the signal and parseability is not at ceiling (accuracy in *up* detection is not perfect). Hence, inhibitory effects of ultra-high-frequency should not be found for word-internal /ʌp/s if they are due to a parseability/salience tradeoff.

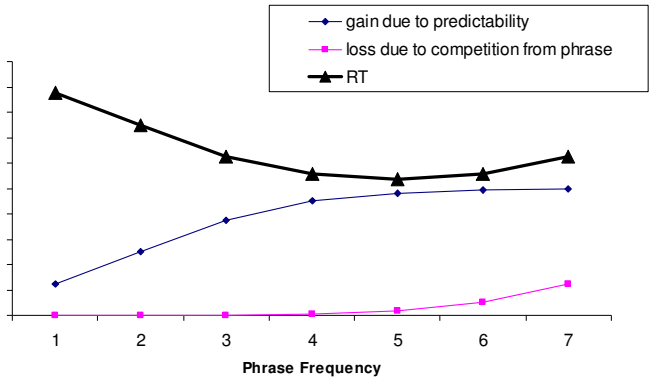


Figure 1. The theoretical relationship between phrase frequency and reaction time (RT) in detecting the second word in the phrase. Here $RT = A + loss - gain$ (predictability makes detection faster while competition from the prefab makes detection slower), where:

$$Gain = \frac{1}{1+B^{2-PhraseFrequency}} \text{ while } Loss = \frac{1}{1+B^{8-PhraseFrequency}}.$$ [4]

On the other hand, if the decrease in parseability of the parts is due to increased competition from the whole, /ʌp/ should be harder to detect in high-frequency words than in low-frequency words. Furthermore, since all words we examine are likely to be stored in the lexicon, there should be a negative correlation between /ʌp/ detectability and word frequency throughout the frequency range.

Methods

Materials

The verb + *up* collocations were chosen for inclusion in the experiment based on having non-zero frequency in the British National Corpus (determined through the online interface at http://view.byu.edu/). The British National Corpus was chosen because of its size and the availability of

[4] B and A are constants. The crucial feature is that the power to which B is raised is larger in the Loss formula than in the Gain formula. A processing interpretation of this mathematical formulation of the theory is that the word and the prefab are nodes with a sigmoid activation function. During recognition, the prefab and its parts compete for a limited amount of activation where the amount of activation received by a node is proportional to its resting activation level.

part-of-speech tagging. To find all verb+*up* constructions, we searched for the following pattern: [v*] up.[avp]. We obtained the frequencies of the verb+*up* collocations from the corpus.

The final sample of collocations used in the study was derived by keeping the 10 collocations closest to each end of the frequency continuum and randomly sampling the remaining collocations. In addition, we took all verbs that occurred with the particle *out* in the corpus and included a sample of such verbs that did not occur with *up* in the corpus but did occur with it on Google (the least frequent of these was *eke up*, as in *Tokyo's Nikkei slipped 0.9% and the FTSE 100 in London eked up 0.1%.*) paired with *up* to create the ultra-low-frequency end of the frequency distribution where *up* is not very predictable.

Most of the verb-particle phrases were presented using the past tense form of the verb. For regular verbs, this ensured that *up* was preceded by /d/ or /t/ (sometimes a flap). This was done to ensure that the location of the vowel onset in *up* can be reliably measured and to minimize the influence of phonological context on detectability of *up*.

The first author created 240 experimental sentences containing the particle *up* and 240 control sentences that were identical to the experimental sentences except for containing a different particle. The sentences were presented to the second author, a native English speaker, in a randomized order. The second author read the sentences aloud, having a fixed amount of time (5 seconds) to produce each sentence.

Thirty-five of the control sentences contained the particle *out*. Since experimental and control sentences were syntactically identical, prosody was not a cue to whether *up* occurs in the sentence. In most sentences, *up* was located immediately after the verb. However, to ensure that the subjects process the entire sentence, there were control sentences in which *up* either followed the direct object (n=20, *He brought it up*) or was sentence-initial (n=10, e.g., *Up he goes*). A verb occurring in these control sentences also occurred in an experimental sentence. The control sentences containing *up* were paired with control sentences of the same syntactic structure that contained a different particle so that the number of sentences containing *up* was equal to the number of sentences not containing *up*. The control sentences in which *up* is not immediately after the verb are not included in the analyses presented in this paper because the frequency of verb+*up* combinations was determined only for the most frequent location of *up*, which is immediately after the verb. The subject of the sentence was almost always a pronoun to ensure lack of co-occurrence-based priming between the subject and the particle. Twenty sentences containing noun-phrase subjects occurred in both the experimental and the control set to increase variability in particle location. Previous research has suggested that the greater the variability in location of the to-be-detected unit, the greater the likelihood of obtaining context effects (Lively & Pisoni, 1990).

In addition to stimuli in which *up* is a particle, we included a set of sentences in which /ʌp/ was inside another word. These sentences increase variability in target location and allow us to examine how word frequency influences detectability of parts of the word. We can then compare the influence of word frequency to the influence of phrase frequency. The words used were found in the MRC Psycholinguistic Database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm, Coltheart, 1981). For the experimental sample, we excluded compounds (e.g., *buttercup*), verb-particle constructions, words in which /ʌp/ was followed by a stop (e.g., *interrupt*), and Internet terms, whose frequency would be elevated in Google counts relative to overall use (*pop-up*, *lookup*, *setup*). We did not exclude nouns and adjectives derived from verb-particle constructions (e.g., *holdup*). If a noun could be used in the plural, we created two sentences, one containing the noun in the plural and one containing it in the singular.

It was ensured that /ʌp/ was equally likely to occur word-finally (e.g., *holdup*, *cup*), word-medially (e.g., *puppy*, *hiccups*) and word-initially (e.g., *upholstery*, *upper*). Morphological and syllabic constituency of /ʌp/ was manipulated. For instance, /ʌp/ is a syllabic constituent (the rime) but not a morphological constituent in *cup* while it is a morphological constituent that crosses a syllable boundary in *upper*. There were 96 /ʌp/-containing words used in the experiment. Each sentence with an /ʌp/-containing word was paired with a control sentence in which the /ʌp/-containing word was replaced by a word containing /aʊt/. The /aʊt/-containing words were also found using the MRC Psycholinguistic Database using the same exclusion criteria as for /ʌp/-containing words.

Subjects and Procedure

Twenty adult native English speakers were recruited from among introductory psychology students. They participated to fulfill a course requirement. The subjects were asked to press the 'present' button as soon as they hear *up*, regardless of whether it is a separate word or is inside another word. If the sentence did not contain 'up', they needed to press the 'absent' button to go on to the next sentence. They were encouraged to respond as soon as they hear *up* without waiting until the end of the sentence. The experiment lasted approximately 25 minutes.

Measurement of Frequency and Duration

For the purposes of deriving frequency-detectability correlations, we obtained phrase frequency estimates from Google. We also obtained frequency estimates from the spoken portion of the British National Corpus (BNC) but, while a U-shaped phrase frequency- word detectability relationship was observed with both counts, the Google-based results exhibited both a larger facilitatory effect on the low-frequency end of the continuum and a larger inhibitory effect at the high-frequency end. Furthermore, the spoken portion of the BNC did not allow us to distinguish between many frequency classes at the low-frequency end of the continuum. Thus only Google results are reported in this paper.

The use of web-based frequency estimates of phrase frequency is supported by the results of Keller and Lapata (2003) who found that plausibility judgments for bigrams that are found only on the Web (and not in the BNC) are reliably predicted by Google frequencies, indicating that Google counts are capturing psychologically relevant variation on the low end of the phrase frequency continuum that the BNC counts are not. Furthermore, even for bigrams found both in the BNC and on Google, correlations with plausibility judgments were higher for web-based frequency counts than for corpus-based ones.

Both base and surface frequency estimates were derived. The surface frequency estimate is the frequency of the verb+*up* combination where the verb is in the particular inflected form used in the experiment. The base frequency estimate is the summed frequency of verb+*up* summed across all forms of the verb. The results did not differ depending on whether base or surface frequency estimates were used. In analyzing the effect of phrase frequency, the frequency continuum was split into seven bins based on natural discontinuities in our sample of frequencies, as shown in Figure 2.

To investigate the effect of phonological reduction on detectability, we measured the durations of each occurrence of *up* in the materials. We also measured the distance between *up* and the beginning of the sentence. All measurements were done in Praat. The release of the stop closure was taken as the end

of the particle. Following stops and fricatives, the beginning of the particle was determined by the beginning of the vowel formants on the spectrogram (since the preceding verb was almost always in the past tense, this was the usual case). When the vowel onset was not readily apparent on the spectrogram, we listened for cues to the identity of the vowel in the preceding speech signal. We took the onset of the vowel to be the latest point at which we could not yet detect cues to the identity of the upcoming vowel. In order to control for possible effects of phonological reduction and measurement error, we measured reaction time both from the onset and the offset of the particle.

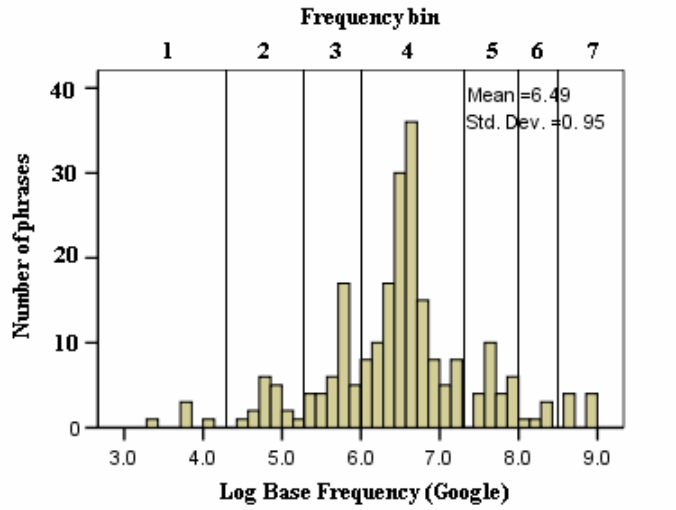


Figure 2. The frequency bins were derived based on discontinuities in the sample of frequencies.

Results

/ʌp/ as a Particle

Accuracy of particle detection in the present study was quite high (error rate is shown in Table 1). This contrasts with Sosa and MacFarlane (2002) where accuracy of *of* detection was at 47% in the lowest-frequency phrases, 60% in medium-low-frequency phrases, 38% in medium-high-frequency phrases, and 37% in the ultra-high-frequency phrases.

In the present study, accuracy in the lowest-frequency group was significantly lower than in any other group (with all other groups combined $p<.0005$, according to one-way ANOVA). Frequency bins 5 and 6 exhibit higher accuracy than either bin 7 ($p=.038$), or bins 2, 3, and 4 ($p=.005$). These results indicate that *up* is easier to detect when it is somewhat predictable than when it is unexpected (Dell & Newman, 1980; Morton & Long, 1976). The data suggest a U-shaped relationship with accuracy steadily increasing with phrase frequency but then dropping for the highest-frequency bin.

frequency bin	1 lowest	2	3	4	5	6	7 highest
error rate	20%	5%	6%	5%	3%	2%	6%

Table 1. Error rate in *up* detection depending on the frequency of the verb+*up* collocation.

Figure 3 presents reaction time (RT) data (correct trials only). As predicted by the hypothesis of between-level competition between prefabs and their component words, detection of *up* is more difficult in ultra-high-frequency verb+*up* collocations than in medium-frequency collocations. The difference in reaction time between frequency bin 7 (the highest-frequency bin containing the collocations *get up*, *sign up*, *go up*, and *set up*) and bin 6 (containing slightly less frequent collocations, including *keep up*, *line up*, *stand up*, *catch up*) is statistically significant according to a one-way ANOVA (for reaction time relative to particle onset, p=.005, for reaction time relative to particle offset, p=.002). Interaction with subject identity is not significant (p>.1). The significance of this effect is further confirmed by the fact that a quadratic function, which is U-shaped, provides a much better fit to the data than a monotonic, logarithmic one (the quadratic function explains 96% of the variance in reaction time as a function of phrase frequency while the logarithmic function explains 57% of the variance in reaction time measured relative to the onset and 46% of the variance in reaction time relative to the offset). The effect is observed regardless of whether we estimate phrase frequency via base frequency or surface frequency (for surface-frequency estimates, the difference between groups 7 and 6 is significant at p<.05, while the difference between groups 7 and 5 is significant at p=.002, interactions with subject identity are not significant, p>.2).

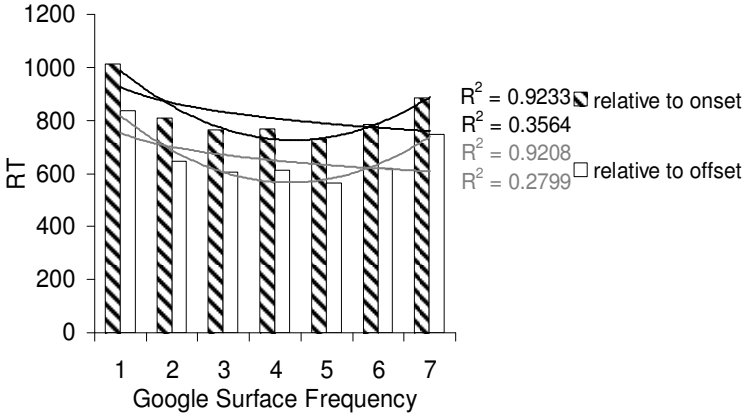


Figure 3. The U-shaped effect of the frequency of verb+*up* collocations on the speed with which *up* is detected.

The difference in fit almost disappears if frequency bin 7 is removed (the fit of the logarithmic function increases to 94-95% of the variance) indicating that throughout most of the frequency range, increased predictability helps to detect the particle. Just like in Sosa and MacFarlane (2002) and consistent with the accuracy results above, effects of phrase-word competition are only observed with extremely high-frequency phrases. Throughout most of the frequency continuum, *up* detection is easier in higher-frequency phrases than in lower-frequency ones, supporting the hypothesis that, other things being equal, predictability of the to-be-detected unit speeds up detection (Dell & Newman, 1980; Morton & Long, 1976).

In order to examine how consistent our results are with the results of Sosa and MacFarlane (2002), we examined where the collocations used in their study fit onto the frequency continuum derived from Google. We obtained a mean log frequency of 8.15 for their lowest-frequency group, 8.36 for the medium-low-frequency group, 8.77 for the medium-high-frequency group and 8.92 for the ultra-high-

frequency group. Thus, their lowest-frequency bin is similar in frequency to our bin 6 (mean log frequency = 8.22) while our group 7 is similar to their medium-high-frequency group (mean log frequency = 8.72). Thus, we find the inhibitory frequency effect at a similar (slightly lower) frequency level than Sosa and MacFarlane. The absence of facilitatory predictability effects in Sosa and MacFarlane's data is consistent with our findings: such effects are found much lower on the frequency continuum (between bin 1 with mean frequency of 3.74 and bin 5 with mean frequency of 7.72) than the range of frequencies used by Sosa and MacFarlane.

Importantly, the duration of the particle does not depend on phrase frequency. As can be seen in Figure 3, the difference between reaction time relative to particle onset and reaction time relative to particle offset is constant throughout the frequency range. Thus, the slow-down in detection observed in ultra-high-frequency phrases is not due to the presence of phonological reduction in those phrases. Thus, the findings of the present study support the hypothesis that phonological reduction is not a precondition for storage (Bybee 2001).

Word-internal /ʌp/

An alternative interpretation of the results in the previous section is a parseability-salience tradeoff: at some point on the phrase frequency continuum, *up* becomes so predictable that it is always parsed out of the signal. Above that point, further increases in phrase frequency can only decrease how surprising the occurrence of *up* is without increasing the likelihood of *up* being parsed out. To test this hypothesis, we turn to data from trials on which /ʌp/ occurs inside another word. In such cases, parseability of /ʌp/ should be decreased, thus /ʌp/ may be easier to detect in high-frequency words than in low-frequency words. On the other hand, since words are stored in the lexicon, the hypothesis of between-level competition predicts that /ʌp/ should be harder to detect in high-frequency words because such words are stronger competitors. A U-shaped function is not predicted because even the lowest-frequency words are expected to be stored in the lexicon.

Since word-internal occurrences of /ʌp/ are not all equal in terms of location within the word, length of the bearing word, morphological and syllabic constituency, stress, and, as it turns out, duration, we tested for effects of each of these variables. While stress and within-word location did not have a significant main effect, morphological and syllabic constituency, word length, and duration did.

Table 2 shows that /ʌp/ is easier to detect when it is a morpheme than when it is not ($p<.0005$ for both accuracy and reaction time). This result is consistent with Zwitserlood et al.'s (1993) findings for syllable monitoring in Dutch.

	Morpheme	Not morpheme
Accuracy	90%	72%
Reaction time	813	1023

Table 2. /ʌp/ is easier to detect when it is a morpheme than when it is not.[5]

[5] Reaction time for word-internal occurrences of /ʌp/ is relative to the onset of /ʌp/.

As shown in Table 3, accuracy of /ʌp/ detection is also affected by the length of the word in which /ʌp/ occurs: /ʌp/ is more likely to be missed in longer words than in shorter ones (p=.002 in a multinomial logistic regression that also included morphological constituency, syllabic constituency, and presence/absence of stress) especially if /ʌp/ is not a morpheme (the interaction is significant at p=.026). Table 3 shows that this is not a side effect of differences in duration of /ʌp/ within long and short words: while in general, longer instances of /ʌp/ are easier to detect (Table 6), instances of /ʌp/ that occur in longer words do not tend to be shorter than those occurring in short words (in fact, instances of /ʌp/ tend to be somewhat longer in longer words).

Length (segments)		3	4	5	6	7	8	10
% correct	Morpheme	N/A	95%	92%	90%	87%	86%	N/A
	Not morpheme	88%	76%	73%	58%	55%	N/A	55%
duration of /ʌp/ (ms)	Morpheme	N/A	93	94	99	102	116	N/A
	Not morpheme	74	64	84	134	112	N/A	47

Table 3. The effect of word length on accuracy of /ʌp/-detection (number of segments by percent correct).

The effect of word length is consistent with the hypothesis of between-level competition. There is a greater chance that not all parts of a word will be fully perceived prior to word identification in a long word than in a short word. Thus, processing of a part is more likely to be interrupted prior to completion in a long word than in a short word. If this hypothesis is correct, then, given that words are processed mostly left-to-right, the effect of word length should be most apparent in the word-final position, less apparent in the word-medial position and least apparent in the word-initial position. This is indeed the case in the data: the effect of word length is highly significant in the word-final position according to a one-way ANOVA (p<.0005 for non-morphemic and p=.008 for morphemic /ʌp/'s), marginally significant in the word-medial position (p=.087 for non-morphemic and p=.063 for morphemic /ʌp/'s), and not significant in the word-initial position (p=.172 for non-morphemic and p=.186 for morphemic /ʌp/'s).

Table 4 shows that detection of /ʌp/ is slower when /ʌp/ straddles a syllable boundary than when it does not (p<.0005). There was no difference between cases in which /ʌp/ is a syllable and when it is the rime (whether or not the rime was followed by an appendix). Syllabic constituency does not have a significant effect on accuracy, although the numerical trend is in the same direction as the effect on reaction times (87% correct when /ʌp/ is a syllabic constituent vs. 85% when it straddles a syllable boundary).

	Morpheme	Not a morpheme
Syllabic constituent	796	960
Not a syllabic constituent	964	1187

Table 4. The effects of morphological and syllabic constituency on the speed of /ʌp/ detection (ms).

The effect of syllabic constituency on sequence monitoring has been previously obtained by Mehler et al. (1981) for French, Bradley et al. (1993) for Spanish, and Zwitserlood et al. (1993) for Dutch. It has not previously been found in English (Cutler et al., 1986; Bradley et al., 1993). A possible reason for why previous studies have not found a syllabic constituency effect for English is that both Cutler et al. (1986) and Bradley et al. (1993) had subjects monitor for sonorant-final targets[6] whereas we used a stop-final target. A post-vocalic sonorant in English is more closely associated with the preceding vowel than an intervocalic stop is (Treiman & Danis, 1988; Derwing, 1992). Thus, previous syllable monitoring studies in English may not have included (many) targets that crossed a syllable boundary. This hypothesis is supported by the results of Ferrand et al. (1997) who failed to observe an effect of prime-target syllable structure consistency in masked priming in English when using Bradley et al.'s (1993) stimuli but were able to obtain it when stimuli with clear syllable boundaries were used.

The findings in Tables 2-4 indicate that /ʌp/ is more detectable when it is a constituent (whether morphological or phonological) than when it is not. These findings support a view of constituency as unithood: constituents are more likely to be parsed out of the signal than phoneme strings that straddle a constituent boundary. Especially in longer words, not all parts of the word are parsed out of the signal. Being a constituent makes a phoneme string more likely to be detected.

There is no interaction between morphological and syllabic constituency for either accuracy or reaction time ($p>.3$), indicating that being a syllabic constituent increases detectability even when /ʌp/ is a morphological constituent. Similarly, being a morpheme increases detectability of units that are syllables or rimes. This suggests that a morphological or syllabic constituent is not always parsed out of the signal. Rather, the fewer constituent boundaries that lie within a phoneme string, the more likely the string is to be parsed out.

However, before we conclude that constituency affects detectability, we need to address the fact that constituency of the particle correlates with particle duration in the stimuli, as shown in Table 5. Main effects of morphological and syllabic constituency are significant ($p<.0005$ in an ANOVA that included morphological constituency, syllabic constituency and word length as fixed factors and subject as random factor). There is no significant interaction.

	Morpheme	Not a morpheme
Syllabic constituent	100	86
Not a syllabic constituent	84	67

Table 5. The effect of constituency on duration of /ʌp/ (ms).

There is a significant correlation between /ʌp/ duration and how easy it is to detect. Shorter, more reduced, instances of /ʌp/ are detected more slowly (Pearson r = -.27, $p<.0005$).[7] Therefore, we conducted a linear regression analysis with logarithmically scaled reaction time as a dependent variable and syllabic constituency (1 vs. 0), morphological constituency (1 vs. 0), presence of stress on /ʌp/, /ʌp/ duration, word length (in segments), distance from sentence onset to /ʌp / onset, log word frequency, and

[6] Cutler et al. (1986) used /l/, Bradley et al. (1993) used mostly /l/ and nasals except for two stimuli containing /s/.
[7] We used $\log_{10}$(reaction time) for correlation analyses.

location of the stimulus in the list of sentences as independent variables. Both of the constituency variables were significant (t=-4.123, p=.001 for syllabic constituency, t=-3.227, p<.0005 for morphological constituency) as was duration of /ʌp/ (t=-4.206, p<.0005). These results suggest that constituency has an effect on detectability above and beyond duration.

In this analysis, the effect of word frequency only approached significance (p=.089, t=1.702). The direction of the trend was as predicted by the hypothesis of between-level competition: /ʌp/ was more difficult to detect in high-frequency words than in low-frequency words. However, we reasoned that the word frequency effect may not manifest itself when /ʌp/ occurs in the word-initial position but only when /ʌp/ occurs word-medially or word-finally. For instance, Lively and Pisoni (1990) observe a much stronger word frequency effect in phoneme categorization when the phoneme was in the final position than when it was in the initial position of a CVC word. In addition, we have observed earlier that the effect of word length on detectability of the word's parts is stronger for non-initial parts.

Thus, we broke the data down by where in the word /ʌp/ was located. Table 6 shows correlations between /ʌp/ duration, log frequency and logarithmically scaled reaction time depending on where in the word /ʌp/ is located. All correlations are significant (p<.001) except the one between word frequency and reaction time in the word-initial position, indicating that while word frequency does not appear to affect detection of word-initial targets, this is not simply because word-initial data is messier. The correlations between word frequency and speed of /ʌp/ detection are in the direction predicted by the between-level competition hypothesis: the higher the frequency of the word, the harder /ʌp/ is to detect when it occurs inside it.

	Initial	**Medial**	**Final**
Word frequency	.052	.285	.221
/ʌp/ duration	-.264	-.231	-.282

Table 6. Correlations (r) between independent variables and reaction time to /ʌp/ depending on the location of /ʌp/ within the word.

When word-initial instances of /ʌp/ are excluded from the regression analysis, word frequency is a significant predictor of reaction time (t=2.999, p=.003). Figure 4 shows that when a variety of functions is fit to the data, all of them display a monotonic relationship between word frequency and reaction time. Thus as word frequency increases, time taken to detect /ʌp/ inside the word rises throughout the frequency range. Unlike the effect of phrase frequency, the effect of word frequency is not U-shaped, as expected if all words we presented to subjects are stored in the lexicon, lexical units compete with their parts during recognition, and high-frequency lexical units are stronger competitors.

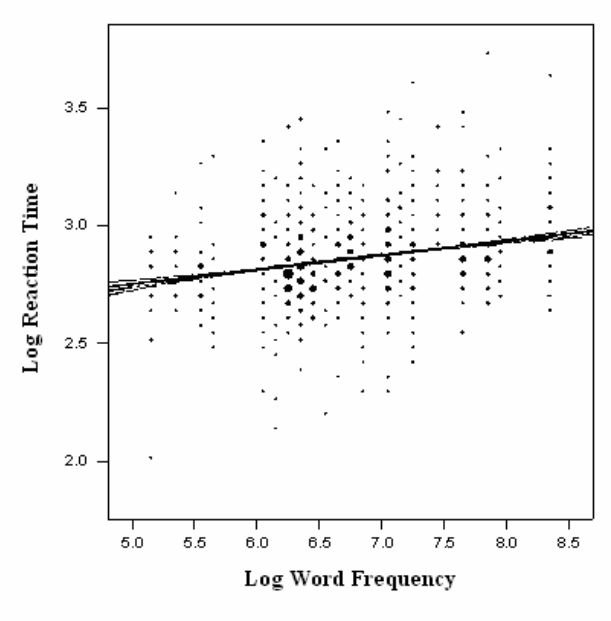


Figure 4. The monotonic relationship between word frequency and detectability of /ʌp/ within the word.[8]

Summary of the Results

In our experiment, /ʌp/ could either be a separate word or be inside another word. When *up* was a separate word, the frequency of the phrase comprising the preceding verb and /ʌp/ influenced detectability of /ʌp/ so that /ʌp/ was easier to detect in medium-frequency phrases than in high- or low-frequency phrases. On the other hand, when /ʌp/ was inside another word (and was not word-initial), detectability of /ʌp/ decreased as the frequency of the carrier word increased. In addition, non-word-initial /ʌp/ was more difficult to detect inside a long word than inside a short word. Regardless of whether /ʌp/ was word-initial, it was easier to detect when it was acoustically long than when it was short and when it was a morphological or syllabic constituent than when it was not.

Discussion

Theoretical Interpretation

The phoneme sequence /ʌp/ is more difficult to detect inside a high-frequency word than inside a low-frequency word. Thus, parts of frequent lexical units are less accessible to detection than parts of rare lexical units. Given this finding, we would predict that, if prefabs are lexical units, parts of frequent prefabs should be harder to detect than parts of rare prefabs. Finding an inverse relationship between frequency of a whole and detectability of its parts should indicate that at least the high-frequency wholes are stored in the lexicon. Such an inverse relationship is found for verb-particle phrases containing *up* but only at the very top of the phrase frequency continuum. These results are consistent with Sosa and MacFarlane's (2002) findings on word+*of* collocations. They indicate that the highest-frequency phrases are stored in memory as lexical units but they also **suggest** that a phrase needs to be extremely frequent to be stored in the lexicon. However, as Figure 1 shows, it is also possible that the activation level of the phrase begins to rise slowly as phrase frequency increases, and that until a certain point these frequency-dependent increases in the amount of competition the phrase generates are not enough to offset increases

[8] Circle size indicates number of data points. The trendlines shown are linear, quadratic, cubic and sigmoid.

in word predictability that are also caused by increases in phrase frequency. If that is the case, a more prudent conclusion is that the phrase representation does not participate in the lexical access process to a significant degree unless the phrase is extremely frequent.

Why are parts of high-frequency lexical units harder to detect than parts of less frequent lexical units? There must be some mechanism that would make activating the prefab interfere with bottom-up activation of the component words and activating a word interfere with bottom-up activation of the component morphemes, syllables, and bigrams. In other words, the results can only be explained if linguistic units in a part-whole relationship compete for activation during the perception process. This hypothesis is also supported by our finding that /ʌp/ is more likely to be missed in a long word, where recognition of /ʌp/ is less likely to be necessary for lexical access.

This idea can be implemented in several non-mutually-exclusive ways. Some possibilities include 1) competition for a limited supply of activation coming from either the acoustic signal or previously perceived context, 2) top-down inhibition, where wholes inhibit their parts when activated beyond a particular threshold (Libben, 2005: 276), or 3) removal of activation source at the completion of lexical access by ceasing to process the acoustic signal that has been parsed into lexical units (Healy, 1994).

It is at present unclear whether the competition process involves competition between lexical units only (within-level competition) or between both lexical and sublexical units. In order to establish whether lexical units compete with their sublexical parts, the present study will need to be replicated with a detection target that is not a (possible) word, e.g., a segment.

Finally, we observe that /ʌp/ is easier to detect when it is a constituent than when it is not a constituent. This finding suggests that the acoustic signal is parsed into morphemes and syllables during speech perception making /ʌp/ easier to detect when it matches one of the units automatically extracted from the signal and more difficult to detect when the component segments of /ʌp/ need to be matched to segments that occur in different, though adjacent, units.

The Facilitatory Effect of Word Frequency on Phoneme Monitoring in Word Lists

In the present study, we observed that sequence detection is easier in low-frequency words than in high-frequency words. This is consistent with letter-detection results observed by Healy (1976) and Minkoff and Raney (2000). However, a word frequency effect in the opposite direction is often observed in phoneme monitoring (Cutler et al., 1987; Eimas et al., 1990; Lively & Pisoni, 1990; and Rubin et al., 1976) and letter monitoring (Howes & Solomon, 1951; Johnston, 1978) where phonemes and letters in high-frequency words are easier to detect than those in low-frequency words.

There is a systematic difference between experiments that find a word-frequency advantage in letter or phoneme detection and those that find a disadvantage: the word-frequency advantage is found with single-word presentation while multi-word presentation yields a word-frequency disadvantage (Hadley & Healy, 1991; Healy et al., 1987).[9]

[9] In the case of Eimas et al. (1990) the words were presented in a sentence context but the sentence context was constant (*the next word is…*) and the target word was always the last word in the sentence.

Healy et al. (1987) explain the difference between single-word and multi-word presentation using the Unitization Hypothesis. According to the hypothesis, readers move on to the next word in text as soon as they have identified the current word, terminating processing of smaller units within the current word. When only a single word is visible, there is no subsequent word, hence the subjects will continue processing the word they have already identified, at which point determining the identity of individual letters will be facilitated by having identified the word because the reader will be able to use his/her knowledge of what the word is to infer whether the target letter has been presented.

This explanation predicts that the word-frequency disadvantage should not be observed when the target word is in the sentence-final position. Our data are consistent with this prediction: there is no significant correlation between log word frequency and log reaction time for words in the sentence-final position even if only words in which /ʌp/ is not word-initial are included (r=.047, p=.569). However, this subset of words is small (12 words), so the reliability of this result is questionable.

It is also possible that the longer the presented acoustic or visual signal, the larger the units to which most attention is paid (Wray, 2002: 271). If this is the case, more attention will be paid to words relative to segments when sentences are presented than when isolated words are presented. As a result, word processing will be more likely to be completed prior to completion of segment or syllable processing and thus interfere with it in connected speech than with isolated words.

Finally, studies that found facilitatory effects of carrier word frequency on target detectability have used targets that are not possible words. Thus, the results of those studies can be reconciled with the results of our study by the theory that spoken word recognition involves within-level but not between-level competition.

Conclusion

Listeners find it more difficult to detect /ʌp/ in a high-frequency lexical unit than in a low-frequency one or, more concisely, **the stronger the whole the weaker the parts** (Bybee & Brewer, 1980; Hay, 2003; Healy, 1976; Sosa & MacFarlane, 2002). While all words are lexical units, leading to a monotonic relationship between word frequency and difficulty of /ʌp/ detection, our results suggest that only high-frequency phrases are stored in the lexicon. Since, other things being equal, predictable units are easier to detect, there is a U-shaped relationship between the frequency of the verb-particle collocation and detectability of the particle. For collocations that are not stored in the lexicon as units, the more probable the particle, the easier it is to detect due to a strong association between the particle and the co-occurring verb. For phrases that are stored in the lexicon, the more frequent the phrase, the more it interferes with the detection of the particle. Finally, /ʌp/ is easier to detect when it matches a morphological or syllabic constituent than when the segments of /ʌp/ are separated by a morpheme or syllable boundary, providing evidence for the hypothesis that syllables and morphemes are extracted from the acoustic signal and take part in the part-whole competition operating during lexical access.

References

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40, 41–61.

Bod, R. (2001). Sentence memory: The storage vs. computation of frequent sentences. Paper presented at the CUNY Sentence Processing Conference. Philadelphia, PA.

Bradley, D. C., Sánchez-Casas, R. M., & García-Albea, J. E. (1993). The status of the syllable in the perception of Spanish and English. *Language and Cognitive Processes*, 8, 197-233.

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón and B. F. Malle (eds.), *The evolution of language out of pre-language*, pp.109-32. Amsterdam: John Benjamins.

Bybee, J. (2001). *Phonology and language use*. Camdridge, UK: Cambridge University Press.

Bybee, J., & Brewer, M. A. (1980). Explanation in morphophonemics: Changes in Provençal and Spanish preterite forms. *Lingua*, 52, 201-42.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Corcoran, D. W. J. (1966). An acoustic factor in letter cancellation. *Nature*, 210, 658.

Cutler, A., Mehler, J., Norris, D. G., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, 19, 141-77.

Cutler, A., Mehler, J., Norris, D. G., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385-400.

Dell, G. S., & Newman, J. E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, 19, 607-23.

Derwing, B.L. (1992). A 'pause-break' task for eliciting syllable boundary judgments from literate and illiterate speakers: Preliminary results for five diverse languages. *Language and Speech*, 35, 219-35.

Eimas, P. D., Marcovitz-Hornstein, S. B., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language*, 29, 160-80.

Ferrand, L., Segui, J., & Humphreys, G. W. (1997). The syllable's role in word naming. *Memory and Cognition*, 25, 458-70.

Hadley, J.A., & Healy, A.F. (1991). When are reading units larger than the letter? Refinement of the unitization reading model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 1062-73.

Hay, J. (2003). *Causes and consequences of word structure*. London: Routledge.

Healy, A.F. (1994). Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin and Review*, 1, 333-44.

Healy, A.F. (1976). Detection errors on the word *the*: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 235-42.

Healy, A.F., Oliver, W. L., & MacNamara, T.P. (1987). Detecting letters in continuous text: Effects of display size. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 413-26.

Howes, D.H., & Solomon, R.L. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41, 401-10.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics*, 40, 431-9.

Johnston, J.C. (1978). A test of the sophisticated guessing theory of word perception. *Cognitive Psychology*, 10, 123-53.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29, 459-84.

Libben, G. (2005). Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics*, 50, 267-83.

Lively, S.E., & Pisoni, D.B. (1990). Some lexical effects in phoneme categorization: A first report. In *Research on Speech Perception Progress Report No. 16* (pp. 327-59). Bloomington, IN: Speech Research Laboratory, Indiana University.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32, 692-715.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, 88, 375-407.

McDonald, S. A., & Shillcock, R. C. (2004). Eye-movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14, 648–52.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20, 298-305.

Minkoff, S. R. B., & Raney, G. E. (2000). Letter-detection errors in the word *the*: Word frequency versus syntactic structure. *Scientific Studies of Reading*, 4, 55-76.

Morton, J., & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, 15, 43-51.

Reali, F., & Christiansen, M. H. (2007). Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology*, 60, 161-70.

Rubin, P., Turvey, M. T., & vanGelder, P. (1976). Initial phonemes are detected faster in words than in non-words. *Perception and Psychophysics*, 19, 394-8.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102, 11629-34.

Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language*, 83, 227-36.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, 27, 87-104.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (ed.), *Formulaic sequences*, pp.153-72. Amsterdam: John Benjamins.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.

Zwitserlood, P., Schriefers, H., Lahiri, A., & vanDonselaar, W. (1993). The role of syllables in the perception of spoken Dutch. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1-12.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Inter-Talker Differences in Intelligibility for Two Types of Degraded Speech

Tessa Bent, Adam Buchwald and Wesley Alford[1]

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by the National Institutes of Health to Indiana University (NIH-NIDCD T32 Grant DC-00012). We thank Vidhi Sanghavi and Jennifer Karpicke for their assistance in data collection, Luis Hernandez for technical assistance and Ann Bradlow for allowing us access to her data.

Inter-Talker Differences in Intelligibility for Two Types of Degraded Speech

Abstract. Are the acoustic-phonetic factors that promote highly intelligible speech invariant across different listener populations and listening environments? Researchers have taken two approaches to investigate differences in intelligibility for a variety of listener populations: examining how speaking style affects intelligibility, and examining how inter-talker differences influence intelligibility. Following the latter approach, we compared the intelligibility of talkers under cochlear implant (CI) simulation (n=200), and in speech mixed with babble (n=200) with their intelligibility under quiet listening conditions (n=200, reported by Karl & Pisoni, 1994). The stimuli consisted of 20 native English talkers producing 100 sentences which were processed to simulate listening with an 8-channel CI or mixed with multi-talker babble. For each condition, stimuli were presented to listeners in a sentence transcription task. The results indicated that the most intelligible talkers in quiet were not the most intelligible talkers under CI-simulation or in babble. Furthermore, listeners demonstrated a greater degree of perceptual learning with the CI-simulated speech compared with the speech mixed with babble. While some of the acoustic-phonetic properties were correlated with intelligibility in all conditions, other properties differed in their degree of correlation among the three conditions. Overall, these results suggest that the acoustic-phonetic parameters that result in highly intelligible speech are dependent on listener characteristics and listening environment.

Introduction

What factors determine speech intelligibility?[2] Traditional views of speech intelligibility hold that intelligibility is a property of the specific words being perceived or of the talker whose speech is being perceived. There is empirical support for each of these views. For example, certain properties of words (e.g., segmental composition; length; frequency) have been shown to influence intelligibility (Black, 1957; Howes, 1952, 1957). Similarly, it has been shown that various properties of a talker's articulations (e.g., speaking rate; vowel dispersion) are essential in determining speech intelligibility (Bond & Moore, 1994; Bradlow, Toretta & Pisoni, 1996; Hood & Poole, 1980). However, recent studies indicate that the speech materials and the talker are not the only relevant factors in determining speech intelligibility. Instead, a variety of research findings suggest speech intelligibility is influenced by properties of the listener, listening environment, linguistic context as well as interactions among these factors. In this paper, we present experimental results indicating that inter-talker variation in speech intelligibility differs for different listeners and in different listening environments.

A number of studies have shown that the intelligibility of talkers varies even under ideal listening conditions (Bond & Moore, 1994; Bradlow et al., 1996; Hazan & Markham, 2004; Hood & Poole, 1980). One aim of this work has been to determine which acoustic-phonetic features correlate with intelligibility, as this may allow us to improve the intelligibility of speech for certain special populations who have particular difficulty in speech perception (e.g., hearing-impaired listeners; second language users). The results obtained in these studies have yielded discrepancies regarding the acoustic-phonetic parameters that are most important for highly intelligible speech. The acoustic-phonetic features that have been reported to correlate with speech intelligibility include increased vowel and word durations (Bond & Moore, 1994; Hazan & Markham, 2004), expanded vowel space (Bond & Moore, 1994; Bradlow et al.,

[2] Speech intelligibility is defined here as the listener's ability to accurately report the words that a talker has produced. This objective measure of speech intelligibility contrasts with other measures in which listeners subjectively rate the "intelligibility" of a speaker (also called comprehensibility) or tests in which the listener must provide an accurate paraphrase of the talker's message in order for the talker's communicative intent to be considered effective.

1996), more pauses (Bond & Moore, 1994), increased F0 range (Bradlow et al., 1996), and more energy in the 1 – 3 kHz region (Hazan & Markham, 2004). Furthermore, talker gender seems to be an important variable for intelligibility. Both Bradlow et al. (1996) and Hazan and Markham (2004) found that female talkers are significantly more intelligible than male talkers. Bond and Moore (1994) did not assess this variable as only male speakers were used. An additional concern with these studies is that the listener populations that are examined tend to be normal-hearing native language listeners (cf. Bond & Moore, 1994). If the ultimate goal of these studies is to improve intelligibility for listeners from special populations then it is important to determine which talkers are most intelligible for these listener populations, and which acoustic-phonetic parameters are important for enhancing intelligibility for the particular listener population. This goal motivates the use of cochlear implant (CI) simulated speech in the present study.

Previous research has also revealed that listener properties help determine which talkers are most intelligible, and thus that different acoustic-phonetic parameters may promote intelligibility for different listener populations. For example, several studies have demonstrated that a shared dialect between the talker and listener may facilitate intelligibility, whereas a mismatch of dialects between the talker and listener may hinder communication (Labov & Ash, 1997; Mason, 1946; cf. Clopper & Bradlow, in press). Similarly, a match or mismatch between talker and listener with respect to nativeness may also affect intelligibility; while native talkers tend to be more intelligible than non-native talkers for native listeners, non-native talkers can be equally intelligible as native talkers for non-native listeners (Bent & Bradlow, 2003; Imai et al., 2003; van Wijngaarden, 2001; van Wijngaarden et al., 2002). In contrast to these findings, Green, Katiri, Faulkner, and Rosen (2007) reported no differences in talker intelligibility among three groups of listeners, which included normal-hearing listeners and actual and simulated CI listeners. However, the lack of a difference in Green et al.'s work may have been an artifact of the small number of talkers used in their study, as discussed more below. Thus, the present investigation examines this issue with a large number of talkers, to determine whether the same relative differences in talker intelligibility are observed under normal listening conditions and degraded listening conditions.

It is also well-known that a listener's experience with a particular talker's idiolect also influences the talker's intelligibility. For example, as listeners become more familiar with the particular acoustic-phonetic properties of a talker's voice, their word recognition skills for that talker will be more accurate (Nygaard, Sommers & Pisoni, 1994). This effect of experience can also be talker-independent, as a beneficial effect of experience on speech intelligibility has been shown for listeners with extensive experience listening to foreign accented speech (Bradlow & Bent, in press; Clarke & Garrett, 2004; Weil, 2001), speech produced by talkers with hearing impairments (McGarr, 1983), computer manipulated speech (Dupoux & Green, 1997; Greenspan, Nusbaum & Pisoni, 1988; Pallier et al, 1998; Schwab, Nusbaum & Pisoni, 1985), and noise-vocoded speech (Davis et al., 2005). Critically, this benefit has been reported to extend to new talkers, and to new speech signals created using the same types of signal degradation. We address this type of perceptual attunement in the present work by comparing performance on an initial group of sentences in a novel listening condition to performance after the listener has been exposed to the condition for many sentences. The results of previous studies would suggest that we will obtain significantly better performance after exposure to a novel listening condition.

Although this review has highlighted how a listener's language background and prior experience may influence inter-talker differences in speech intelligibility, other studies suggest that listener properties are largely unimportant for determining intelligibility when compared to talker characteristics. For example, Hazan and Markham (2004) reported that intelligibility differences between male and female adult and child talkers were the same for listeners of all ages. Similarly, Bond and Moore (1994) reported that intelligibility rankings among several native talkers were the same for native and non-native listeners, suggesting that – at least for native talkers – the language background of the listener is less important than talker-based characteristics. Further, several studies of intelligibility among native and

non-native talkers and listeners have found that native and non-native talkers demonstrate the same relative intelligibility for native and non-native listeners (Major et al., 2002; Munro, Derwing & Morton, 2006), and that certain cue enhancement strategies (i.e., amplification of regions of the speech signal that are thought to carry more information) enhance intelligibility for both native and non-native listeners (Hazan & Simpson, 2000).

In a study that examined similar populations to the present work, Green et al. (2007) argued that a listener's hearing status is also relatively unimportant in determining relative intelligibility among talkers. They presented words from six talkers to CI users and normal-hearing listeners. Normal-hearing listeners heard the speech either mixed with babble at a very favorable signal to noise ratio or under cochlear implant simulation. The stimuli were from two adult male, two adult female and two child female talkers. In each group, one talker was characterized as a high intelligibility talker and one was characterized as a low intelligibility talker based on results from Hazan and Markham (2004). Green et al. reported that intelligibility was relatively consistent across listeners and degradation types, which suggests that at least some talker characteristics are beneficial across listener populations and listening conditions. However, the small number of talkers included in the study (six total), the use of word length stimuli (except for sentences by the adult male talkers), and the choice of talkers at the extremes of the intelligibility distribution limits the extent of generalization of these results. The current study addresses these limitations by using a larger number of talkers, sentence length stimuli and talkers with a wide range of intelligibility scores.

In addition to properties of the talker and the listener, the listening environment also contributes to intelligibility. Overall, speech in noise is less intelligible than speech in quiet (See Assmann & Summerfield, 2004, for a review). However, different types of noise affect speech intelligibility differently, both in overall intelligibility as well as determining what aspects of the signal are difficult to identify. For example, low frequency noise tends to reduce the intelligibility of speech more than high frequency noise (Miller, 1947), and broadband noise tends to impair listeners' abilities to identify place of articulation more than other consonant features (Miller & Nicely, 1955). Furthermore, some listeners are more affected by noise than others; bilinguals or second language users may perform similarly to native listeners on speech identification tasks in the quiet, but their performance decreases more than natives in the presence of noise (Mayo, Florentine & Buus, 1997; Meador, Flege & Mackay, 2000; Nabelek & Donahue, 1984; Rogers, Listeer, Febo, Besing & Abrams, 2006; Takata & Nabelek, 1990). Likewise, those with hearing loss may show relatively unimpaired speech perception performance under quiet listening conditions, but will have much more difficulty in the presence of background noise (e.g., Moore, 2003; Nabelek, 1988). Results from these studies demonstrate a clear interaction of listener characteristics and listening environment in determining intelligibility. Whether differences across talkers are maintained in different listening environments is an issue that has not been extensively studied. In one of the few extant studies, Cox, Alexander and Gilmore (1987) found that relative intelligibility rankings among six talkers were generally maintained across four levels of noise degradation (speech mixed with babble). Their results suggest that the same talkers may be least and most intelligible across listening environments, but the types of degradation studied were similar, with differences only in signal-to-noise ratio and reverberation characteristics. Therefore, the present study compares intelligibility of speech mixed with babble with the intelligibility of CI simulated speech and speech in quiet listening conditions, to determine whether relative intelligibility among talkers will change more extensively for these different types of degradation.

The Present Study

In this paper, we report on an investigation of how talker characteristics interact with listener characteristics and listening environment to determine speech intelligibility. The central aim of this experiment is to determine whether and how inter-talker differences in intelligibility change depending on

listener characteristics (e.g., status as a simulated cochlear implant listener) and listening environment (quiet environment versus noisy environment). Understanding how the interaction of talker and listener characteristics and listening environment influences intelligibility is an important goal in characterizing the factors that contribute to speech intelligibility. While many experimental paradigms and clinical tests use only one talker, it remains largely unknown whether talker specific acoustic-phonetic features that are beneficial to one listener population are beneficial to all listener populations. Investigating the responses of listeners from special populations will further contribute to our knowledge about how listener-related variables can interact with inter-talker differences in intelligibility.

In addition to providing a richer characterization of the factors that contribute to speech intelligibility, this research may have practical applications for people with cochlear implants. Specifically, identifying speech features which are beneficial to these listeners can help guide talkers in improving their intelligibility when communicating with a person with a cochlear implant. Additionally, the results may have applications for the selection of talkers used in clinical tests.

In the current experiment, intelligibility scores for 10 male and 10 female talkers were compared across three listening conditions: Quiet, CI simulation, and Babble. Listeners were presented with speech from only one talker in one listening condition. Six hundred listeners were tested in total: two-hundred listeners for each listening condition. Intelligibility scores were compared across listening conditions and the extent of adaptation to the speech across the course of the experiment was assessed. Lastly, acoustic-phonetic correlates of intelligibility for the two degradation conditions were identified.

Method

Stimuli

The sentences from the Indiana Multi-talker Sentence Database were used. This database includes recordings of 100 Harvard sentences (IEEE, 1969) produced by 20 talkers (10 male and 10 female), with a total of 2000 sentences. Sentences included in this database are shown in Appendix A. The sentences were processed in two ways to assess the intelligibility of these sentences for the simulated listener population as well as when mixed with noise.

CI Simulation. For the CI simulation condition, each sentence was processed through an 8-channel sinewave vocoder using the cochlear implant simulator TigerCIS (*http://www.tigerspeech.com/*). The 8-channel simulation was chosen because normal-hearing listeners perform similarly to CI-users when listening to 8-channel simulations compared to greater or fewer numbers of channels (Dorman et al., 1997). Furthermore, a sine-wave vocoder was employed rather than noise-band vocoder for the same reason. Additionally, when single electrodes are stimulated in CI-users they subjectively report that they hear a sound more like a pure tone than noise (Dorman et al., 1997).

Babble. For the babble condition, the original sentences were mixed with 6-talker babble at a signal to noise ratio of 0. This signal-to-noise ratio was chosen based on pilot data in which the intelligibility of the sentences mixed with babble was matched with intelligibility of the 8-channel CI-simulated sentences. The speech in this condition was not vocoded.

Participants

Four hundred normal-hearing listeners participated (268 females and 132 males with an average age of 21.4 years). All listeners were native speakers of English and reported no current speech or hearing impairments. Listeners were either paid $5.00 for their participation or received course credit in an

introductory psychology course. Participants were undergraduate students at Indiana University or members of the greater Bloomington community.

Task

In each condition, a talker's intelligibility was assessed by examining the performance of 10 normal-hearing listeners on a sentence transcription task (20 talkers x 2 degradation conditions x 10 listeners = 400 listeners total). Each listener was presented with speech from one condition (i.e. quiet, CI simulation or Babble) and heard only one talker during the course of the experiment. During testing, each participant wore Beyer Dynamic DT-100 headphones while sitting in front of a Power Mac G4. Each sentence was played over the headphones followed by a dialogue box presented on the screen which prompted the listener to type in what he or she heard. Each sentence was presented once in a randomized order, and the experiment was self-paced so participants could take as long as needed to enter a response. Listeners were not provided with feedback as to the accuracy of their responses. Prior to the first experimental trial, participants were familiarized with the type of degradation by hearing two familiar nursery rhymes ("Jack and Jill" and "Star Light, Star Bright") which had been processed in the same manner as the sentences in their experimental condition. During familiarization, listeners were not required to make any responses.

Scoring

The responses were scored based on number of keywords and sentences correct. Each sentence has five keywords (underlined words in Appendix A). Keywords were only counted as correct if all and only the correct morphemes were present. Therefore, words with added or deleted morphemes were counted as incorrect. Obvious misspellings and homophones were counted as correct. A sentence was counted as correct if all five keywords were correctly transcribed.

Results

The results will be presented separately for the CI simulation condition and the babble condition. Each of these sections contains several critical comparisons. First, the data in the experimental conditions reported here are compared with intelligibility scores from these same talkers under quiet listening conditions (reported in Karl & Pisoni, 1994). Second, male speakers are directly compared with the female speakers in terms of intelligibility; this was shown to be a significant predictor of intelligibility under quiet listening conditions (results reported in Bradlow et al., 1996). Third, we examined the rate of perceptual attunement under each experimental condition by comparing performance on the first 20 sentences with performance on the last 20 sentences. Each talker's proportion improvement was then compared to their overall intelligibility under quiet listening conditions. The final section of the results considers findings from the CI simulation and babble conditions with respect to a variety of acoustic-phonetic parameters.

Intelligibility of Cochlear-Implant Simulated Speech

Four subjects' data were removed as they were determined to be outliers (their keyword correct score was at least three standard deviations below the mean for that talker). Their data was replaced by data from four new listeners. The data reported below are *keywords correct* except when noted, as this is a more fine-grained measure of intelligibility than *sentences correct*.

Comparison to Intelligibility in Quiet. The intelligibility scores for each talker in the CI simulation condition were computed and compared to intelligibility scores in the quiet (gathered by Karl & Pisoni, 1994). This comparison is shown in Figure 1. For this initial analysis, *sentence* intelligibility

was considered rather than *keyword* intelligibility as Karl and Pisoni only reported sentence intelligibility scores (due to a lack of variation in keyword correct scores). Overall, intelligibility scores in Quiet were not significantly correlated with CI Simulated intelligibility (r=0.347, *ns*). As can be seen in the figure in which talkers are arranged from least to most intelligible in Quiet, talkers who were most and least intelligible in quiet listening conditions were not necessarily the talkers who were most and least intelligible under CI-simulated listening conditions.

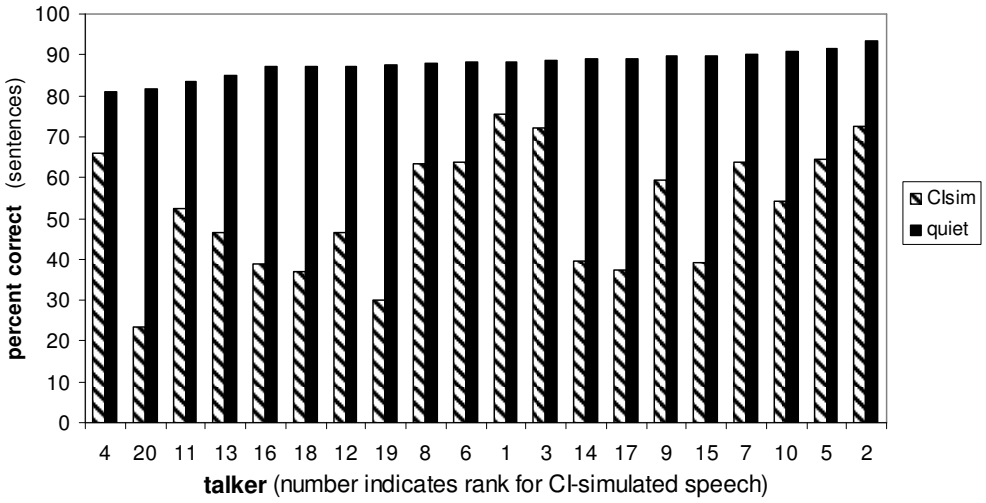


Figure 1: Comparison of intelligibility scores in quiet and under CI-simulated listening conditions. Talkers are ordered on the x-axis by their intelligibility in quiet. The intelligibility of talkers in the quiet and under CI-simulation was not correlated.

Gender Differences. The data from the CI simulation condition revealed that female talkers are more intelligible than their male counterparts. Using keywords correct as the dependent variable, female talkers (mean = 84%, SD = 11) were significantly more intelligible than male talkers (mean = 77%, SD = 11; t(198)=4.61, p<0.001). This is consistent with the findings of a gender difference in speech intelligibility in the Quiet condition.

Perceptual Attunement. In addition to overall intelligibility, the adaptation to the CI-simulated speech was assessed by examining improvement from the first 20 sentences to the last 20 sentences, a measure of perceptual attunement. This analysis was conducted using keywords correct as the dependent variable, and revealed rapid adaptation, with significantly more keywords correct in the last 20 sentences (mean = 84%, SD = 11) than in the first 20 sentences (mean = 73%, SD = 15; t(199)=16.6, p<0.001). Thus, listeners rapidly adapted to the CI simulated speech from all talkers without explicit feedback. Additionally, we found a great deal of variation in the extent of adaptation across talkers, with proportion improvement ranging from 0.21 to 0.56. These data are shown in Figure 2, sorted by the Karl and Pisoni (1994) measure of intelligibility in quiet. The rank-ordered correlation[3] between attunement scores and intelligibility in quiet was not significant (rho = 0.023, *n.s.*) indicating that the talkers with the greatest attunement were not necessarily the talkers with the highest intelligibility scores in quiet.

[3] We used a rank-ordered correlation because the two dependent variables are on different scales, with the Karl and Pisoni (1994) data measured in sentences correct and percent attunement in the CI simulation condition based on keywords correct.

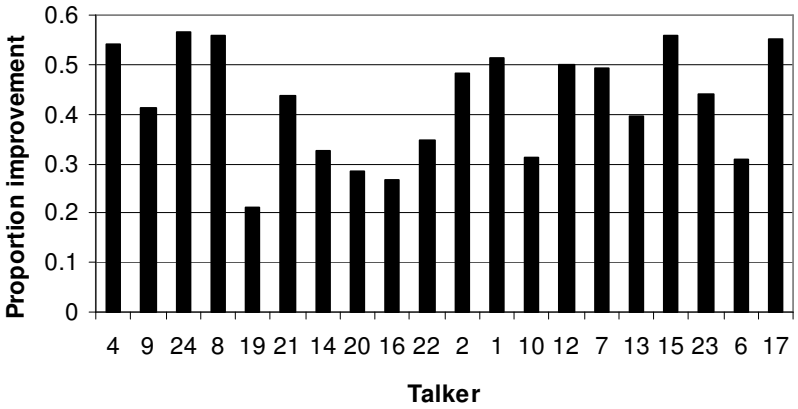


Figure 2: Proportion improvement from first 20 sentences to final 20 sentences for CI-simulated listening conditions. Talkers are ordered on the x-axis by their intelligibility in quiet. While listeners adapted to the speech from all talkers, the extent of adaptation depended on the particular talker.

Intelligibility of Speech Mixed with Multi-Talker Babble

Comparison to Quiet and CI Simulation. The sentence intelligibility scores in babble were compared with the sentence intelligibility scores under quiet listening conditions (from Karl & Pisoni, 1994). As with the CI-simulated intelligibility scores, the intelligibility scores in the babble condition were not correlated with the scores from the quiet listening condition (r=0.36, *ns*). This result indicates that talkers who were highly intelligible in quiet were not necessarily highly intelligible under noisy listening conditions. Comparisons of individual talker scores in the babble condition and in quiet are shown in Figure 3, sorted by intelligibility in quiet.

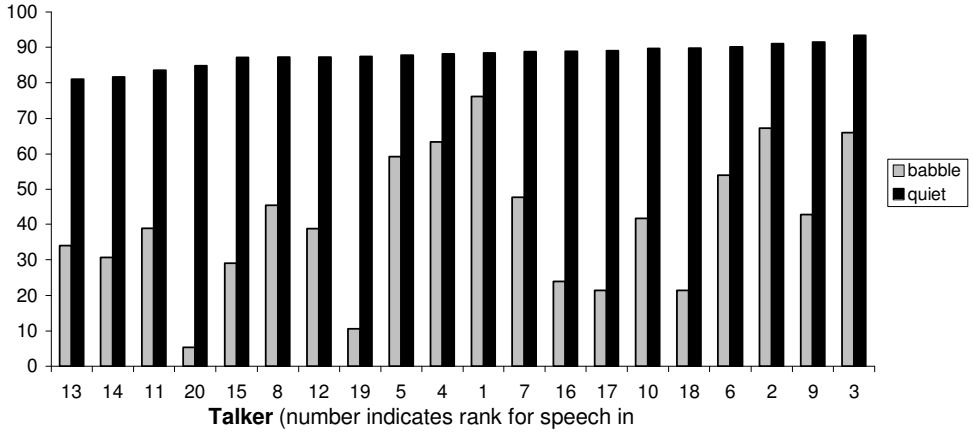


Figure 3: Comparison of sentence intelligibility scores in quiet and in noisy listening conditions (i.e. speech mixed with multi-talker babble). Talkers are ordered on the x-axis by their intelligibility in quiet. Scores in the quiet and babble conditions are not significantly correlated.

We also compared the intelligibility scores from the two experimental conditions: CI Simulation and Babble. This analysis was conducted using keyword accuracy as the dependent variable, as this finer grained measure is more appropriate; although ceiling effects were observed with keyword accuracy under quiet listening conditions, keyword intelligibility scores were not at ceiling in either of the

degradation conditions. The keyword accuracy scores for the CI-simulated condition and the babble condition were significantly correlated (r=0.73, $p < 0.001$). Therefore, while intelligibility under quiet listening conditions was not significantly correlated with intelligibility in either of the two experimental conditions, the CI simulation intelligibility scores were significantly correlated with the Babble intelligibility scores, suggesting that acoustic-phonetic parameters that promote intelligibility under one type of degradation may also promote intelligibility with the other type of degradation. Comparisons of keyword intelligibility scores in the two degradation conditions are shown in Figure 4.

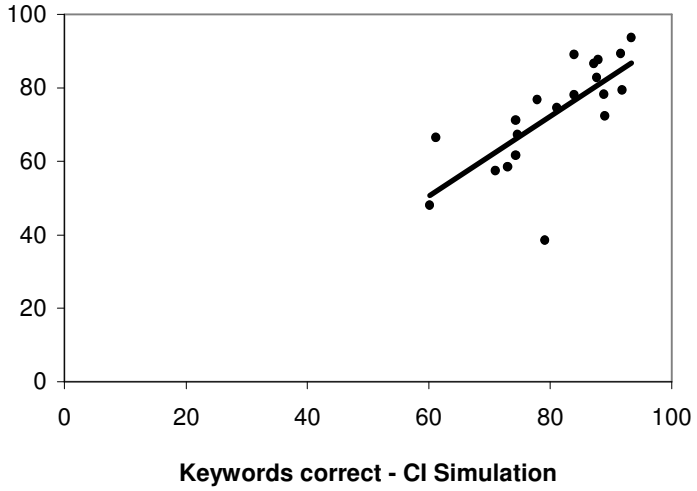


Figure 4: Comparison of keyword intelligibility for the two degradation conditions, CI-simulated speech and speech mixed with babble. Intelligibility scores under these two conditions were significantly correlated.

Gender Differences. The data from the Babble condition revealed that female talkers were more intelligible than their male counterparts in this listening condition. Using keywords correct as the dependent variable, female talkers (mean = 81%, SD = 14) were significantly more intelligible than male talkers (mean = 65%, SD = 13; $t(198)=8.47$, $p < 0.001$). This is consistent with the findings of a gender difference in speech intelligibility in the Quiet and CI Simulation conditions.

Perceptual Attunement. As with the CI simulation condition, perceptual adaptation to the speech in the Babble condition was assessed by examining improvement from the first 20 sentences to the last 20 sentences, a measure of perceptual attunement. This analysis was conducted using keywords correct as the dependent variable, and revealed rapid adaptation, with significantly more keywords correct in the last 20 sentences (mean = 75%, SD = 13) than in the first 20 sentences (mean = 69%, SD = 16; $t(19)=6.45$, $p<0.001$). Overall, listeners rapidly adapted to the speech from all talkers without explicit feedback. Additionally, a large amount variation was observed in the extent of adaptation for the talkers, with proportion improvement ranging from 0.01 to 0.40. These data are shown in Figure 5, sorted by the Karl and Pisoni (1994) measure of intelligibility in quiet. A rank-ordered correlation indicated that the talkers with the greatest attunement were not correlated with the talkers with the highest intelligibility scores in quiet (rho = -0.12).

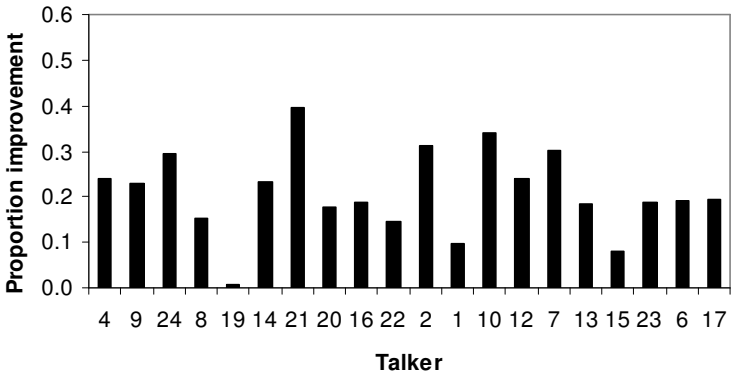


Figure 5: Proportion improvement from first 20 sentences to final 20 sentences for speech mixed with multi-talker babble. While listeners adapted to the speech from all talkers, the extent of adaptation depended on the particular talker.

In addition to comparing the perceptual adaptation in Babble to intelligibility in Quiet, we also compared the extent of adaptation in the two degradation conditions. A paired t-test revealed that listeners showed greater perceptual attunement improvement in the CI-simulated listening condition (mean = 0.43, SD = 0.11) than in the babble condition (mean = 0.21, SD = 0.09; $t(38) = 6.66$, $p<0.001$). However, the extent of adaptation in the two conditions was not correlated ($r=0.18$, *n.s.*).

Correlations among Acoustic-Phonetic Parameters and Intelligibility

To determine whether the acoustic-phonetic parameters that correlate with intelligibility in the Quiet condition also correlate with intelligibility under degraded conditions, the keyword intelligibility scores for the two degradation conditions were correlated with a variety of global acoustic-phonetic parameters measured from these sentences: fundamental frequency range (F0 range), mean fundamental frequency (F0 mean), vowel dispersion, first formant range (F1 range), second formant range (F2 range), and sentence duration. These measures were reported in Bradlow et al. (1996). We reanalyzed the data from Bradlow et al. using Pearson correlations rather than the Spearman correlations that they reported since we believe comparing the numerical values rather than rank orderings is more appropriate. With this reanalysis, under quiet listening conditions there were trends for F0 range and F0 mean to correlate with sentence intelligibility. No other acoustic-phonetic parameter was significantly correlated with intelligibility.

The parameters that showed a trend to correlate with intelligibility in quiet also tended to be correlated with intelligibility under the two degradation conditions. F0 range, which showed a trend to correlate with intelligibility in quiet, was significantly correlated with intelligibility under the CI-simulated conditions, ($r=0.549$, $p < 0.05$) and in the babble condition ($r=0.71$, $p < 0.001$). F0 mean, which also showed a trend to correlate with intelligibility in quiet, showed a trend for a correlation with intelligibility under CI-simulation ($r=0.361$, $p = 0.12$) and was significantly correlated in the babble condition ($r=0.58$, $p < 0.01$). F1 range which was not significantly correlated with intelligibility in quiet showed a trend to be correlated with intelligibility under CI simulation ($r=0.435$, $p = 0.06$) and was significantly correlated with intelligibility in babble ($r=0.53$, $p < 0.05$). F2 range and sentence duration were not correlated with intelligibility in any of the three listening conditions.

For the correlations between vowel dispersion and intelligibility, in addition to using overall intelligibility scores, intelligibility was measured with just a subset of sentences that included the point vowels (i.e., /i, a, o/) (see Bradlow et al. (1996) for more specifics). /o/ was chosen as the back vowel

rather than /u/ due to the high degree of allophonic variation in American English for this phoneme. In all three listening conditions, overall intelligibility did not correlate with vowel dispersion. In quiet, intelligibility for the 18 sentence-subset was not correlated with vowel dispersion. For the CI-simulated listening conditions, there was a trend for a correlation between keyword intelligibility and vowel dispersion (r=0.42, p = 0.07) while for the babble condition the correlation was not significant (r=0.04, *ns*). A summary of these results is shown in Table 1.

	Listening condition		
Acoustic measure	**Quiet**	**CI Simulation**	**Babble**
F0 range	0.39[t]	0.55*	0.71**
F0 mean	0.40[t]	0.36	0.58**
F1 range	0.32	0.44[t]	0.53*
F2 range	0.09	0.25	0.18
Vowel dispersion (all sentences)	0.11	0.37	0.05
Vowel dispersion (18 sentences)	0.27	0.42[t]	0.04
Sentence duration	-.01	0.33	0.34

Table 1: Correlations between acoustic-phonetic parameters and intelligibility in the three listening conditions (Quiet, CI-simulation and Babble). Correlation values are listed with asterisks indicating significance levels: one asterisk indicates a p-value of 0.05 or less, two asterisks indicates a p-value of 0.01 or less and a "t" indicates a trend with a p-value of 0.10 or less.

In both degradation conditions, talkers differed substantially in the degree to which listeners could adapt to their speech. The same acoustic-phonetic parameters that were correlated with intelligibility were also correlated with the proportion improvement scores to assess if certain acoustic-phonetic parameters can explain the large range in the extent of adaptation for individual talkers. For the CI-simulated listening conditions, vowel dispersion (r=0.46, $p < 0.05$) and F2 range (r=0.45, $p < 0.05$) correlated with proportion improvement whereas for the speech mixed with babble condition, none of the measured acoustic-phonetic parameters correlated with the proportion improvement scores.

Discussion

Results from the current study provide clear and consistent evidence that differences in intelligibility among talkers are not absolute; rather, inter-talker intelligibility scores are strongly dependent on the characteristics of the listener (CI simulation) and of the listening environment (Babble). This overall pattern converges with previous studies examining relative intelligibility among talkers, indicating that intelligibility rankings may change depending on listener language background (Bent & Bradlow, 2003; Imai et al., 2003; van Wijngaarden, 2001; van Wijngaarden et al., 2002). We add to this literature by demonstrating that other characteristics of the listener and the listening environment affect talker intelligibility. Our findings diverge from those of Green et al. (2007) who suggest that inter-talker differences are maintained across different listener groups (i.e., CI users, normal-hearing listeners presented with speech in a low level of babble or with CI simulated speech). The discrepancy between the current results and their results may be primarily due to the small number of talkers they tested who demonstrated intelligibility scores at the high and low ends of the intelligibility distribution. That is, we may expect that talkers of particularly high or low intelligibility in quiet (or very favorable signal to noise ratios as with Green et al.) will also be of high and low intelligibility for CI listeners or in CI simulated listening conditions as can be seen for talkers 17 and 6 in our study who were the top two most intelligible talkers in the quiet and were 2nd and 5th most intelligible in CI simulated listening conditions.

On the other hand, talkers with more moderate intelligibility scores may show more variation across different listener populations and listening environments, as is the case for talkers 10 and 2 in our study who were 8th and 10th most intelligible in quiet but were 14th and 1st most intelligible in CI simulated listening conditions. Furthermore, while intelligibility was tested here with sentences, Green et al. used word length materials. It remains possible that the factors that make words more or less intelligible may be more consistent across listener groups and listening conditions than the factors that influence sentence intelligibility. Future studies should test both word and sentence intelligibility for a large number of talkers to assess how the type of linguistic material interacts with talker characteristics.

The results of the Babble condition reveal that intelligibility under quiet listening conditions is not correlated with intelligibility under noisy listening conditions. However, the extent to which this result can be extended to other types of signal degradation remains an empirical issue. It is worth noting that the finding of a strong correlation between the two signal degradation conditions suggests the existence of features that enhance intelligibility in a wide range of difficult listening conditions. Significant correlations (or trends) between intelligibility and several of the acoustic-phonetic parameters measured (i.e., F0 range and F1 range) in both degradation conditions also suggest that certain acoustic-phonetic parameters may be important for enhancing intelligibility in multiple difficult listening situations.

Although mixing speech with babble is typically considered an ecologically valid noise-addition process, it should be noted that the same recordings – collected in quiet conditions – were used in the quiet and noise-added listening conditions. Therefore, modifications which talkers make when they are in noisy environments (e.g., Lombard speech, Lombard, 1911) are not performed in these recordings. It may be the case that certain talkers are more effective at making modifications that help listeners in noisy environments when they are producing speech with noise present. This issue remains a topic for future research.

The remainder of this section explores several issues raised by the data reported here. In particular, we address the issue of the acoustic-phonetic parameters that facilitate speech intelligibility, and discuss gender differences, perceptual adaptation, and the potential clinical implications of this work.

Acoustic-Phonetic Parameters

Three types of acoustic-phonetic parameters were examined in this study: those relating to fundamental frequency, vowel space characteristics, and measures of duration. Previous studies (Bond & Moore, 1994; Bradlow et al., 1996; Hazan & Markham, 2004) have measured other acoustic-phonetic parameters such as total energy in specific frequency regions, amplitude characteristics and specific cues to consonant contrasts. The comparison of the results from the current study with previous studies suggests that the particular talkers and materials used in a study may influence the acoustic-phonetic parameters that are significantly correlated with intelligibility. Furthermore, Hazan and Markham (2004) as well as studies investigating the acoustic-phonetic correlates of clear speech (Bradlow, Kraus & Hayes, 2003) suggest that speakers may be able to achieve highly intelligible speech through the manipulation of different combinations of acoustic-phonetic characteristics.

The results reported above support the claim that the acoustic-phonetic properties of a talker's speech that enhance intelligibility differ to some extent when listener and listening environment characteristics are changed. That is, only F0 range was correlated with intelligibility both for normal-hearing listeners in quiet listening conditions and under CI simulation, while others were only correlated with intelligibility under CI-simulation (F1 range and vowel dispersion). Similarly, F0 range was also correlated with intelligibility in the Babble condition as were F0 mean and F1 range, but vowel dispersion was not correlated with intelligibility in this condition.

In comparison with previous studies, fundamental frequency characteristics were a larger factor in the determination of speech intelligibility in the current study. In particular, both Hazan and Markham (2004) and Bond and Moore (1994) failed to find the correlations between fundamental frequency characteristics and intelligibility that were reported above. The other large discrepancy between the current study and previous studies came from measures of sentence or word duration. In the current study, as well as in Bradlow et al. (1996), intelligibility was not correlated with sentence duration. However, both Hazan and Markham (2004) and Bond and Moore (1994) found correlations between word duration and intelligibility. Because in naturally produced speech speakers manipulate multiple acoustic-phonetic parameters at the same time, it is difficult to definitively determine which parameters are most essential for highly intelligible speech. Future studies should, therefore, assess the contribution of individual acoustic-phonetic parameters by synthetically manipulating them and determining how changes in each parameter affect intelligibility. Furthermore, future studies should address how these synthetic manipulations interact with the linguistic materials (e.g., comparing words and sentences).

Perceptual Attunement

Listeners are able to quickly adapt their internal speech categories to more accurately perceive speech in a variety of different listening conditions. For example, listeners can adjust their category boundaries for phoneme contrasts (e.g., Eisner & McQueen, 2005). Also, listeners have shown both talker-dependent and talker-independent perceptual learning of speech such that experience or training with specific talkers, talker populations or synthesis conditions improves listeners' ability to accurately identify words from familiar talkers (e.g., Nygaard & Pisoni, 1998), new talkers from the same special population as they were exposed to in training (e.g., Bradlow & Bent, in press; McGarr, 1983) or speech that has been degraded in the same way as the training materials (e.g., Schwab, Nusbaum & Pisoni, 1985; Davis et al., 2005). Investigating listeners' adaptation to new talkers or speech patterns and the conditions that allow for this adaptation provides information about the extent of neural plasticity in the speech perception system. Furthermore, testing the conditions under which the most learning occurs can potentially help in the development of training programs for listeners with speech perception difficulties such as the hearing impaired or second language learners.

In the present experiment, the analysis of adaptation to the degraded speech revealed the flexibility of the speech perception system. Even in the absence of feedback, listeners interpreted the talker's utterances more accurately after several minutes of exposure to the experimental stimuli (i.e., last 20 sentences) compared to the beginning of exposure to these stimuli (i.e., first 20 sentences). The extent of adaptation varied for each talker and in each type of degradation, but the present data did not allow us to determine the source of this variability.

In the present research, the correlation between proportion improvement for each talker in the two degradation conditions was not significant. In addition, the analyses examining the relationship between perceptual attunement and acoustic-phonetic properties of the speaker did not yield conclusive results, and there was not a significant correlation between a talker's intelligibility in quiet and the perceptual attunement on that talker's speech in the experimental conditions. Furthermore, the extent of adaptation did not depend on overall intelligibility as proportion improvement was not correlated with overall intelligibility scores in either degradation condition. This result differs from the findings regarding adaptation to foreign accented speech (Bradlow & Bent, in press) in which listeners were better able to adapt to individual talkers who demonstrated high overall intelligibility than talkers of low intelligibility.

One likely cause of the greater adaptation in the CI Simulation condition compared to the Babble condition is the novelty of the former type of degradation. The listeners in the study had never experienced CI Simulated listening conditions before participating in the experiment, but each listener has

perceived speech with competing talkers daily. Thus, listeners are already practiced at picking out a given talker in noisy listening environments that are similar to the Babble condition, and must only adapt to the specifics of the multi-talker babble added to the speech in the experiment. We suggest here that this leaves listeners with less room to improve in the Babble condition than in the CI Simulation condition.

It is worth noting that the listeners in this experiment did not receive feedback, which suggests that they could have taken advantage of semantic and syntactic cues to enable them to learn how to perceive the speech under the two degradation conditions. It remains an open question whether perceptual learning would be as robust with anomalous sentences or nonsense words.

Gender

Previous studies have found that for normal-hearing adult and child listeners, adult female talkers are more intelligible than adult male talkers (Bradlow et al., 1996; Hazan & Markham, 2004). This result was consistent across different types of materials (i.e., both words and sentences).

The findings from the current study are consistent with previous results and support the claim that female talkers tend to be more intelligible then male talkers. The present study adds to the previous data by demonstrating that this result holds in a variety of listening environments (quiet and Babble) and for different listener populations (normal-hearing listeners and CI Simulation).

The source of this difference is not known at this point. It is possible that female talkers are generally more intelligible than their male counterparts because of physical differences in the vocal tracts For example, the higher mean fundamental frequency for most female talkers compared to male talkers will result in wider spacing of their vowel formants. This increased spacing between formants may lead to fewer formants being collapsed into one spectral channel that would presumably hinder vowel intelligibility. However, it remains possible that the gender differences come from a learned source of behavior. For example, female talkers could make articulatory adjustments that result in more intelligible speech. If this latter type of explanation is the source of this difference, it would suggest that male talkers may be able to be taught to alter their articulatory patterns to increase their intelligibility. It is clear that more work is critical to resolving this issue.

Cochlear Implant Users

One of the practical goals of this research was to develop strategies to increase the intelligibility of speech for listeners who have particular difficulty in speech communication. The current findings suggest several modifications that talkers could make when communicating with a CI user, and clinicians could make these suggestions to the family members and friends of cochlear implant recipients. First, talkers should be encouraged to increase their fundamental frequency range. Second, talkers should increase the distances between vowel categories. They can be instructed to do this by being told to speak clearly as tests of clear speech have shown that talkers tend to have greater vowel dispersion in clear speech than conversational speech (e.g. Bradlow, 2002). According to the findings from this study, talking at a slower speaking rate will not be particularly helpful to CI users. Liu et al (2004) also found that sentence produced in a clear speaking style were more intelligible to CI users than those produced in a conversational style. However, note that while generally clear speech is more intelligible to hearing impaired listeners (Picheny, Durlach & Braida, 1985), Ferguson and Kewley-Port (2002) found that certain clear speech modifications to vowels can actually be detrimental to hearing impaired listeners.

While the finding that normal-hearing listeners in quiet and listeners under CI simulated listening conditions find different talkers more and less intelligible is suggestive, it is necessary to test the validity of this approach by testing CI-users with the same stimuli. We currently have a study underway to test

whether the talkers who are of high and low intelligibility under CI-simulation will also be of high and low intelligibility for CI-users.

Conclusion

The current study suggests that intelligibility minimally needs to be characterized by a combination of talker-, listener- and listening environment- factors. This conclusion is in contrast to those studies that have suggested that acoustic-phonetic features of a talker's voice are the primary determinants of intelligibility levels (e.g. Hazan & Markham, 2004; Green et al., 2007). Furthermore, the acoustic-phonetic correlation analysis here suggests that while certain parameters may be beneficial for a wide range of listeners and listening environments, the importance of other parameters may vary depending on the listener and listening environment. Lastly, listeners were shown to adapt rapidly to speech in both the CI simulated and Babble conditions although the extent of adaptation differed widely across talkers.

References

Assmann, P.F & Summerfield, A.Q. (2004). The perception of speech under adverse conditions. In S. Greenberg, W.A. Ainsworth, A.N. Popper & R. Fay (Eds.), *Speech Processing in the Auditory System*. Springer-Verlag: New York.

Bent, T. & Bradlow, A.R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, *114,* 1600-1610.

Black, J.W. (1957). Multiple-choice intelligibility tests. *Journal of Speech and Hearing Disorders, 22,* 213-235.

Bond, Z.S. & Moore, T.J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication, 14,* 325-337.

Bradlow, A.R. (2002). Confluent talker- and listener-related forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (Pp. 241-273). Berlin & New York: Mouton de Gruyter.

Bradlow, A.R. & Bent, T. (in press). Perceptual adaptation to non-native speech. *Cognition.*

Bradlow, A.R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with learning disabilities: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, *46,* 80-97.

Bradlow, A.R., Toretta, G.M. & Pisoni, D.B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20,* 255-272.

Clarke, C.M. & Garrett, M.F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116,* 3647-3658.

Clopper, C.G. & Bradlow, A.R. (in press). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech.*

Cox, R.M., Alexander, G.C. & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *Journal of the Acoustical Society of America*, *81,* 1598-1608.

Davis, M.H., Johnsrude, I.., Hervais-Ademan, A., Taylor, K. & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134,* 222-241.

Dorman, M.F., Loizou, P.C. & Rainey, D. (1997). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding, *Journal of the Acoustical Society of America, 102,* 2993-2996.

Dupoux, E. & Green, K.P. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23,* 914-927.

Eisner, F. & McQueen, J.M. (2005). The specificity of perceptual learning in speech processing, *Perception and Psychophysics, 67,* 224-238.

Ferguson, S.H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, *112,* 259-271.

Green, T., Katiri, S., Faulkner, A., & Rosen, S. (2007). Talker intelligibility differences in cochlear implant listeners. *JASA Express Letters*, *121,* EL223-EL229.

Greenspan, S.L., Nusbaum, H.C. & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14,* 421-433.

Hazan, V. & Markham, D. (2004) Acoustic-phonetic correlates of talker intelligibility in adults and children. *Journal of the Acoustical Society of America, 116,* 3108-3118.

Hazan, V. & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects, *Language and Speech*, *43,* 273-294.

Hood, J.D. & Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, *19,* 434-455.

Howes, D. (1952). The intelligibility of spoken messages. *Journal of Psychology, 65,* 460-465.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29,* 296-305.

IEEE (1969). IEEE recommended practices for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*. *17,* 227-46.

Imai, S., Flege, J.E., & Walley, A. (2003) Spoken word recognition of accented and unaccented speech: Lexical factors affecting native and non-native listeners. In *Proceedings of the International Congress on Phonetic Science*, Barcelona, Spain.

Karl, J. & Pisoni, D.B. (1994). The role of talker-specific information in memory for spoken sentence. *Journal of the Acoustical Society of America*, 95, 2873.

Labov, W. & Ash, S. (1997). Understanding Birmingham. In C. Bernstein, T. Nunnally & R. Sabino (Eds.) *Language Variety in the South Revisited*. Tuscaloosa, AL: University of Alabama Press.

Liu, S., Del Rio, E., Bradlow, A.R., & Zeng, F-G. (2004) Clear speech perception in acoustic and electric hearing. *Journal of the Acoustical Society of America*, *116,* 2374-2383.

Lombard E. 1911. Le signe de l'elevation de la voix. *Annales de Maladies d L'oreille et du Larynx*, *37,* 101-119.

Major, R., Fitzmaurice, S., Bunta, F. & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36,* 173-190.

Mason, H.M. (1946). Understandability of speech in noise as affected by region of origin of speaker and listener. *Speech Monographs*, *13,* 54-68.

Mayo, L.H., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language and Hearing Research, 40,* 686-693.

McGarr, N.S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research*, *26,* 451-458.

Meador, D., Flege, J.E., & MacKay, I.R.A. (2000). Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition, 3,* 55-67.

Miller, G.A. (1947). The masking of speech. *Psychological Bulletin*, *44*, 105-129.

Miller, G.A. & Nicely, P.E. (1955). An analysis of perception confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338-352.

Moore, B.C.J. (2003). Speech processing for the hearing-impaired: successes, failures and implication for speech mechanisms. *Speech Communication*, *41,* 81-91.

Munro, M., Derwing, T., & Morton, S. (2006). The mutual intelligibility of foreign accents. *Studies in Second Language Acquisition*, 28, 111-131.

Nabelek, A.K. (1988). Identification of vowels in quiet, noise, and relationships with age, and hearing loss. *Journal of the Acoustical Society of America*, *84,* 476-484.

Nabelek, A.K. & Donohue, A.M. (1984). Perception of consonants in reverberation by native and non-native listeners. *Journal of the Acoustical Society of America*, *75,* 632-634.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5,* 42-46.

Pallier, C., Sebastian, Gallés, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition*, *26,* 844-851.

Picheny, M.A., Durlach, N.I., & Braida, L.D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of the Acoustical Society of America*, *28*, 96-103.

Rogers, C.L., Lister, J.J., Febo, D.M., Besing, J.M., & Abrams, H.B. (2006). Effects of bilingualism, noise and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics, 27,* 465-485.

Schwab, E.C., Nusbaum, H.C., & Pisoni, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27,* 395–408.

Takata, Y. & Nabelek, A.K. (1990). English consonant recognition in noise and in reverberation by Japanese and American listeners. *Journal of the Acoustical Society of America, 88,* 663-666.

van Wijngaarden S.J. (2001) Intelligibility of native and non-native Dutch speech. *Speech Communication*, *35,* 103-113.

van Wijngaarden, S.J., Steeneken, H.J.M, & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *Journal of the Acoustical Society of America*, *111*, 1906-1916.

Weil, S.A. (2001) Foreign-accented speech: Encoding and generalization. *Journal of the Acoustical Society of America*, *109,* 2473 (A).

Appendix A

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. Large size in stockings is hard to sell.
11. The boy was there when the sun rose.
12. A rod is used to catch pink salmon.
13. The source of the huge river is the clear spring.
14. Kick the ball straight and follow through.
15. Help the woman get back to her feet.
16. A pot of tea helps to pass the evening.
17. Smoky fires lack flame and heat.
18. The soft cushion broke the man's fall.
19. The salt breeze came across from the sea.
20. The girl at the booth sold fifty bonds.
21. The small pup gnawed a hole in the sock.
22. The fish twisted and turned on the bent hook.
23. Press the pants and sew a button on the vest.
24. The swan dive was far short of perfect.
25. The beauty of the view stunned the young boy.
26. Two blue fish swam in the tank.
27. Her purse was full of useless trash.
28. The colt reared and threw the tall rider.
29. It snowed, rained, and hailed the same morning.
30. Read verse out loud for pleasure.
31. Hoist the load to your left shoulder.
32. Take the winding path to reach the lake.
33. Note closely the size of the gas tank.
34. Wipe the grease off his dirty face.
35. Mend the coat before you go out.
36. The wrist was badly strained and hung limp.
37. The stray cat gave birth to kittens.
38. The young girl gave no clear response.
39. The meal was cooked before the bell rang.
40. What joy there is in living.
41. A king ruled the state in the early days.
42. The ship was torn apart on the sharp reef.
43. Sickness kept him home the third week.
44. The wide road shimmered in the hot sun.
45. The lazy cow lay in the cool grass.
46. Lift the square stone over the fence.
47. The rope will bind the seven books at once.
48. Hop over the fence and plunge in.
49. The friendly gang left the drug store.

50. Mesh wire keeps chicks inside.
51. The frosty air passed through the coat.
52. The crooked maze failed to fool the mouse.
53. Adding fast leads to wrong sums.
54. The show was a flop from the very start.
55. A saw is a tool used for making boards.
56. The wagon moved on well oiled wheels.
57. March the soldiers past the next hill.
58. A cup of sugar makes sweet fudge.
59. Place a rosebush near the porch steps.
60. Both lost their lives in the raging storm.
61. We talked of the side show in the circus.
62. Use a pencil to write the first draft.
63. He ran half way to the hardware store.
64. The clock struck to mark the third period.
65. A small creek cut across the field.
66. Cars and busses stalled in snow drifts.
67. The set of china hit, the floor with a crash.
68. This is a grand season for hikes on the road.
69. The dune rose from the edge of the water.
70. Those words were the cue for the actor to leave.
71. A yacht slid around the point into the bay.
72. The two met while playing on the sand.
73. The ink stain dried on the finished page.
74. The walled town was seized without a fight.
75. The lease ran out in sixteen weeks.
76. A tame squirrel makes a nice pet.
77. The horn of the car woke the sleeping cop.
78. The heart beat strongly and with firm strokes.
79. The pearl was worn in a thin silver ring.
80. The fruit peel was cut in thick slices.
81. The Navy attacked the big task force.
82. See the cat glaring at the scared mouse.
83. There are more than two factors here.
84. The hat brim was wide and too droopy.
85. The lawyer tried to lose his case.
86. The grass curled around the fence post.
87. Cut the pie into large parts.
88. Men strive but seldom get rich.
89. Always close the barn door tight.
90. He lay prone and hardly moved a limb.
91. The slush lay deep along the street.
92. A wisp of cloud hung in the blue air.
93. A pound of sugar costs more than eggs.
94. The fin was sharp and cut the clear water.
95. The play seems dull and quite stupid.
96. Bail the boat, to stop it from sinking.
97. The term ended in late June that year.
98. A tusk is used to make costly gifts.
99. Ten pins were set in order.
100. The bill as paid every third week.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Hearing Impairment and Correlations with Neuropsychological Function in Alzheimer's Disease, Mild Cognitive Impairment and Older Adults with Cognitive Complaints[1]

Vanessa Taler, Kashif Shaikh,[2] John D. West,[2] David B. Pisoni and Andrew J. Saykin[2]

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] The present research was supported by a postdoctoral fellowship to V.T. from the Fonds de la Recherche en Santé du Québec and by National Institutes of Health DC00012.
[2] Currently at the Indiana University School of Medicine, Indianapolis.

Hearing Impairment and Correlations with Neuropsychological Function in Alzheimer's Disease, Mild Cognitive Impairment and Older Adults with Cognitive Complaints

Abstract. We examined hearing status in groups of participants diagnosed with Alzheimer's disease (AD), mild cognitive impairment (MCI), or presenting with subjective cognitive complaints (CC), as well as healthy elderly individuals (HE). Baseline hearing status differed across the groups, with AD individuals showing higher pure-tone thresholds than HE, CC and MCI groups. MCI individuals who went on to develop AD showed higher thresholds than those who remained stable, although this finding did not reach statistical significance. Hearing thresholds correlated significantly with verbal and non-verbal memory performance in HE participants as well as the patient groups.

Introduction

Communication impairments are well-documented in Alzheimer's disease (AD) and are a major source of stress for both caregivers and patients. While a good deal of research has focused on standardized language tests in this population, less work has been done examining low-level hearing performance in this population, and the links between hearing performance and cognitive decline.

A second patient group that is of major interest in delineating the factors contributing to cognitive decline in AD are individuals with mild cognitive impairment (MCI). Individuals diagnosed with MCI present with subjective and objective memory impairment in the absence of dementia (Petersen et al., 1999). Patients with MCI have a greatly elevated risk of developing AD, with a conversion rate of approximately 10-15% per annum, versus 1-2% in the general elderly population (Chertkow, 2002). This population thus represents pre-clinical AD in the majority of cases. Recent research has revealed the opportunity to study AD at an even earlier stage. Individuals presenting with subjective cognitive complaints ((CC) but normal neuropsychological performance have been shown to exhibit similar patterns of cortical atrophy to those seen in MCI (Saykin et al., 2006), and thus may represent very early AD in many cases.

The present study investigates hearing and neuropsychological function in these populations using data from a five-year longitudinal study of healthy elderly adults, as well as groups with AD, MCI or CC. As part of the initial screening, patients underwent a pure-tone audiometric hearing assessment. Previous research has indicated that hearing loss is greater in AD patients than in healthy elderly (Weinstein & Amsel, 1987) and that this correlates with cognitive performance (Uhlmann et al., 1989). Thus, we examined correlations between neuropsychological performance and hearing thresholds in this population.

The aims of the present study were threefold. First, we wished to determine whether extent of cognitive decline was related to hearing performance; that is, whether differences in hearing thresholds would be observed across AD, MCI and CC groups. Second, we wanted to explore the relationship between hearing loss and longitudinal cognitive performance, with the goal of determining whether baseline hearing thresholds predicted cognitive decline. Finally, we examined correlations between hearing loss and performance on verbal and non-verbal memory tests, as well as several other neuropsychological measures. We predicted that verbal tasks, which rely heavily on auditory input, would correlate better than non-verbal tasks with hearing thresholds.

Methods

Participants

Four groups took part in the present study: participants with probable Alzheimer's disease (AD), mild cognitive impairment (MCI), or cognitive complaints (CC), and healthy elderly (HE). Classification of groups reflects initial diagnosis. All participants were aged ≥ 60 years, were right handed and were fluent speakers of English. Participant characteristics are presented in Table 1.

	Healthy Elderly - mean (SD)	**CC[a] Participants - mean (SD)**	**MCI[b] Participants - mean (SD)**	**AD[c] Participants - mean (SD)**
N	39	37	44	8
Age (years)	70.77 (5.34)	72.59 (6.19)	71.72 (8.51)	73.88 (6.33)
Education (years)	16.92 (2.60)	16.32 (2.92)	16.22 (3.06)	15.00 (3.59)
Sex	28F/11M	21F/16M	20F/23M[e]	4F/4M
MMSE (/30)[d]	29.05 (1.05)	29.03 (1.09)	26.91 (2.10)	24.38 (2.92)
Mother's Educational Level (years)	12.15 (3.13)	12.32 (3.21)	12.07 (3.49)	10.88 (3.48)
Father's Educational Level (years)	13.10 (4.46)	12.41 (4.08)	13.24 (5.12)	11.13 (3.80)
IQ	116.82 (4.34)	115.91 (5.61)	115.95 (5.87)	112.86 (9.61)

Table 1. Demographic characteristics of the 4 participant groups.
[a]CC – cognitive complaints. [b]MCI – mild cognitive impairment. [c]AD – Alzheimer's disease. [d] HE = CC > MCI > AD. [e] One missing data point.

AD Participants. Eight AD patients took part in the study. The diagnosis of dementia was established by a neurologist or neuropsychologist according to standard diagnostic criteria (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; American Psychiatric Association, 1994), and the diagnosis of AD was established according to NINCDS-ADRDA criteria (McKhann et al., 1984).

MCI Participants. Forty-four MCI patients took part in the study. Individuals were classified on the basis of neuropsychological assessment, self and informant reports, and geropsychiatric and neurologic evaluation. Diagnosis of MCI was based on clinical consensus according to the following

criteria: (1) abnormal memory performance; (2) significant memory complaints, corroborated by an informant; (3) relatively preserved general cognitive functioning; (4) generally normal activities of daily living; (5) no dementia; (6) no depression or other major psychiatric disorder. The MCI participants performed 1.5 SDs below the adjusted mean of HEs on at least one verbal memory test score (CVLT Total 1-5, Short Delay, Long Delay, WMS-III LM I or LM-II).

CC Participants. Thirty-seven individuals with CC took part in the study. CC participants were also classified by consensus, fulfilling criteria (2)-(6) outline above but exhibiting normal performance on memory tests.

Healthy Elderly. Thirty-nine HE participants took part in the study. Healthy elderly fulfilled criteria (3)-(6) above but exhibited no subjective or objective memory impairment.

Neuropsychological Battery

Subjects underwent an extensive neuropsychological battery that examined general cognitive function, intelligence, memory, verbal learning, executive function and language. The following tests were included in the battery:

Mini-Mental State Examination (MMSE; Folstein et al., 1975). The MMSE is a brief test of cognitive function assessing orientation, registration, attention and calculation, recall, and language. It takes about 10 minutes and generally serves as a first measure in assessing cognitive decline. The MMSE is scored out of 30, with a score of 23 or below indicating cognitive impairment.

Mattis Dementia Rating Scale (DRS; Mattis, 1976). In order to obtain a more sensitive measure of dementia severity, participants completed this more extensive test that measures cognitive functioning across five subscales: attention, initiation-perseveration, construction, conceptualization, and memory. Scores range from 0 to 144, with higher scores representing better cognitive function.

Weschler Adult Intelligence Scale-III (WAIS-III; Wechsler, 1997a). The WAIS is a general test of intelligence that includes 14 measures of verbal and performance IQ. The verbal subtests include information, comprehension, arithmetic, similarities, vocabulary, digit span, and letter-number sequencing. The performance subtests include picture completion, digit symbol, block design, matrix reasoning, picture arrangement, symbol search, and object assembly. The battery of tests assesses verbal comprehension, perceptual organization, working memory, and processing speed.

Weschler Memory Scale-III (WMS-III; Wechsler, 1997b). The WMS-III includes eleven subtests assessing auditory immediate memory, auditory delayed memory, visual immediate memory, visual delayed memory delayed auditory recognition, and working memory.

California Verbal Learning Test (CVLT; Delis et al., 1987). The CVLT is a test that assesses verbal learning and memory. Participants listen to 16 words from 4 categories (4 items per category). They must then either repeat them or recognize them from a list of 44 items including distractors.

Delis Kaplan Executive Function System (DKEFS; Delis et al., 2001). The DKEFS is a set of standardized tests comprising nine subtests: trail making, verbal fluency, design fluency, color-word interference, sorting, twenty questions, word context, the Tower test, and a proverb test. These tests assess the integrity of executive functions and determine if deficits in abstract thinking impact the patient's daily life.

Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948). The WCST is a set-shifting test that evaluates participants' ability to adapt to constantly changing requirements. Participants must sort 64 cards according to criteria that switch periodically during testing. The WCST assesses executive functions.

Boston Naming Test (BNT; Kaplan et al., 1983). The BNT is a picture naming task in which participants name 60 line drawings of decreasing frequency.

Hearing Screen

Assessment of hearing ability was conducted using pure-tone audiometry. Each subject was tested separately in each ear at 500Hz, 1kHz, 2kHz, 3kHz, 4kHz, 6kHz, and 8kHz. A pure-tone average (PTA) for each ear was derived from averaging the audiogram data for 500Hz, 1kHz, and 2kHz.

Results

The data were analyzed to determine: (1) whether differences would be observed across the groups in baseline hearing status; (2) whether hearing status correlated with performance on verbal and non-verbal memory tests; and (3) whether baseline hearing status predicted cognitive decline. The results of each analysis are presented separately.

Group Differences in Baseline Hearing Status

Pure tone average thresholds in right and left ears for each group are presented in Figure 1. Visual inspection of the figures reveals increased hearing thresholds in both ears in the AD group relative to the CC and MCI groups, and in the CC and MCI groups relative to the HE group. A 4 (group) x 2 (left vs. right ear) ANOVA revealed a main effect of group ($F(3,124) = 3.02$, $p < 0.03$). LSD posthocs indicated a significant difference in both ears between AD participants and the remaining three groups ($p < 0.05$ in all cases), but no significant differences between HE, CC and MCI groups.

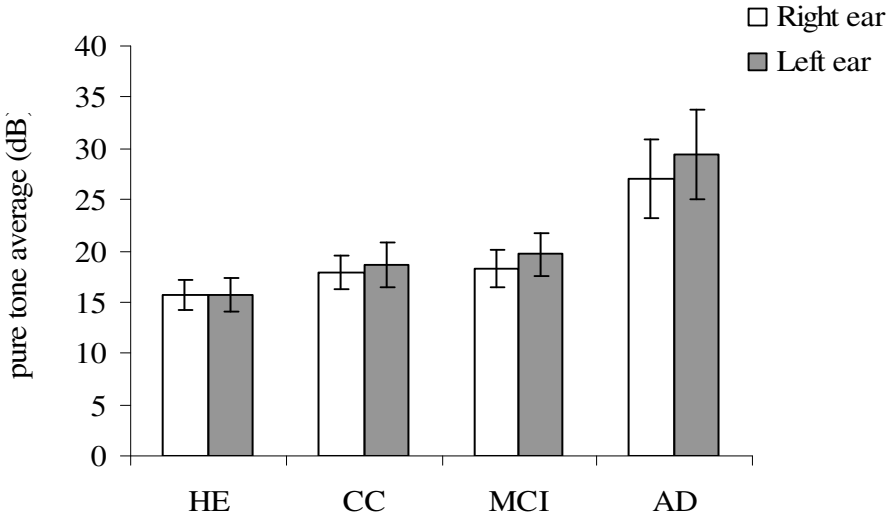


Figure 1. Pure tone average thresholds in the right and left ears for each participant group. Error bars represent standard error.

Correlations with Neuropsychological Function

A second issue of interest is whether hearing performance correlates with neuropsychological function in these populations. To address this question, we conducted Pearson correlations between hearing thresholds and scores on verbal (CVLT) and non-verbal (visual reproduction) memory tests. All participant groups were pooled for this analysis. Correlations were observed between CVLT scores in both ears (left ear: r = -0.24; $p < 0.005$; right ear: r = -0.32; $p < 0.001$), indicating lower performance on the CVLT with increasing hearing thresholds. The relationship between performance on non-verbal memory tests (the immediate and delayed visual reproduction subtests of the WMS) and hearing function was also assessed, with all participants pooled. Again, negative correlations were observed between performance on both measures and thresholds in both ears (immediate visual reproduction: right ear, r = -0.34 $p < 0.001$, left ear, r = -0.30, $p < 0.005$; delayed visual reproduction: right ear, r = -0.22 $p < 0.01$, left ear, r = -0.25, $p < 0.005$). A second set of correlation analyses including only HE participants found significant negative correlations between the CVLT and immediate visual reproduction and the right ear but not the left ear (CVLT: right ear, r = 0.38, $p < 0.02$, left ear: r = 0.27, $p < 0.09$; immediate visual reproduction: right ear, r = -0.38 $p < 0.02$, left ear, r = -0.22, $p < 0.18$; delayed visual reproduction: right ear, r = -0.20, $p > 0.23$, left ear, r = -0.26, $p > 0.11$).

Correlations were also performed for the remaining tests in the neuropsychological battery. Means across patient groups and correlations with hearing function are presented in Tables 2-4. When all participants are pooled, strong correlations are observed between hearing function and several neuropsychological measures, although these correlations do not remain significant when groups are analyzed separately.

	HE Participants	CC Participants	MCI Participants
MMSE (/30)	29.10 (1.09)	28.98 (1.09)	26.90 (1.86)
CVLT (	50.13 (8.85)	46.95 (9.32)	31.22 (6.29)
VRI (%ile)	78.10 (10.90)	75.02 (13.06)	63.70 (18.60)
VRD (%ile)	52.85 (20.85)	43.36 (17.11)	29.82 (18.53)
BNT (/60)	57.9 (1.90)	56.7 (2.90)	55.7 (3.00)
WCST	3.804 (1.37)	3.523 (1.28)	2.816 (1.23)
DRS (/144)	141.00 (2.34)	141.15 (2.39)	136.14 (5.34)
DGSY	62.02 (14.47)	63.68 (13.17)	50.78 (12.28)
DGSP	16.87 (3.09)	17.59 (3.91)	16.37 (3.91)
DTR1sc	23.93 (4.72)	23.83 (6.75)	28.83 (7.50)
DTR1er	0.13 (0.35)	0.22 (0.53)	0.28 (0.51)

Table 2. Group performance on neuropsychological measures.
MMSE = Minimental State Examination; CVLT = California Verbal Learning Test; VRI = Visual Reproduction – Immediate; VRD = Visual Reproduction – Delayed; BNT = Boston Naming Test; WCST = Wisconsin Card Sorting Test; DRS = Mattis Dementia Rating Scale; WAIS-DS = WAIS – Digit Symbol; WAIS-DSP = WAIS – Digit Span; DTR1sc = DKEF visual scanning, seconds; DTR1er = DKEF visual scanning, errors.
* indicates significance at $p < 0.05$

	All Participants	HE Participants	CC Participants	MCI Participants
MMSE	**-0.19***	-0.05	-0.10	-0.04
CVLT	**-0.24***	-0.27	-0.06	-0.18
VRI	**-0.30***	-0.22	-0.21	-0.26
VRD	**-0.25***	-0.26	-0.11	-0.11
BNT	-0.10	-0.17	0.01	-0.04
WCST	-0.12	0.04	-0.11	0.05
DRS (/144)	**-0.18***	-0.18	-0.14	-0.10
WAIS-DS	**-0.22***	0.02	-0.31	-0.17
WAIS-DSP	-0.06	0.04	-0.22	0.08
DTR1sc	0.18	-0.08	0.28	**0.38***
DTR1er	**0.22***	**0.49***	0.17	0.12

Table 3. Correlations between neuropsychological measures and left ear pure tone average threshold.
MMSE = Minimental State Examination; CVLT = California Verbal Learning Test; VRI = Visual Reproduction – Immediate; VRD = Visual Reproduction – Delayed; BNT = Boston Naming Test; WCST = Wisconsin Card Sorting Test; DRS = Mattis Dementia Rating Scale; WAIS-DS = WAIS – Digit Symbol; WAIS-DSP = WAIS – Digit Span; DTR1sc = DKEF visual scanning, seconds; DTR1er = DKEF visual scanning, errors.
* indicates significance at $p < 0.05$

	All Participants	HE Participants	CC Participants	MCI Participants
MMSE (/30)	**-0.21***	-0.021	-0.14	-0.15
CVLT	**-0.32***	-0.38	-0.28	**-0.34***
VRI	**-0.34***	-0.38	-0.23	-0.28
VRD	**-0.22***	-0.20	-0.10	-0.10
BNT	-0.13	-0.13	0.04	-0.01
WCST	-0.12	0.03	-0.19	0.09
DRS (/144)	**-0.22***	-0.23	-0.06	-0.17
DGSY	**-0.22***	-0.05	-0.30	-0.09
DGSP	-0.05	0.08	-0.18	0.06
DTR1sc	**0.34***	0.11	**0.36***	0.29
DTR1er	**0.18***	**0.49***	0.05	0.12

Table 4. Correlations between neuropsychological measures and right ear pure tone average threshold.
MMSE = Minimental State Examination; CVLT = California Verbal Learning Test; VRI = Visual Reproduction – Immediate; VRD = Visual Reproduction – Delayed; BNT = Boston Naming Test; WCST = Wisconsin Card Sorting Test; DRS = Mattis Dementia Rating Scale; WAIS-DS = WAIS – Digit Symbol; WAIS-DSP = WAIS – Digit Span; DTR1sc = DKEF visual scanning, seconds; DTR1er = DKEF visual scanning, errors.
* indicates significance at $p < 0.05$

Prediction of Conversion to AD

Seven of the 44 MCI participants converted to probable AD over the course of this study. Average pure tone thresholds of converters and non-converters are shown in Figures 2 (pure tone average) and 3 (averages from 500-4000 Hz).

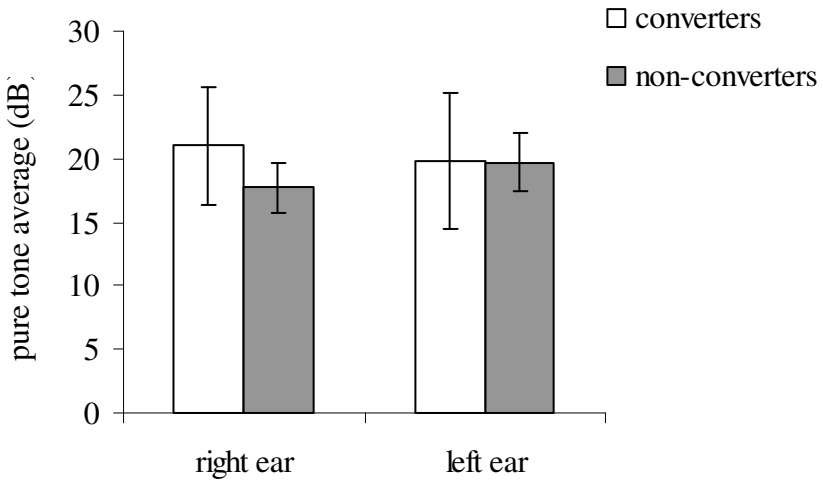


Figure 2. Pure tone average thresholds in the right and left ears for MCI participants who went on to develop probable AD (converters) and those who remained stable (non-converters). Error bars represent standard error.

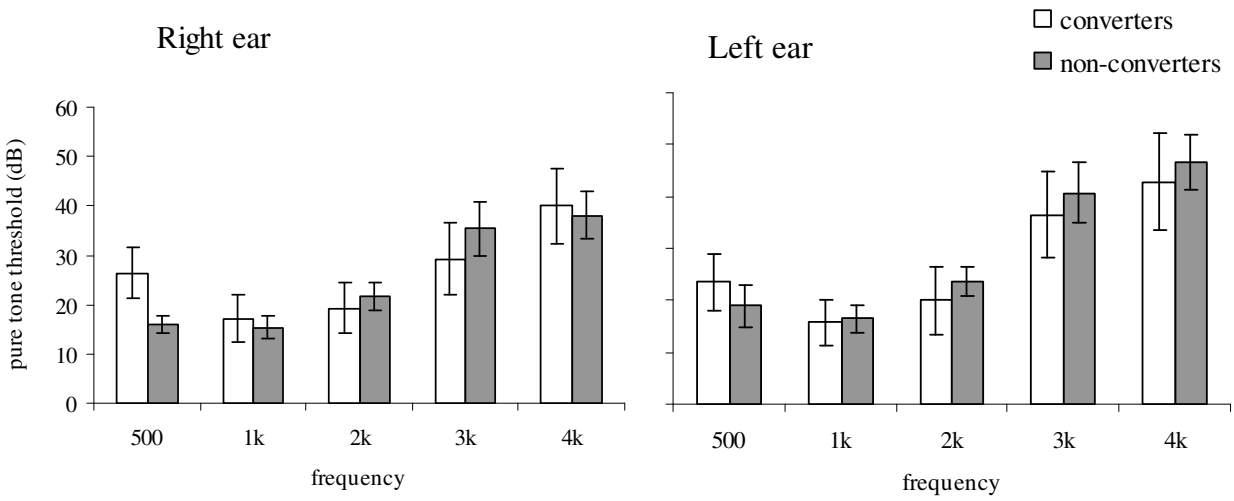


Figure 3. Pure tone average thresholds (500-4000 Hz) in the right and left ears for MCI participants who went on to develop probable AD (converters) and those who remained stable (non-converters). Error bars represent standard error.

While the figures demonstrate that the converter subgroup had higher thresholds than the non-converter subgroup in the pure tone average for the right ear and in lower frequencies (500 Hz), this comparison did not reach statistical significance.

Discussion

The present study demonstrated higher pure-tone hearing thresholds in AD participants than in HE, CC and MCI individuals. MCI and CC participants had higher thresholds than the healthy elderly, although this finding did not reach significance. Participants' thresholds were found to correlate with performance on verbal and non-verbal memory tests, in addition to a number of other cognitive functions. Moreover, hearing thresholds were higher in those MCI participants who went on to convert to AD, albeit not significantly so.

The finding that hearing loss is more severe in AD participants than in healthy elderly is consistent with a number of previous studies (e.g., Gates et al., 1995; Sinha et al., 1993; Uhlmann et al., 1989). Additionally, MCI and CC groups show a similar albeit statistically insignificant pattern, with hearing thresholds intermediate between those of AD and HE participants, suggesting that AD-related hearing loss may already be underway in this population. These data clearly indicate a relationship between hearing impairment and cognitive function.

A number of studies have found altered auditory evoked potentials in AD (Cancelli et al., 2006; Pekkonen et al., 1999) and MCI (Golob et al., 2007), reflecting deficits in sensory gating, which may be due to diminished hearing function in this population. It has been argued that these alterations in sensory gating may be related to dysfunction in the α-7 subunit of the cholinergic nicotinic receptor (Jessen et al., 2001), providing a possible neural substrate for the hearing impairment seen in this and other studies. Additionally, a recent study indicates alterations in dendritic arborization and loss of dendritic spines in the auditory cortex of early AD patients (Baloyannis et al., 2007); it is possible that the hearing impairments observed in our patients may be related to this auditory cortex pathology.

Correlations were also found with hearing function and a number of neuropsychological tests, including verbal and non-verbal memory. While it is possible that hearing impairment is compromising performance on these tasks, it is likely the case that AD-related hearing dysfunction is progressing in tandem with cognitive impairment but not influencing neuropsychological task performance, given that verbal and non-verbal memory are affected equally.

Interestingly, the HE group also showed a correlation between right-ear hearing loss and cognitive performance. This result was entirely unexpected, given the subclinical hearing loss seen in this population as well as the fact that cognition is not impaired in this group. Given that neither cognitive decline nor AD-related hearing decline is expected in this population, this finding suggests that hearing function may indeed impact upon neuropsychological performance. Another possibility is that the healthy elderly population contains individuals who will soon develop cognitive complaints or MCI, and these individuals are driving the correlations between hearing decline and cognitive performance. Future research should examine the possible causal link between hearing loss and neuropsychological performance in these populations.

Finally, we were interested in examining the possible predictive value of measures of hearing threshold in MCI. To this end, we compared the hearing thresholds of those MCI participants who converted to probable AD and those who did not. Converters had higher baseline thresholds, although

these differences did not reach significance, likely due to the small number of converters in the patient sample (7 of 44). Further research is clearly necessary to explore the question of whether hearing decline predicts conversion from MCI to AD.

In sum, the significantly higher hearing thresholds in AD compared to healthy elderly participants may be due to cortical pathology in AD. The finding that MCI and CC participants showed similarly elevated thresholds at baseline suggests that this auditory decline may be occurring even in very early AD, although these results did not reach significance. Additionally, the negative correlations between hearing thresholds and neuropsychological test performance may be due to concomitant decline in the two domains. However, the finding that this correlation also holds for healthy elderly opens the possibility that there may be a causal link, something that should be explored further. Additionally, our results point toward the possibility that auditory function is poorer in those MCI patients who will go on to develop AD, although this hypothesis should be tested with a larger sample.

Hearing declines in early AD have important implications at a number of levels, including neuropsychological assessment, as well as speech communication with these individuals. Communication impairments in AD are a major source of caregiver and patient stress, and contribute to increasing caregiver burden and breakdown of social relationships. As such, a better understanding of hearing impairment in this population is crucial to improve quality of life and care for this vulnerable population. Our results indicate that hearing thresholds are significantly higher in AD participants than in the healthy elderly, and suggest that impairments may be present even in very early AD. As such, a hearing screen should form part of any routine clinical examination for these patients.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th Ed.). Washington, DC: Author.

Baloyannis, S.J., Costa, V., Mauroudis, I., Psaroulis, D., Manolides, S.L., & Manolides, L.S. (2007). Dendritic and spinal pathology in the acoustic cortex in Alzheimer's disease: morphological and morphometric estimation by Golgi technique and electron microscopy. *Acta Oto-Laryngologica, 127,* 351-354.

Cancelli, I., Pittaro Cadore, I., Merlino, G., Valentinis, L., Moratti, U., Bergonzi, P., et al. (2006). Sensory gating deficit assessed by P50/Pb middle latency event related potential in Alzheimer's disease. *Journal of Clinical Neurophysiology, 23,* 421-425.

Chertkow, H. (2002). Mild cognitive impairment. *Current Opinion in Neurobiology, 15,* 401-407.

Delis, D.C., Kaplan, E., & Kramer, J.H. (2001). *Delis-Kaplan Executive Function System (D-KEFS).* San Antonio, TX: The Psychological Corporation.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (1987). *California Verbal Learning Test.* San Antonio, TX: Psychological Corporation.

Folstein, M.J., Folstein, S.E., & McHugh, P.R. (1975). Mini-mental state: a practical method for grading the cognitive state of the patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Gates, G.A., Karzon, R.K., Garcia, P., Peterein, J., Storandt, M., Morris, J.C., et al. (1995). Auditory dysfunction in aging and senile dementia of the Alzheimer's type. *Archives of Neurology, 52,* 626-634.

Golob, E.J., Irimajiri, R., & Starr, A. (2007). Auditory cortical activity in amnestic mild cognitive impairment: relationship to subtype and conversion to dementia. *Brain, 130,* 740-752.

Grant, D.A., & Berg, E.A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card sorting problem. *Journal of Experimental Psychology, 38,* 404-411.

Jessen, F., Kucharski, C., Fries, T., Papassotiropoulos, A., Hoenig, K., Maier, W., et al. (2001). Sensory gating deficit expressed by a disturbed suppression of the P50 event-related potential in patients With Alzheimer's disease. *American Journal of Psychiatry, 158,* 1319–1321.

Kaplan, E.F., Goodglass, H., & Weintraub, S. (1983). *Boston Naming Test.* Philadelphia, PA: Lea & Febiger.

Mattis, S. (1976). Mental status examination for organic mental syndrome in the elderly patient. In L.B.T.B Karasu (Ed), *Geriatric Psychiatry,* Pp. 77-121. New York: Grune & Stratton.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E.M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34,* 939-944.

Pekkonen, E., Jaaskelainen, I.P., Hietanena, M., Huotilainen, M., Naatanen, R., Ilmoniemic, R.J., et al. (1999). Impaired preconscious auditory processing and cognitive functions in Alzheimer's disease. *Clinical Neurophysiology, 110,* 1942-1947.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology, 56,* 303-308.

Saykin, A.J., Wishart, H.A., Rabin, L.A., Santulli, R.B., Flashman, L.A., West, J.D., et al. (2006). Older adults with cognitive complaints show brain atrophy similar to that of amnestic MCI. *Neurology, 67,* 834-842.

Sinha, U.K., Hollen, K.M., Rodriguez, R., & Miller, C.A., (1993). Auditory system degeneration in Alzheimer's disease. *Neurology, 43,* 779-785.

Uhlmann, R.F., Larson, E.B., Rees, T.S., Koepsell, T.D., & Duckert, L.G. (1989). Relationship of hearing impairment to dementia and cognitive dysfunction in older adults. *Journal of the American Medical Association, 261,* 1916-1919.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale—Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale (WMS-III).* San Antonio, TX: The Psychological Corporation.

Weinstein, B.E., & Amsel, L. (1987). Hearing impairment and cognitive function in Alzheimer's disease. *Journal of the American Geriatric Society, 35,* 273-275.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Links Between Implicit Learning and Spoken Language Processing: Some Preliminary Data[1]

Christopher M. Conway and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH Grant DC00012. We wish to thank Jennifer Karpicke and Luis Hernandez for their invaluable assistance on this project.

Links Between Implicit Learning and Spoken Language Processing: Some Preliminary Data

Abstract. Spoken language consists of a complex, time-varying signal that contains sequential patterns that can be described in terms of statistical relations among language units. Previous research has suggested that a domain-general ability to learn structured sequential patterns may underlie language acquisition. To test this prediction, we examined the extent to which implicit sequence learning of probabilistically-structured patterns in normal-hearing adults is correlated with performance on a spoken sentence perception task under degraded listening conditions. Our data revealed that performance on the sentence perception task correlated with implicit sequence learning, but only when the sequences were composed of stimuli that were easy to encode verbally. The evidence is consistent with the hypothesis that implicit learning of phonological sequences is an important cognitive ability that contributes to spoken language processing abilities.

Introduction

It has long been recognized that language comprehension involves the coding and manipulation of sequential patterns (Lashley, 1951; see also Conway & Christiansen, 2001). Spoken language can be thought of as patterns of sound symbols occurring in a sequential stream. Many of the sequential patterns of language are fixed, that is, they occur in a consistent, regular order (e.g., words are fixed sequences of phonemes). Thus, being able to encode and store in memory fixed sequences of sounds would appear to be a key aspect of language learning. Empirical work with normal-hearing adults and children supports this view, showing a strong link between sequence memory, word learning, and vocabulary development (for a review, see Baddeley, 2003).

Although short-term verbal memory is undoubtedly important for learning *fixed* sequences in language, such as words or idioms, the learning of more complex, highly variable patterns in language may require a different kind of cognitive mechanism altogether (Conway & Christiansen, 2001). For instance, in addition to fixed sequential patterns of sounds, spoken language also contains sequences that can be described in terms of complex statistical relations among language units. Rarely is a spoken utterance perfectly predictable; most often, the next word in a sentence can only be partially predicted based on the preceding context (Rubenstein, 1973). It is known that sensitivity to such probabilistic information in the speech stream can improve the perception of spoken materials in noise; the more predictable a sentence is, the easier it is to perceive it (Kalikow et al., 1977). Therefore, the ability to extract probabilistic or statistical patterns from the speech stream may be a factor that is important for language learning and spoken language processing: the better able one is at implicitly learning the sequential patterns in language, the better one should be at processing upcoming spoken materials in an utterance, especially under highly degraded listening conditions.

In this paper, we examine the hypothesis that a domain-general ability to implicitly encode complex sequential patterns underlies aspects of spoken language processing. This kind of incidental, probabilistic sequence learning has been investigated in some depth over the last few years under the rubrics of "implicit", "procedural", or "statistical" learning (Cleeremans, Destrebecqz, & Boyer, 1998; Conway & Christiansen, 2006; Saffran, Aslin, & Newport, 1996; Stadler & Frensch, 1998). To help

elucidate the link between implicit learning and language processing, we used a new experimental methodology that was developed to assess sequence memory and learning based on Milton Bradley's Simon memory game (e.g., Pisoni & Cleary, 2004). In this task, participants see sequences of colored lights and/or sounds and are required to simply reproduce each sequence by pressing colored response panels in correct order.

Not only can the Simon task be used to assess learning and memory of fixed sequences, but it can also be used to measure implicit sequence learning of more complex rule-governed or probabilistic patterns (Karpicke & Pisoni, 2004). In the present experiment, we used a version of the Simon task that incorporates visual-only stimuli that contained structural regularities, and correlated participants' performance on the implicit learning task with their ability to perceive spoken sentences that varied in terms of the final word's predictability, under degraded listening conditions. Before describing the study in full, we first briefly review previous evidence related to implicit learning and language processing.

Implicit Sequence Learning and Language

Implicit learning involves automatic, unconscious learning mechanisms that extract regularities and patterns that are present across a set of exemplars, typically without direct awareness of what has been learned. Many researchers believe that implicit learning is one of the primary mechanisms through which children learn language (Cleeremans et al., 1998; Conway & Christiansen, 2001; Dominey, Hoen, Blanc, & Lelekov-Boissard, 2003; Ullman, 2004): language acquisition, like implicit learning, also involves the incidental, unconscious learning of complex sequential patterns. This perspective on language development is supported by recent findings showing that infants engage implicit learning processes to extract the underlying statistical patterns in language-like stimuli (Gómez & Gerken, 2000; Saffran et al., 1996).

Although it is a common assumption that implicit learning is important for language processing, the evidence directly linking the two processes is mixed. One approach is to assess language-impaired individuals on a putatively non-linguistic implicit learning task; if the group shows a deficit on the implicit learning task, this result is taken as support for a close link between the two cognitive processes. Using this approach, some researchers have found an implicit sequence learning deficit in dyslexics (Howard, Howard, Japikse, & Eden, 2006; Menghini, Hagberg, Caltagirone, Petrosini, & Vicari, 2006; Vicari, Marotta, Menghini, Molinari, & Petrosini, 2003) while others have found no connection between implicit learning, reading abilities, and dyslexia (Kelly, Griffiths, & Frith, 2002; Rüsseler, Gerth, & Münte, 2006; Waber et al., 2003). At least with regard to reading and dyslexia, the role of implicit learning is not clear (also see Grunow, Spaulding, Gómez, & Plante, 2006).

One complication with establishing an empirical link between implicit learning and language processing is that implicit learning itself may involve multiple subsystems that each handles different types of input (e.g., Conway & Christiansen, 2006; Goschke, Friederici, Kotz, & van Kampen, 2001). For instance, Conway and Christiansen (2006) used a novel modification of the artificial grammar learning paradigm (Reber, 1967), with participants exposed to sequential patterns from two grammars interleaved with one another. Participants learned both grammars well when the stimuli were in two different sense modalities (vision and audition) or were in two different perceptual dimensions within the same sense modality (colors and shapes or tones and nonsense words). However, when the grammars were instantiated using the same perceptual dimension (two sets of shapes or two sets of nonsense words), participants demonstrated much lower implicit learning performance. These results suggest the possible existence of multiple learning mechanisms that operate in parallel, each over a specific kind of input (tones, speech-like material, shapes, etc.).

A similar conclusion was reached by Goschke et al. (2001). They found that aphasics were impaired on the learning of phoneme sequences but not visual sequences, suggesting the involvement of dissociable domain-specific learning systems. The existence of multiple implicit learning systems may help explain why some studies have demonstrated a link between implicit learning and language and other studies have not: some implicit learning systems (e.g., perhaps those handling phonological patterns) may be more closely involved with language acquisition and processing than others.

The empirical study described below was designed to elucidate some of the complex issues regarding the nature of implicit sequence learning and its involvement in spoken language processing. In the present experiment, we used two versions of the Simon game task – one using color patterns and the other using non-color spatial patterns -- in order to examine possible differences in visual stimuli that can be easily or not easily encoded verbally. We also used a spoken language task under degraded listening conditions. In this way, we were able to assess whether implicit sequence learning that is or is not phonologically-mediated is correlated with spoken language perception under degraded listening conditions. Our hypothesis was that performance on the Simon implicit sequence learning task would be significantly and strongly correlated with performance on the spoken sentence perception task, but only when the Simon task uses stimuli that are easy to encode verbally.

Method

Participants

Twenty undergraduate students (age 18-36 years old) at Indiana University received either monetary compensation or course credit for their participation. All subjects were native speakers of English and reported no history of a hearing loss or speech impairment.

Apparatus

A *Magic Touch*® touch-sensitive monitor displayed visual sequences for the two implicit learning tasks and recorded participant responses.

Stimulus Materials

Spoken Sentence Perception Task. For the language perception task, we used English "SPIN" sentences created by Kalikow et al. (1977) and subsequently modified by Clopper and Pisoni (2006). The sentences varied in terms of the final word's predictability. Three types of sentences were used, 25 of each type: high-predictability (HP), low-predictability (LP), and anomalous (AN). All sentences were 5 to 8 words in length and were balanced in terms of phoneme frequency. HP sentences have a final target word that is predictable given the semantic context of the sentence (e.g., "*Her entry should win first prize*"); LP sentences have a target word that is not predictable given the semantic context of the sentence (e.g., "*The man spoke about the clue*"). On the other hand, AN sentences follow the same syntactic form and use the same carefully constructed set of phonetically balanced words as the HP and LP sentences, but the content words have been placed randomly (e.g., "*The coat is talking about six frogs*").

All 75 sentences were spoken by a single male speaker, a life-time resident of the "midland" dialect region of the United States, whose spoken recordings were chosen from amongst a set of recordings taken from multiple speakers developed as part of the "Nationwide Speech Project" (see

Clopper & Pisoni, 2006). The sentences were then degraded by processing them with a sinewave vocoder (www.tigerspeech.com) that simulates listening conditions for a user of a cochlear implant with 6 spectral channels. All sentences were leveled at 64 dB RMS.

Implicit Learning Tasks. For the sequence learning tasks, we used three different artificial grammars to generate the sequences. Grammar A was taken from Karpicke and Pisoni (2004) while Grammars B and C were from Knowlton and Squire (1996). An artificial grammar is a Markovian finite-state machine that consists of a series of nodes connected by various transitions (see Figure 1). The grammars can generate sequences of various lengths that obey certain rules that specify the order that sequence elements can occur. To use the grammar to generate a sequence, one begins at the arrow marked "start", and traverses through the various states to determine the elements of the sequence, until reaching the "end" arrow. For example, this grammar can generate the sequence: 3-4-3-1.

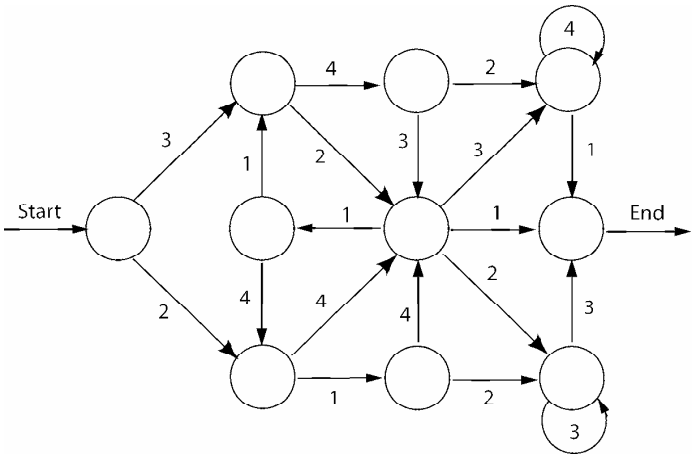


Figure 1. Artificial grammar used to create sequences for the implicit learning tasks. To generate a sequence, the experimenter follows the paths of the grammar and notes the sequence of numbers that are encountered. For instance, the sequence 3-4-2-4-1 is grammatical with respect to this grammar, whereas the sequence 4-1-3-2-2 is not.

We used each grammar to generate 22 unique exemplars (2 exemplars of length 3, and 4 exemplars each of lengths 4-8) that were used for the Learning Phase of the task. Twenty additional exemplars were also generated by each grammar (4 exemplars each of lengths 4-8), for use in the Test Phase. Twenty ungrammatical sequences were also generated for the Test Phase. Ungrammatical sequences were created by taking each grammatical sequence and randomly shuffling the elements that comprise it. For example, the ungrammatical sequence 2-2-3-3 is a randomized version of the Grammar A grammatical sequence 3-2-2-3. Using this method, ungrammatical sequences differ from grammatical sequences only in terms of the *order* of elements within a sequence, not in terms of the actual elements themselves.

Procedure

All participants engaged in three tasks: a spoken sentence perception (SSP) task which occurred under degraded listening conditions; and two visual sequence learning tasks, "Colored-Sequence" (Color-

Seq) and "Non-Colored-Sequence" (Non-Color-Seq). The order of these three tasks varied according to random assignment, but in all cases the SSP task always occurred as the middle of the three tasks.

Spoken Sentence Perception Task. In the SSP task, participants were told they would listen to sentences that were distorted by a computer, making them difficult to understand. Their task was to identify the last word in each sentence and write the word down on a sheet of paper provided to them. Sentences were presented over headphones using a self-paced format. The 75 sentences described above were presented in a different random order for each subject. A written response was scored as correct if the written word matched the intended spoken target word; misspellings (e.g., "valt" instead of "vault") were counted as correct responses.

Implicit Learning Tasks. For the two sequence learning tasks, Color-Seq and Non-Color-Seq, we used a touchscreen version of the Simon game device. Participants were told that they would see visual sequences on the computer screen and would be required to reproduce what they saw using the response panels on the touch screen. Unbeknownst to participants, the sequences were generated according to one of the three artificial grammars previously described. Each sequence learning task consisted of two parts, a Learning Phase and a Test Phase. The procedures for both phases were identical and in fact from the perspective of the subject, there was no indication of separate phases at all. The only difference between the two phases was which sequences were used. In the Learning Phase, the 22 Learning Sequences were presented randomly, two times each. After completing the sequence reproduction task for all of the learning sequences, the experiment seamlessly transitioned to the Test Phase, which used the 20 novel grammatical (G) and 20 ungrammatical (U) Test Sequences.

Sequence presentation consisted of colored (for Color-Seq) or black (for Non-Color-Seq) squares appearing one at a time, in one of four possible positions on the screen (upper left, upper right, lower left, lower right). Each square appeared on the screen for a duration of 700 msec, with a 500 msec ISI. For Color-Seq, the four elements (1-4) of each grammar were randomly mapped onto each of the four screen locations as well as four possible colors (red, blue, yellow, green). The assignment of grammar element to position/color was randomly determined for each subject; however, for each subject, the mapping remained consistent across all trials. Likewise, for Non-Color-Seq, the four elements of each grammar were mapped onto each of the four screen locations, randomly determined for each subject. The spatial mapping in this condition also remained invariant for a given subject.

Each element of a sequence was presented for 700 msec and was separated from the next element by 500 msec of blank screen. After the entire sequence had been presented, there was a 2000 msec delay and then five panels appeared on the touch screen to signify the beginning of the response phase. Four of those panels were the same-sized and same-colored as the four locations that were used to display each sequence. The squares were appropriately colored (red, green, blue, and yellow for Color-Seq and all black for Non-Color-Seq). The fifth panel was a long horizontal bar placed at the bottom of the screen, which acted as the equivalent of the "Enter" button. The subject's task was to watch a sequence presentation and then to reproduce the sequence they saw by pressing the appropriate buttons in the correct order as dictated by the sequence. When they were finished with their response, they were instructed to press the long black bar at the bottom, and then the next sequence was presented after a 2-sec delay.

Participants were not told that there was an underlying grammar for any of the Learning or Test sequences, nor that there were two 2 types of sequences in the Test phase. From the standpoint of the participant, the task in Color-Seq and Non-Color-Seq was solely one of observing and then reproducing a series of unrelated sequences. Finally, following the experiment, all participants filled out a debriefing

form that asked whether they used a verbal strategy when doing the Non-Color-Seq task, such as verbally coding the four different locations in terms of numbers "one", "two", etc.

Results

For the SSP task, subjects accurately perceived target words in HP sentences (M=18.2) significantly more often than LP or AN sentences (M=12.9 and 13.3, respectively): HP vs. LP, $t(19) = 10.8$, $p < .001$; HP vs. AN, $t(19) = 7.1$, $p < .001$.

For Color-Seq and Non-Color-Seq, a sequence was scored correct if the participant correctly reproduced the sequence in its entirety. Span scores were calculated using a weighted method, in which the total number of correct sequences at a given length was multiplied by the length, and then scores for all lengths added together. We calculated separate span scores for grammatical and ungrammatical test sequences for each subject. Performances on the two sequence learning tasks are shown in Table 1, which depicts weighted span scores for grammatical (G) and ungrammatical (U) sequences.

	Sequence Type					
	G		**U**		**LRN**	
Sequence Task	***M***	***SE***	***M***	***SE***	***M***	***SE***
Color-Seq	64.9	5.13	56.4	5.77	8.55	4.62
Non-Color-Seq	55.3	5.70	43.9	4.35	11.5	3.08

Table 1. Weighted span scores for grammatical (G) and ungrammatical (U) sequences, as well as the difference between G and U (LRN), which is a measure of learning.

A 2x2 ANOVA contrasting Task (Color-Seq vs. Non-Color-Seq) and Sequence Type (grammatical vs. ungrammatical) revealed a main effect of Task ($F(1, 76) = 4.4$, $p < .05$) and a marginal main effect of Sequence Type ($F(1, 76) = 3.6$, $p = .061$) and no significant interaction. These results indicate that overall, participant's span scores were better for the Color-Seq task, which is not surprising considering that the Color-Seq task has an extra cue (color) over and beyond the spatio-temporal cues available in the Non-Color-Seq task. The marginal effect of Sequence Type indicates that participants had higher span scores for the grammatical sequences and thus suggests that overall, participants showed implicit learning of the underlying grammatical regularities in the sequence patterns.

For each subject, we also calculated the difference between G and U on each task, which served as a measure of implicit learning (LRN; see Table 1). To confirm that learning occurred in both tasks, we compared the LRN scores to chance levels using one-tailed t-tests. Both comparisons were statistically significant (Color-Seq: $t(19) = 1.85$, $p < .05$; Non-Color-Seq: $t(19) = 3.72$, $p < .001$), indicating that participants in both tasks on average showed implicit learning for the grammatical regularities of the sequences, demonstrated by having better memory spans for test sequences that were consistent with the

grammars used during the learning phase. Finally, we compared the two LRN scores between tasks and found no differences between them, $t(19) = .60$, $p = .56$.

We next investigated the size of the learning effect for individual subjects. Although on average, subjects showed a learning effect, there was wide variation in LRN scores across these two tasks (Seq-Color: -18 to 71; Non-Color-Seq: -14 to 33). Because of the variability in the scores, it is possible to determine to what extent individual differences in implicit learning abilities for sequential patterns correlates with spoken sentence perception under degraded listening conditions.

To assess the relations between implicit sequence learning and spoken language perception, we computed correlations among the following dependent measures: HP, LP, AN, Color-Seq grammatical (C-G), Color-Seq ungrammatical (C-U), Color-Seq LRN (C-LRN), Non-Color-Seq grammatical (NC-G), Non-Color-Seq ungrammatical (NC-U), and Non-Color-Seq LRN (NC-LRN). If probabilistic sequence learning is an important underlying source of variance that contributes to spoken language perception, we would expect that the LRN scores will be strongly correlated with the spoken sentence perception scores.

The correlation analyses, shown in Table 2, revealed several interesting patterns. None of the G and U scores correlated significantly with the SSP scores. However, as expected, the LRN scores, which measure implicit learning of the underlying sequence patterns, revealed a different pattern altogether. The results showed that LRN for Color-Seq correlated significantly with HP ($r = .48$, $p < .05$) and LP ($r = .56$, $p < .01$) but not with AN ($r = .36$, $p = .12$), whereas LRN for Non-Color-Seq did not correlate significantly with any of the SSP measures (r's $< .38$). Moreover, neither of the two LRN scores correlated significantly with one another ($r = .26$, $p = .28$)[2].

Measure	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1.HP	--	**.83**	**.60**	.26	-.2	**.48**	.01	-.2	.33
2.LP		--	.39	.29	-.2	**.56**	.03	-.2	.28
3.AN			--	.37	.01	.36	.30	.13	.38
4.C-G				--	**.65**	.30	**.61**	.42	**.53**
5.C-U					--	**-.5**	**.52**	**.49**	.27
6.C-LRN						--	.03	-.1	.26
7.NC-G							--	**.85**	**.66**
8.NC-U								--	.15
9.NC-LRN									--

Table 2. Correlations between dependent measures for the sentence processing and implicit sequence learning tasks (see above text for abbreviations). Significant correlations at $p < .05$ are in bold; those at $p < .01$ are also underlined.

[2] With a sample size of n=20, there is only enough power to identify "large" correlation/effect sizes (Cohen, 1988); thus, a non- significant correlation in this data may not signify no correlation at all, but it does suggest that if a correlation exists, it is substantially weaker than the significant effects reported here.

Additionally, we ran a principal component analysis (PCA) on all nine variables to reduce the data set to a smaller set of components. The results of the analysis revealed two components that explained 69% of the total variance. Interestingly, the second component (31.4% of total variance) includes HP, LP, and Color-Seq LRN, whereas the first component (37.6% of total variance) includes the six other DV's.

In sum, the results can be summarized as follows. First, participants on average showed implicit learning in both the Color-Seq and Non-Color-Seq task, as demonstrated by the LRN scores being statistically greater than zero. Second, only LRN for Color-Seq, but not Non-Color Seq, was significantly correlated with the high (HP) and low probability (LP) sentences in the SSP task; neither LRN scores were correlated with the anomalous (AN) sentences. Finally, a PCA analysis showed that HP, LP, and LRN for Color-Seq all loaded on a common component. These data suggest a strong link between visual implicit sequence learning and spoken language processing abilities.

Discussion

Our hypothesis was that participants' abilities on a visual, implicit sequence learning task, especially one that incorporated stimuli that could be easily encoded verbally, would be correlated with their performance on a spoken sentence perception task under degraded listening conditions. Building on previous empirical and theoretical work suggesting that spoken language processing depends upon domain-general implicit sequential learning skills, our results provide the first empirical demonstration of individual variability in implicit learning performance correlating with language processing in typically-developing subjects. The results are particularly striking given that the sequence learning and language tasks involved stimuli in two different sensory modalities (vision and audition, respectively).

A few observations are important to highlight. First, performance on the SSP task was not correlated with span scores for G or U sequences. That is, the contribution to language processing that we have demonstrated is not due merely to serial recall abilities. It was only when we assessed how much memory span *improved* for grammatically-consistent sequences did we find a significant correlation. Thus, it is the ability to extract knowledge about structured sequential patterns over a set of sequences that is important, not just the ability to encode and recall a sequence of items from memory.

Second, performance on the Color-Seq task correlated much more strongly with the high (HP) and low (LP) predictable sentences compared to the anomalous (AN) sentences. To do the HP (and to a lesser extent, LP) sentence perception tasks successfully, the listener needs to use the context of the preceding material in the sentences to help predict and identify the final target word. This sequential context is not available for the AN sentences because they were semantically anomalous. In turn, successful performance on the Color-Seq task also requires sensitivity to sequential, probabilistic context. That is, the greater one's sensitivity to sequential structure in the grammatical sequences, the better chance one has of correctly recalling a novel grammatical sequence that contains the same kind of probabilistic structure. Thus, we believe we have identified a key link between implicit sequence learning and spoken language perception: *both require the ability to acquire and use probabilistic information distributed across temporal patterns*.

Third, we note that only the Color-Seq task, not the Non-Color-Seq task, was correlated with SSP. From a procedural standpoint, the only difference between Color-Seq and Non-Color-Seq was that the Color-Seq task included not only spatiotemporal information, but also the presence of color cues. One account of these differences is that the sequences from the Color-Seq task are very readily verbalizable

and codable into phonological form (e.g., "Red-Blue-Yellow-Red") whereas those from the Non-Color-Seq task are not. Thus, Color-Seq but not Non-Color-Seq might involve implicit learning of phonological representations, and it could be this basic learning ability that contributes to success on the SSP task.

To examine this prediction further, we used the post-experiment debriefing questionnaire to identify 12 participants ("phonological coders") who attempted to encode sequences in the Non-Color-Seq task using some kind of verbal code, such as labeling each of the four spatial positions with a digit (1-4). The remaining 8 subjects ("non-phonological-coders") indicated they did not use a verbal code during the task. We assessed correlations between these two groups' LRN scores and SSP measures and found that although none of the correlations quite reached statistical significance (presumably due to a lack of statistical power), the difference in the correlations between the two groups was quite striking: phonological coders' performance on the sequence task correlated with HP (r = .43), LP (r = .28), and AN (r = .44) whereas the correlations for non-coders were r = -.31 for HP, r = -.17 for LP, and r = .14 for AN.

Thus, for those participants who explicitly used a phonological-coding strategy on the Non-Color-Seq task, their performance was positively correlated with SSP task performance, whereas for participants who did not use such a strategy, their performance was much less or even negatively correlated with SSP task performance. Although not statistically significant at this time, this pattern of results for the Non-Color-Seq task may suggest that a crucial aspect of implicit sequence learning that contributes to spoken language processing is the learning of structured patterns from sequences that can be easily represented using a verbal code.

To summarize, we believe the evidence points to an important factor underlying spoken language processing: the ability to implicitly learn complex sequential patterns, and perhaps especially those that can be represented phonologically. Using a visual implicit sequence learning task, we found that sequence learning performance correlated with performance on a spoken sentence perception task requiring one to capitalize on sequential context. These results suggest a strong link between implicit sequence learning and spoken language processing and not only provide important new theoretical insights, but also have practical implications regarding the nature of language processing in both typical and clinical populations.

References

Baddeley, A.D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189-208.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences, 2*, 406-416.

Clopper, C.G. & Pisoni, D.B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication, 48*, 633-644.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Conway, C.M. & Christiansen, M.H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science*, *17*, 905-912.

Conway, C.M., & Christiansen, M.H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, 5, 529-546.

Dominey, P.F., Hoen, M., Blanc, J.-M., & Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies. *Brain and Language, 86*, 207-225.

Gómez, R.L. & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178-186.

Goschke, T., Friederici, A. D., Kotz, S. A., & van Kampen, A. (2001). Procedural learning in broca's aphasia: Dissociation between the implicit acquisition of spatio-motor and phoneme sequences. *Journal of Cognitive Neuroscience, 13*(3), 370-388.

Grunow, H., Spaulding, T.J., Gómez, R.L., & Plante, E. (2006). The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *Journal of Communication Disorders, 39*, 158-170.

Howard, J.H., Jr., Howard, D.V., Japikse, K.C., & Eden, G.F. (2006). Dyslexics are impaired on implicit higher-order sequence learning, but not on implicit spatial context learning. *Neuropsychologia, 44*, 1131-1144.

Kalikow, D.N., Stevens, K.N., & Elliott, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, *61*, 1337-1351.

Karpicke, J. D., & Pisoni, D. B. (2004). Using immediate memory span to measure implicit learning. *Memory & Cognition, 32(6)*, 956-964.

Kelly, S.W., Griffiths, S., & Frith, U. (2002). Evidence for implicit sequence learning in dyslexia. *Dyslexia, 8*, 43-52.

Knowlton, B.J. & Squire, L.R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*, 169-181.

Lashley, K.S. (1951). The problem of serial order in behavior. In L.A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112-146). New York: Wiley.

Menghini, D., Hagberg, G.E., Caltagirone, C., Petrosini, L., & Vicari, S. (2006). Implicit learning deficits in dyslexic adults: An fMRI study. *Neuroimage, 33,* 1218-1226.

Pisoni, D.B. & Cleary, M. (2004). Learning, memory, and cognitive processes in deaf children following cochlear implantation. In F.G. Zeng, A.N. Popper & R.R. Fay (Eds.), *Springer handbook of auditory research: Auditory prosthesis*, SHAR Volume X (pp. 377-426).

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Behavior, 6*, 855-863.

Rubenstein, H. (1973). Language and probability. In G.A. Miller (Ed.), *Communication, language, and meaning: Psychological perspectives* (pp. 185-195). New York: Basic Books, Inc.

Rüsseler, J., Gerth, I., & Münte, T.F. (2006). Implicit learning is intact in developmental dyslexic readers: Evidence from the serial reaction time task and artificial grammar learning. *Journal of Clinical and Experimental Neuropsychology, 28*, 808-827.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Stadler, M.A. & Frensch, P.A. (Eds.) (1998). *The handbook of implicit learning*. London: Sage Publications.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*, 231-270.

Vicari, , S., Marotta, , L., Menghini, D., Molinari, M., & Petrosini, L. (2003). Implicit learning deficit in children with developmental dyslexia. *Neuropsychologia, 41*, 108-114.

Waber, D. P., Marcus, D. J., Forbes, P. W., Bellinger, D. C., Weiler, M. D., Sorensen, L. G., et al. (2003). Motor sequence learning and reading ability: Is poor reading associated with sequencing deficits? *Journal of Experimental Child Psychology, 84*, 338-354.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Cochlear Implant Simulations: A Tutorial on Generating Acoustic Simulations for Research[1]

Jeremy L. Loebach

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work supported by NIH-NIDCD Training Grant T32-DC00012 and NIH-NIDCD Research Grant R01-DC00111. I would like to thank Dr. Qian-Jie Fu for developing and maintaining TigerCIS, a tool that has become invaluable to my research.

Cochlear Implant Simulations: A Tutorial on Generating Acoustic Simulations for Research

Abstract. Acoustic simulations of cochlear implants have become a common tool available to researchers interested in many aspects of speech perception and cognition. Although many ways exist to create such simulations, all are derived from a common philosophical and physiological base. The primary goal of this tutorial is to instruct the reader how to make cochlear implant (CI) simulations for use in research. In order to realize this goal, the various methodologies and signal processing techniques will be reviewed, since the varied techniques used will determine the behavioral outcomes. Finally, the reader will be led through a systematic demonstration using a stand-alone simulator (TigerCIS).

Introduction

Although explicit cochlear implant simulations have been around for only about a decade, the technology and signal processing techniques that gave rise to them have existed for the better part of a century. Most modern simulations are based on the principle of the vocoder, pioneered by Homer Dudley at Bell Labs (Dudley, 1939). The vocoder was a speech synthesizer that passed the acoustic signal through a series of band pass filters, which derived the energy profiles for each band. The spectrum of each band was replaced with a synthetic source (tones and noisy hisses) which were then modulated using the energy profile appropriate for the original band. The result was a stimulus that sounded highly artificial, but was surprisingly intelligible as speech. Most modern CI simulations are based, in part, on the philosophy of Dudley's vocoder.

Although multi-channel synthesis techniques have been around since the 1930's, the development of multi-channel cochlear implants did not occur until far later. Research into the electrical stimulation of the cochlear nerves has had a long history in both animal and human models. Experiments with single electrodes inserted into the cochlear partition had been ongoing since the early 1960's, but the development of single channel cochlear implants for humans did not occur until the early 1970's, finally gaining FDA approval in the United States in 1984. While much of the research into single channel implants occurred in the U.S., the development of multi-channel implants first occurred in Australia. In fact, in 1985, the Cochlear multi-channel implant received U.S. FDA approval, barely a year after the approval of the House/3M single channel implant. This effectively put an end to single channel implants, though there are still advocates of the single channel short electrode implants (c.f. House, 1994).

Today, only 3 companies have FDA approval to produce and market cochlear implants in the United States: Med-EL, Cochlear (makers of the Nucleus series of implants) and Advanced Bionics (makers of the Clarion series of implants). Other models may be available for experimental purposes and clinical trials, and other brands do exist in European and other foreign markets. The basic design of each implant is similar, although different brands use different numbers of electrodes and different signal processing techniques to provide electrical stimulation. Most currently available implants contain six to twenty-four channels. Theoretically, increasing the number channels that are available in the electrode array will increase the amount of acoustic information that will be available to the user. Practically, however, more channels do not necessarily translate to better performance due to surgical, anatomical, and physiological constraints.

How a Cochlear Implant Works

Cochlear implants rely on the anatomy and physiology of the cochlea for their functionality. In the normal hearing ear, vibrations in the air are translated into pressure waves in the fluid of the cochlea. The basilar membrane is differentially displaced by these pressure waves depending on their frequency. The width, thickness and stiffness of the basilar membrane (BM) vary depending on longitudinal position. At the base of the cochlea (closest to the stapes) the BM is thin, narrow and tight requiring high frequency oscillations to displace it. At the apex (farthest from the stapes), the BM is wide, thick and loose, requiring low frequency oscillations to displace it. The mode of basilar membrane displacement bears some similarity to the strings of a guitar. Thin tight strings like the high E vibrate very fast to produce high-pitched sounds, whereas thick loose strings like the low E vibrate very slowly to produce low pitch sounds. The displacement patterns of the BM essentially work in reverse: high frequency sounds best displace the BM at the base, whereas low frequency sounds best displace the BM at the apex.

On top of the basilar membrane sits the organ of Corti, which contains the hair cells necessary for transduction. The cell bodies of the hair cells are embedded in the surrounding tissue so that only the stereocilia are exposed to the fluid inside the organ of Corti. These stereocilia move back and forth with the motion of the fluid (like seaweed moving back and forth with ocean waves) opening mechanically gated ion channels. The influx of potassium ions (K+) changes the resting potential of the cell, stimulating the release of neurotransmitter onto the neurons of the spiral ganglion at the base of the hair cell, effectively transducing the sound into neural impulses. The neurons of the spiral ganglion (known by many different names: auditory nerve, cochlear nerve, vestibulocochlear nerve, VIII (8th) cranial nerve) then carry the electrical impulses to the brainstem for processing.

Although the hair cells themselves do not have a frequency preference, they are situated along the length of the basilar membrane and will be displaced in a frequency dependent fashion. This means that the specific hair cells, and consequently the neurons of the spiral ganglion, will respond to specific frequencies. If the apical region of the BM is displaced, indicating a low frequency sound, the hair cells overlying this region of the BM will be stimulated, transmitting information to the brain that a low frequency sound occurred. If, on the other hand, the basal region of the BM is displaced indicating a high frequency sound, the hair cells overlying this region of the BM will be stimulated, sending information to the brain that a high frequency sound occurred. Thus, one could examine the responses of the neurons in the spiral ganglion, and based on the frequency of tones they respond best to, trace them back to the region of the cochlea and indeed the hair cells that stimulated them. This is a very important property of the spiral ganglion: it has a tonotopic organization that arises from the location of the hair cells in the organ of Corti. Due to the law of specific nerve energies, one could stimulate a single neuron in the spiral ganglion and produce the sensation of a tone of a particular pitch in the brain. This is the fundamental mechanism that cochlear implants exploit.

In sensorineural hearing loss, the normal processes of transduction are disrupted. While a myriad of etiologies can lead to specific deficits in the mechanisms of transduction, one of the most common causes of deafness occurs when the hair cells are damaged or destroyed. Since hair cells cannot regenerate, they can no longer stimulate the neurons of the spiral ganglion, so information about sound cannot get into the brain. Despite the destruction of their primary sources of input, the neurons of the spiral ganglion are far more robust, and typically do not atrophy following cochlear insult. This is a key mechanism that cochlear implants utilize: although the mechanisms of transduction have failed, the neurons of the spiral ganglion are intact, and can be electrically stimulated. The electrode array of the cochlear implant stimulates the surviving spiral ganglion neurons in order to produce the sensation of

sound in the brain. More importantly, cochlear implants do not rely on extensive pre-processing; rather they simply provide an alternate form of input, allowing the ascending auditory pathway to function as it normally would.

An electrode array inserted into the round window of the cochlea brings electrical contacts close to spiral ganglion neurons. Pulses of electrical current stimulate the neurons to produce a sensation of sound in the brain. An external microphone takes in the sound, and sends it to a speech processor, which divides the acoustic information into a number of channels, changing the energy in each channel into a digital code. The coded signal is then transmitted through the skin to a receiver implanted behind the ear, which in turn sends the information to the appropriate electrodes in the cochlea as a series of current pulses. Electrodes are organized tonotopically, so that high frequency channels are located basally, and low frequency channels are located apically. Thus, the electrical impulses are delivered to the appropriate tonotopic region of the cochlea to evoke the sensation of the appropriate pitch. In the normal hearing ear, a single hair cell will stimulate a limited number of spiral ganglion neurons (typically 3-7), resulting in a very discrete encoding of frequency. Such discrete stimulation is impossible with modern implants, however, since there is some distance between the stimulation site and the target neurons. Due to spread of electrical current in the fluid and tissue of the cochlea, many more neurons are stimulated by each electrode than would be stimulated in the intact cochlea. This causes a widening of the frequency representation, which removes much of the spectral fine structure and frequency resolution of the normal auditory signal. Despite such spectral reduction, many modern cochlear implant users demonstrate very high levels of speech recognition under quiet conditions, suggesting that the limited frequency information provided by electrical stimulation is sufficient to transmit information about speech.

Acoustic Simulations of Cochlear Implants

Although no one with normal hearing can truly know what the world sounds like through a cochlear implant, simulations can approximate the experience by processing the acoustic signal in a manner similar to that of an implant's speech processor. There are three main components to such simulations: the frequency channel, the amplitude envelope and the carrier signal. Each of these can be altered to produce differential effects in perception.

The Channel. Since cochlear implant processors divide the acoustic information into a limited number of channels, simulations typically filter the acoustic signal into different frequency bands using band pass filters. The number of frequency bands that are used depends on the type of implant being modeled. For example, one-band simulations can be used to model single channel implants (Van Tasell, Soli, Kirby, & Widin, 1987), and multiple band simulations can be used to model multi-channel implants (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995; Fu & Shannon, 1999). The band pass filters are typically broad in order to limit the spectral detail in manner similar to that of a cochlear implant, and the filter bandwidth will ultimately depend on the total number of channels used (Figure 1). In single channel simulations, the band pass filter will typically cover most of the frequency range of the stimulus. In multiple channel simulations, the band pass filters are as broad as necessary to cover the range of the spectrum. In some cases, the center frequencies are selected based on known measurements such as the articulation index (AI) (French & Steinberg, 1947), or locations of critical bands in the cochlea (Greenwood, 1961; Greenwood, 1990). Channel selection is not limited by any known function, so the cutoff frequencies may vary according to the task one wishes to design. The overall selection of cutoff frequencies and width of the band pass filters should approximate reasonable values for a cochlear implant processor in order to be a maximally applicable model, however. Specific models can be made to simulate particular implant processors (CIS processor, SPEAK processor) or even specific implantees (based on their frequency sensitivity).

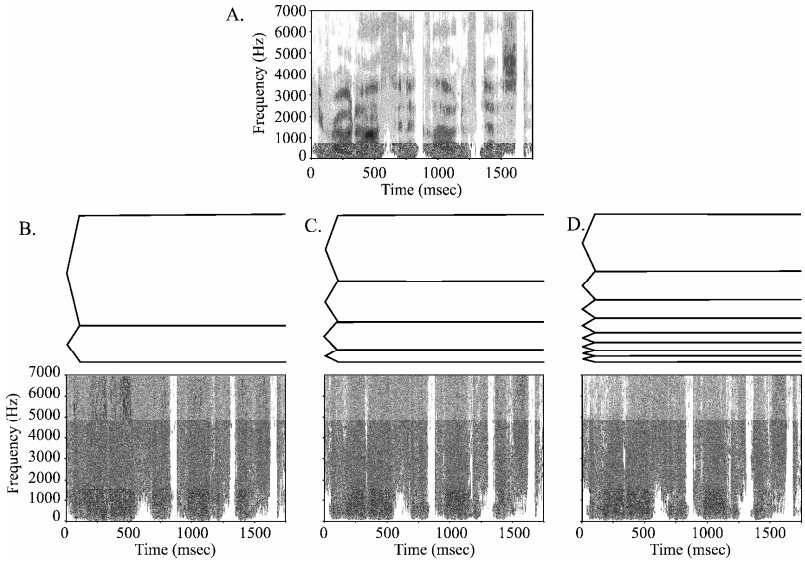


FIGURE 1: Band pass filtering allows one to simulate the limited number of channels available in a cochlear implant. The spectrogram for the naturally produced version of the sentence "He rode off in a cloud of dust" as produced by a male speaker is presented in (A). The band pass filters required to generate noise vocoded stimuli with 2 (B), 4 (C) and 8 (D) spectral channels and the spectrographic representation of the stimuli generated by them appear below. Although the majority of the spectral detail is absent from the noise vocoded stimuli, the overall spectral profiles become increasingly similar to the naturally produced stimulus as the number of bands increases from two to eight. Changes in the temporal information, however, remain unchanged across the noise vocoded stimuli.

The Envelope. Although the output of a cochlear implant is spectrally degraded, it is temporally precise. This means that the temporal modulations are not limited by any known mechanism. Typically, individuals with cochlear implants can discriminate temporal modulations upwards of 300 Hz (i.e. modulations changing at a rate of 300 times per second) (Fu & Shannon, 2000). The amplitude envelope can be thought of as a trace around the time domain waveform of a sound, and is extracted by using a low pass filter (Figure 2). How closely you trace the sound will depend on the frequency cutoff of the low pass filter. If you choose a filter with an upper frequency limit of 160 Hz, only the changes in amplitude that occur at a rate of less than 160 times per second will be preserved. Envelope derivation will also determine how well subjects perform; if you make the cutoff too low, you will undermine performance (see Rosen, 1992 for a review of the temporal cues in the envelope).

Whether modeling single or multi channel implants, the amplitude envelope must be derived from each band. One must first divide the spectrum into frequency bands using the band pass filters, and then use a low pass filter to derive the amplitude envelope from each. This means that if one were designing a 22-channel simulation, one would filter the signal into 22 channels and then extract the amplitude envelope from each. Since the energy in the spectrum will change depending on the stimulus (e.g. as is the case for consonants and vowels, where a high frequency fricative may be followed by a low frequency vowel), deriving the envelope for each band will preserve the energy in the appropriate spectral regions.

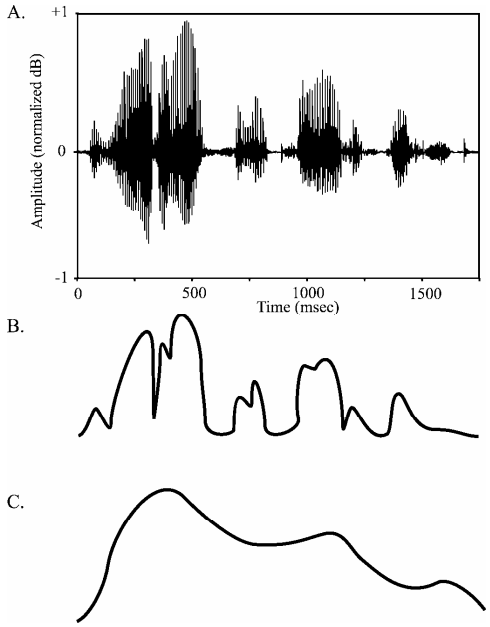


FIGURE 2: Envelope detection is based on low pass filtering the time domain waveform of the stimulus (A). The amount of temporal information preserved depends on the upper limit of the low pass filter. Higher frequency cutoffs (~300 Hz) preserve smaller changes in the time intensity information (B) than lower frequency cutoffs (~25 Hz), which only preserve course temporal changes (C). Low pass filters with sufficiently high cutoff frequencies can preserve rudimentary pitch information, if the pitch falls within the pass band.

The Carrier. Although the spectrum has been filtered into a discrete number of bands, the residual spectral information must be removed to make an effective CI simulation. This is typically done by replacing the spectral content of each band with some other signal. Two carriers have been traditionally used: white noise and sinusoids (Figure 3). White noise has proven to be a successful carrier signal for CI simulations, and the earliest pioneering work with CI simulations typically used noise based carriers to model single channel implants (Van Tasell, Soli, Kirby & Widin, 1987; Van Tasell, Greenfield, Logemann & Nelson, 1992) and multi channel implants (Blamey, Dowell, Tong, Brown, Luscombe & Clark, 1984a; Blamey, Dowell, Tong & Clark, 1984b; Shannon, Zeng, Kamath, Wygonski & Elekid, 1995). Noise is an effective way to remove the spectral detail from the frequency channels, but it may over-represent the information in the band. Recall that the electrodes in the cochlear implant do not stimulate all neurons in a given region of the cochlea evenly. If the stimulation patterns were reflective of the noise-based carrier then each electrode would be very broad, and each channel would be contiguous, with neither gaps nor overlaps between each. Physiologically, this may not be the case, since the electrodes are typically narrow, and stimulation is more likely to be focused at a given frequency, and roll off to both sides due to current spread and electrical diffusion. As such, other simulations have opted to use sinusoids as their carrier signals (Dorman & Loizou, 1997; Dorman & Loizou, 1998; Loizou Dorman, & Tu, 1999). Each sinusoid is focused at the channel center frequency, and rolls off in intensity with a given slope on either side. Indeed, some evidence for using sinewaves as carriers come from CI users themselves, who have described the sensation of electrical stimulation as being a series of beep tones rather than noise pulses (Dorman, Loizou & Rainey, 1997).

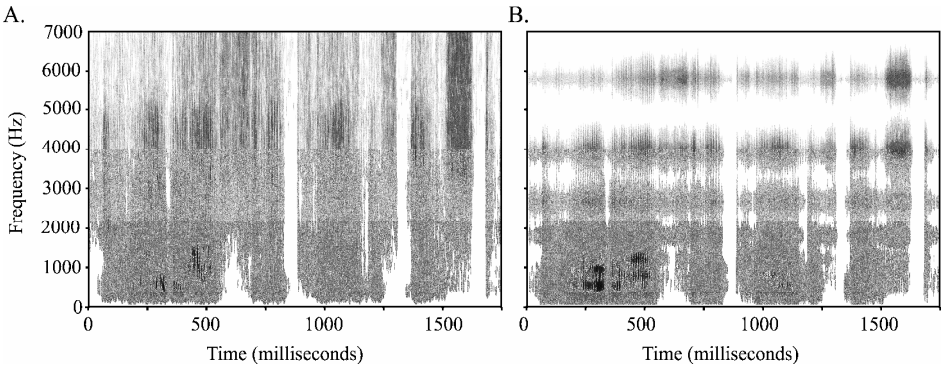


FIGURE 3: Spectrograms of the sentence "He rode off in a cloud of dust." produced using noise band carriers (A) and sinusoidal carriers (B). Although the stimuli sound qualitatively distinct, and appear to differ in overall spectral profile, studies have demonstrated that speech recognition in quiet does not appear to differ between stimuli produced with noise band and sinewave carriers.

No matter which type of carrier signal is selected, speech recognition under each type of stimulation is virtually identical (Dorman, Loizou & Rainey, 1997). Although sinusoids may offer a more realistic simulation, and indeed people with cochlear implant report that the world sounds more metallic than noisy (c.f. Chorost, 2005), the behavioral results are identical whether using noise or sinusoids, at least for speech in quiet.

Software. There are many different methods of simulating cochlear implants. Applications exist for MatLAB and stand-alone programs such as TigerCIS have also been developed for use. Any program that is capable of band and low pass filtering and can generate noise and pure tone sinusoids can be used (indeed I have done this all by hand using a Cool Edit Pro, MS Excel, and MS Notepad, see Loebach, 2005). The basic concept is simple: divide the spectrum into bands, derive the envelope from each band, replace the spectrum in each band with your carrier signal, modulate the carrier with the amplitude envelope appropriate for that band and reassemble the bands. Although similar in concept, the exact steps that you will take will depend on the specific program that you use.

One of the easiest simulators to use is TigerCIS, which was developed by Dr. Qian-Jie Fu at the House Ear Institute. It produces quality simulations using either noise band or sinewave carrier signals, and is very user friendly. It is also very malleable, allowing one to change the settings for the amplitude envelope derivation, band pass filtering, number of channels, and output filters allowing one to "shift" the electrodes in any manner. Moreover, TigerCIS is freely available on the internet at http://www.tigerspeech.com/.

How To

TigerCIS can process any Windows PCM .wav file of a standard format. Typical formats include 22,050 Hz, 16-bit stereo or mono files. If you are editing your wav files in Adobe Audition or Praat, be sure that no additional proprietary information is being placed in the header, as this will cause the file to be unreadable in TigerCIS. Occasionally, errors will be introduced if each track of the stereo .wav file contains slightly different information. This can result in a doubling of the frequency information, and an apparent upward shift in pitch. Anecdotally, I have found that mono .wav files yield the best results, and avoid processing artifacts. In addition, it is critical to level the stimuli prior to processing them with the simulator. The simulator will process whatever information is present, and if the stimuli have different

RMS amplitude values, you may under specify some portions of the signal, and over specify others. Ensuring that the stimuli are at a comparable level prior to simulation will provide you with the most accurate and appropriate simulations. The systematic demonstration below was generated using TigerCIS version 1.04.03. Although the software is frequently updated with additional features, the basic structure of the following steps will be preserved in some manner in most versions.

Step 1: Select Processor This step allows you to specify which type of processor you want to use. For a standard CI simulation, be sure the "Noise or sinewave vocoder" box is checked, and the "FFT-based" box is un-checked.

Step 2: Select Carrier Type for Vocoder This step allows you to specify the carrier that you wish to use, either sinusoids centered at the bands center frequency or noise bursts.

Step 3: Select Number of Channels This step allows you to set your analysis and synthesis channel filters. Although the number of channels is unlimited, the more that you use, the longer it will take to synthesize the stimuli. If you want a straightforward CI simulation the number of channels simulated should match the number of spectral channels. By varying the degree and direction of mismatch, one can make compressive simulations (the number synthesized is smaller than the number of spectral channels) or expansive (the number synthesized is larger than the number of spectral channels) simulations.

Step 4: Set Analysis Filter Type This step determines the method used to divide the spectrum. The default setting is based on Greenwood's function (c.f. Greenwood, 1990), which is essentially a pitch to distance map of the basilar membrane based on critical bandwidth. You can create your own mapping function as well, if you want to model a specific type of processor with specific band cutoffs. You can even program your own simulator based on patient data using their frequency channel cutoffs and electrode mappings.

Step 5: Set Analysis Filter Variables This step runs hand and hand with the previous. If you chose a custom filter in step 4, you should load that file here. If not, you can set your lower frequency cutoff (the lowest frequency information that you want to include), upper frequency cutoff (the highest frequency information that you want to include) and the roll off function (dB/octave) which tells the program how sharp to make the filters. For example, if you have a center frequency of 100 Hz and you roll off at 24 dB/octave, it means that as you increase the frequency by 1 octave (from 100 to 200 Hz in this case) the signal strength would drop by 24 dB. This avoids the summation of information where the band pass filters overlap (which would distort the signal, possibly causing clicks). The larger this number is the faster you roll off, and the steeper your filter slopes will be. If your roll off is too slow, channel overlap and summation can occur. If your roll off is too fast you may get transients and other distortions.

Step 6: Envelope Detection This step allows you to design the low pass filter to derive the amplitude envelope from each band. The first number is the cutoff frequency for the upper limit. This number should ideally be below 400 Hz, since this is the upper limit detectible by individuals with cochlear implants. If your low pass cutoff is too low (6-12 Hz) you may not adequately be representing the temporal information available to the cochlear implant user. The roll off function is similar to that used in the previous step and determines how quickly the information drops off as you increase in frequency. Select the roll off carefully, since it will affect the intelligibility (too shallow and you can include more information than is available in a CI).

Step 7: Set Carrier Filter Type This step is the same as in Step 4, except you are now selecting how you want to process the carrier signal. If you are using noise, this will divide the noise spectrum into bands in

a manner similar to the analysis filters in Sep 4. You can either use the Greenwood function, or model a specific implant or patient by specifying a custom filter.

Step 8: Set Carrier Filter Variables This step is the same as in Step 5, except you are now specifying the frequency range for the carrier signal. You can use values similar to or different from the filters used in Step 5, depending on what type of simulation you wish to create. If you want to make a 1:1 simulation, the frequency range should be identical to that in Step 5. If you want to simulate a basal shift of 2 millimeters based on electrode insertion depth, for instance, you would simply change the upper and lower frequency limits by a function that would equate a 2 mm shift along the basilar membrane. Since the tonotopic organization of the basilar membrane is non-uniform across frequency, the lower frequency cutoff would shift upward by about 180 Hz, and the upper frequency cutoff would shift upward by about 1800 Hz. See Zwicker for a more complete treatment (Zwicker, Flottorp & Stevens, 1957).

Step 9: Processing Now that all of the information has been specified, it is time to process the stimuli. Stimuli can be processed individually by loading a single wave file at the top, or in a batch by specifying input and output directories at the bottom. For processing single stimuli simply press the "Start Simulation" button to begin. You can then press "Play Simulation" to listen to what you just created, or "Save Simulation" to save it as a .wav file. Each time you update any of the information in Steps 1-8 you will have to reprocess the stimuli with the current settings (TigerCIS will maintain the previous settings until you exit the program).

Although there are several other useful options available in TigerCIS, the ones presented here are necessary for making standard cochlear implant simulations. The program is very malleable, allowing you to make your own creations (or even to make real time simulations of your own voice), and an online forum is available at http://www.tigerspeech.com/ for support.

Conclusions

Modern cochlear implant simulations may allow the research scientist some insight into the perceptual experiences of a person with a cochlear implant. One must never assume, however, that they are in fact reproducing those actual perceptual experiences. To that end, the closer the variables used to design the simulations are to the specifications of real cochlear implants, the more readily the results in normal hearing subjects will generalize to cochlear implant users themselves. One should be thoughtful in the selection of synthesis variables, and be able to defend their choices in publication. Whenever possible, results should be normalized by comparing findings in normal hearing subjects with those from cochlear implant users, so that a maximum benefit be can achieved for researcher, clinician and patient alike.

References

Blamey, P.J., Dowell, R.C., Tong, Y.C., Brown, A.M., Luscombe, S.M. & Clark, G.M. (1984a). Speech processing studies using an acoustic model of a multiple-channel cochlear implant, *Journal of the Acoustical Society of America, 76*, 104-110.

Blamey, P.J., Dowell, R.C., Tong, Y.C. & Clark, G.M. (1984b). An acoustic model of a multiple-channel cochlear implant, *Journal of the Acoustical Society of America, 76,* 97-103.

Chorost, M. (2005). *Rebuilt: How becoming part computer made me more human*. Houghton Mifflin: New York.

Dorman, M.F. & Loizou, P.C. (1997). Speech intelligibility as a function of the number of channels of stimulation for normal-hearing listeners and patients with cochlear implants, *American Journal of Otology, 18,* S13-S114.

Dorman, M.F. & Loizou, P.C. (1998). The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels, *Ear and Hearing, 19,* 162-166.

Dorman, M.F., Loizou, P.C. & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs, *Journal of the Acoustical Society of America, 102,* 2403-2411.

Dudley, H. (1939). Remaking speech, *Journal of the Acoustical Society of America, 11,* 169-177.

French, N.R. & Steinberg, J.C. (1947). Factors governing the intelligibility of speech sounds, *Journal of the Acoustical Society of America, 19,* 90-119.

Fu, Q.J. and Shannon, R.V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing, *Journal of the Acoustical Society of America, 105,* 1889-1900.

Fu, Q.J. & Shannon, R.V. (2000). Effect of stimulation rate on phoneme recognition by nucleus-22 cochlear implant listeners, *Journal of the Acoustical Society of America, 107,* 589-597.

Greenwood, D.D. (1990). A cochlear frequency-position function for several species—29 years later, *Journal of the Acoustical Society of America, 87,* 2592-2605.

House, W.F. (1994). *Cochlear Implants: My perspective*. All Hear Incorporated: Portland. http://www.allhear.com/monographs/m-95-htm.html

Loebach, J.L. (2005). Temporal aspects of speech: The encoding of naturally produced and spectrally reduced synthetic speech by the auditory nerve. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Loizou, P.C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech, *Journal of the Acoustical Society of America, 106*, 2097-2103.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects, *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 336*, 367-373.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science, 270,* 303-304.

Van Tasell, D.J., Greenfield, D.G., Logemann, J.J. & Nelson, D.A. (1992). Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants, *Journal of the Acoustical Society of America, 92,* 1247-1257.

Van Tasell, D.J., Soli, S.D., Kirby, V.M. & Widin, G.P. (1987). Speech waveform envelope cues for consonant recognition, *Journal of the Acoustical Society of America, 82,* 1152-1161.

Zwicker, E., Flottorp, G. & Stevens, S.S. (1957). Critical band width in loudness summation, *Journal of the Acoustical Society of America, 29,* 548-557.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

A Cross-Language Familiar Talker Advantage?[1]

Susannah V. Levi,[2] Stephen J. Winters,[3] and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We would like to thank Melissa Troyer for her help with data collection.
[2] Currently at the University of Michigan.
[3] Currently at the University of Alberta.

A Cross-Language Familiar Talker Advantage?

Abstract. Previous research has shown that familiar talkers are more intelligible than unfamiliar talkers. In the current study, we tested the source of this familiar talker advantage by manipulating the type of talker information available in the signal. Two groups of listeners were trained to identify the voices of five German-English bilingual talkers; one group learned the voices from German stimuli and the other from English stimuli. After three days of training, all listeners performed a word recognition task in English. Consistent with previous findings, English-trained listeners found the speech of trained talkers to be more intelligible than untrained talkers, as measured by whole words and phonemes correct. German-trained listeners, however, showed no familiar talker advantage, suggesting that listeners must have knowledge of talker-specific, linguistically relevant information to elicit the familiar talker advantage.

Introduction

The speech waveform conveys both indexical and linguistic information. Indexical information includes details about the talker, such as gender, age, sociolinguistic background, and personal identity (Abercrombie, 1967). Linguistic information forms the content of the utterance. Although listeners can selectively attend to either one of these two dimensions, a growing body of literature shows that these two types of information interact in speech processing. Linguistic experience has been shown to affect indexical processing; listeners are better able to identify and discriminate talkers in their native language (Goggin, Thompson, Strube, & Simental, 1991; Thompson, 1987) or in a second language (Köster & Schiller, 1997; Schiller & Köster, 1996; Schlichting & Sullivan, 1997; Sullivan & Schlichting, 2000) than in an unfamiliar language. Similarly, aspects of the indexical dimension can affect linguistic processing; linguistic processing is faster and/or more accurate in single-talker conditions compared to multiple-talker conditions (e.g., Goldinger, Pisoni, & Logan, 1991; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989), in same-talker compared to different-talker conditions (Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, 1992), with acoustically similar talkers compared to acoustically different talkers (Magnuson & Nusbaum, 2007), and with familiar compared to unfamiliar talkers (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). In this study we investigated the factors that are responsible for this latter effect – the familiar talker advantage – by training listeners to learn the voices of bilingual talkers either in English or in an unknown language. This manipulation allowed us to control the type of indexical information that listeners received and test whether language-specific indexical information is necessary to elicit the familiar talker advantage.

What do listeners know about familiar talkers that they do not know about unfamiliar talkers that facilitates linguistic processing? When listening to a talker, listeners have access to both language-specific indexical information and language-independent indexical information. Language-specific indexical properties are tied to the linguistic information encoded in the speech signal, such as dialectal and idiolectal articulations of the talker. Because these indexical properties are associated with linguistic contrasts in the language, they are not available as cues to talker identity in other languages. In contrast, language-independent indexical properties are cues to talker identity that are available across different languages. These properties include size and shape of the vocal tract, gender, and age.

The existence of language-independent indexical properties has been demonstrated recently in a study using cross-language talker identification and discrimination tasks (Winters, Levi, & Pisoni, submitted). Monolingual English listeners learned to identify the voices of bilinguals speaking in either

English or German and were later tested on their ability to generalize this knowledge to the other language. Both groups of listeners were able to identify talkers above chance in the untrained language, even when it was an unknown language. In a second experiment that measured cross-language talker discrimination, untrained English listeners were asked to judge whether two words were spoken by the same or different talker in both matched language (both English or both German) and mismatched (one stimulus in English, one in German) conditions. Results from this experiment revealed that language-independent indexical properties exist and are sufficient to support accurate talker discrimination across different languages. From these results Winters et al. concluded that some aspects of a talker's identity must be retained when speaking different languages (i.e. language-independent indexical properties) and that listeners are able to reliably use those acoustic attributes of speech to perform voice identification and discrimination tasks.

When listeners know both types of indexical information about a set of talkers, they show facilitation in linguistic processing tasks. Nygaard, Sommers, and Pisoni (1994) trained native English listeners to learn the voices of ten unfamiliar talkers speaking in English and then had them complete a speech intelligibility task with words from both familiar (i.e., trained) and novel talkers mixed with noise. Results revealed a familiar talker advantage with greater word recognition accuracy for familiar talkers compared to unfamiliar talkers. Because listeners learned the novel voices using English words, they were able to learn both language-specific and language-independent indexical information about the talkers and could use this knowledge to facilitate linguistic perception. Other research confirms that listeners learn and store subphonemic, talker-specific, linguistically relevant articulations and use this knowledge in a talker-contingent manner when performing phoneme categorization tasks (Allen & Miller, 2004; Eisner & McQueen, 2005; Kraljic & Samuel, 2005). In these studies, listeners learned to associate a potentially acoustically ambiguous segment with a particular phoneme (t/d, f/s, or s/ʃ) and then generalized this knowledge about the category boundary to new stimuli spoken by the same talker but not by different talkers.

While knowledge of both language-specific and language-independent indexical properties can elicit the familiar talker advantage, it remains unclear whether knowledge of language-specific indexical properties is necessary for talker-contingent effects to be observed. To examine this issue, we controlled the type of indexical information available in the signal by familiarizing listeners with talkers in different languages. Two groups of monolingual English listeners were trained on the voices of L1 German/L2 English bilingual talkers. One group learned the voices from German stimuli, while the other group learned the voices from English stimuli produced by the same set of talkers. After training, listeners performed a word recognition task in English with both familiar talkers and unfamiliar talkers. With this manipulation, we were able to isolate the type of indexical information that listeners were able to learn from the talkers. Listeners trained with German stimuli could only learn talker-general, language-independent characteristics (and possibly some German-specific characteristics), whereas listeners trained with English stimuli not only learned language-independent indexical properties of the talker's voice but also acquired detailed, English-specific properties. If listeners require knowledge of language-specific indexical properties to exhibit a familiar talker advantage, then listeners in the German training condition should not show a facilitation during English word recognition for familiar talkers.

Experiment

Methods

Stimulus Materials. Twelve female German L1/English L2 speakers living in Bloomington, IN, were recorded in a sound-attenuated IAC booth at the Speech Research Laboratory at Indiana University.

Speech samples were recorded using a SHURE SM98 head-mounted unidirectional (cardioid) condenser microphone with a flat frequency response from 40 to 20,000 Hz. Utterances were digitized into 16-bit stereo recordings via Tucker-Davis Technologies System II hardware at 22,050 Hz and saved directly to an IBM-PC. A single repetition of 360 English and 360 German words was produced by each speaker. Each word was of the form consonant-vowel-consonant (CVC) and was selected from the CELEX English and German databases (Baayen, Piepenbrock, & Gulikers, 1995). Stimulus materials were presented visually to speakers in random order and blocked by language. (See Levi, Winters, & Pisoni, 2007 for additional details about the recording methods.) The recording session lasted approximately one hour per language. The silent portions before and after each stimulus were removed by hand using Praat sound editing software, and the resulting tokens were normalized to a uniform RMS amplitude of 66.5 dB. German was selected as the second language in the experiment because it has a sufficient number of CVC words with the same syllabic structure as the English words and because uniformly calculated frequency counts for both the English and German words were available in the CELEX database.

The bilingual speakers were given the option of recording the materials in two sessions, but all speakers elected to record all stimuli in a single recording session. Bilingual speakers were paid $10 per hour for their time. Two speakers were eliminated (speech disorder, N=1; greater age difference: N=1), yielding 10 bilingual speakers. Based on data collected in a pilot word-recognition study, talkers were divided into two groups ("Group 1 talkers," "Group 2 talkers") of roughly equal intelligibility. Average intelligibility scores, as well as other demographic data, are provided in Table 1.

Talker Group	Speaker	Age of Acquisition	Years of English	Length of Residence	Fluency	Intelligibility
1	**F3**	10	14	1	5	49.0
	F4	13	13	3	4.5	43.8
	F7	9	12	1	5	33.5
	F9	9	16	2	4	48.4
	F10	13	11	5	5	38.5
	Mean (SD)	10.8 (2.1)	13.2 (1.9)	2.4 (1.7)	4.7 (.4)	42.7 (6.6)
2	**F2**	12	9	1	4	37.1
	F5	10	14	5	3	41.1
	F8	13	16	4	5	54.8
	F11	--	--	2	4.5	41.3
	F12	7	26	5	5	54.8
	Mean (SD)	10.5 (2.6)	16.2 (7.1)	3.4 (1.8)	4.3 (.8)	45.8 (8.3)

Table 1. Demographic variables for the bilingual speakers. "Years of English" refers to the number of years speakers have been learning/using English (current age – age of acquisition). "Fluency" is a self-reported measure of English proficiency (1=poor, 5=fluent). The final column provides a measure of each speaker's intelligibility as measured by average number of words correctly perceived under four signal-to-noise ratios by a set of untrained listeners.

Seven female native speakers of American English were also recorded producing only the list of English words under the same conditions as the bilingual speakers. Productions from two of the female speakers were not included in the study due to problems these speakers had with completing the task accurately. The remaining five speakers were between the ages of 18-25 and reported no history of a speech or hearing disorder. These speakers received partial course credit for their participation.

Participants. Forty-two listeners participated in the German-training condition (21 in each training group) and 41 in the English-training condition (19 trained on group 1 talkers, 22 on group 2 talkers). All listeners were native speakers of American English attending Indiana University. In the German-training condition, 10 listeners were eliminated (did not reach criterion, N=5; did not complete the experiment, N=3; nonnative speaker of American English, N=1; lived in Germany, N=1), resulting in 32 usable listeners. Nine listeners were eliminated in the English-training condition (did not complete the experiment, N=4; nonnative speaker of American English, N=2; German-speaking parent, N=1; last participants to complete the experiment, N=2) yielding 32 usable listeners. None of the remaining 64 listeners reported any knowledge of German, had ever lived in Germany, or had any German-speaking friends or family members. All were between the ages of 18-25 and reported no history of speech or hearing impairments. Listeners were paid $10/hour for their participation. In each training condition, half of the listeners were trained on Group 1 Talkers ("Group 1 Listeners") and half on Group 2 Talkers ("Group 2 Listeners").

The data from listeners who did not correctly identify at least 40% of the talkers in 3 (half) or more testing phases during training were excluded from analysis. This level of performance was selected to mirror the criterion used in Winters, Levi, and Pisoni (submitted). In addition, listeners were divided into "good learners" and "poor learners" following the criterion used in Nygaard and Pisoni (1998) who found that listeners who did not reach 70% accuracy in voice identification did not show the familiar talker advantage. In the German-training condition, 9/16 Group 1 Listeners and 7/16 Group 2 Listeners were classified as good learners. In the English-training condition, 8/16 Group 1 Listeners and 12/16 Group 2 Listeners were good learners.

Procedure. During the four days of the study, participants were seated in a quiet room at individual testing stations. All stimuli were presented to participants over Beyer Dynamic DT-100 headphones on PowerMac G4 computers running a customized SuperCard (version 4.1.1) stack. Participants were trained to identify one of two sets of five different bilingual voices by name in six training sessions spanning three days. Each training session consisted of seven distinct phases, summarized in Table 2. Each talker was associated with a common female name in both English and German and each name was presented in a different color in a unique position on the screen.

During each training session, listeners completed two training blocks followed by a testing block. Each training block began with two familiarization phases where listeners heard the same words from each of the five talkers. After familiarization, listeners completed a recognition task in which they heard five different tokens from each of the five talkers presented twice in random order. During recognition, listeners selected a talker by clicking an on-screen button next to the appropriate talker's name and then received feedback by seeing the correct talker's name and hearing the same stimulus token repeated again. After two training blocks, listeners completed a testing phase with no feedback. The testing phase consisted of 10 words produced once by each of the five speakers in random order. The only difference between the English and German training sessions was that the word used during familiarization B was the same as the last word used during familiarization A for the English trained listeners, but was a novel word for the German trained listeners. Each training session (consisting of two training blocks plus the test phase) lasted approximately 20 minutes. Participants completed two training sessions per day for three days. Participants were required to take a short (approximately five minute) break between consecutive sessions on each day of training.

Training Session

	Phase	Stimuli	Task
Training Block I	Familiarization A	Same 5 words produced by each talker (500 ms ISI)	Listen and attend to talker-name pair
	Familiarization B	Same 1 word produced by each talker	Listen and attend to talker-name pair
	Recognition	5 different words produced by each talker, presented twice in random order	Identify speaker (feedback)
Training Block II	Familiarization A	same procedure as above	same procedure as above
	Familiarization B	same procedure as above	same procedure as above
	Recognition	same procedure as above	same procedure as above
	Test	10 different words produced by each talker, presented once in random order	Identify speaker (no feedback)

Table 2. Training procedure used for each session. Two training sessions were completed each day.

On the fourth day of the experiment, listeners completed a word recognition task in which they heard monosyllabic CVC English words and were asked to type what they heard. Stimuli in the word recognition test were presented to listeners at four different signal-to-noise ratios (SNR): Clear (no noise added), +10, +5, and 0 dB SNR. Each stimulus was mixed with white noise which included a 200 ms linearly increasing ramp from 0 dB to the appropriate noise level at the beginning of the stimulus and a similar 200 ms decreasing ramp of the noise at the end. One quarter of the stimuli were presented at each SNR. No more than two days intervened between any of the four days of testing.

German-trained listeners heard all 360 English words during the word recognition task. One third of the stimuli were spoken by Group 1 talkers, one third by Group 2 talkers, and one third by the five native speakers of English. The English-trained listeners heard only 180 words during the word recognition task. These 180 words were randomly selected for each listener and they did not occur during any training sessions to avoid any lexical priming between the training and testing phases. For each listener, one third of these words were spoken by Group 1 Talkers, one third by Group 2 Talkers, and one third by the native English speakers.

Results

Training. An Analysis of Variance (ANOVA) was conducted on the response data from the test phases of the six training sessions. This ANOVA assessed the effects that Training Session (1, 2, 3, 4, 5, 6) and Training Language (English, German) had on the percentage of talkers correctly identified in each testing phase. The ANOVA revealed a significant main effect of training session ($F(5,62) = 84.34$; $p < .001$), but no effect of training language, nor an interaction between training session and training language. The main effect of training session indicated that talker identification accuracy improved across the six training sessions. In other words, listeners were able to learn the voices of the bilingual talkers across the three days of training. The lack of a main effect of training language or an interaction between training language and training session suggests that listeners learned the talkers to the same degree and at the same rate regardless of the training language. These results are illustrated in Figure 1.

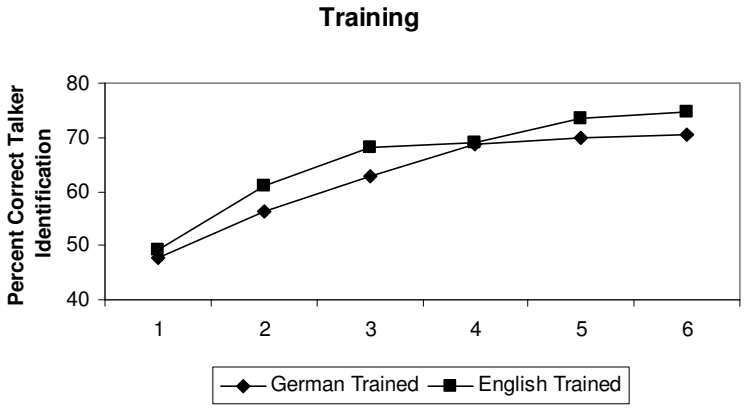


Figure 1. Talker identification accuracy during the six training sessions for both German-trained and English-trained listeners. Two training sessions were completed on each day of training.

Word Recognition. As previously mentioned, Nygaard and Pisoni (1998) found that "poor" learners did not exhibit a familiar talker advantage when performing a word recognition task similar to the one used in the current study. Using their criterion of 70% correct talker identification accuracy, listeners from both language training groups (English, German) and both talker training groups (trained on Group 1 talkers, trained on Group 2 talkers) were divided into "good learners" (those listeners who achieved 70% or greater on the last day of training) and "poor learners" (those listeners who did not reach 70% accuracy on the last day of training). Typed responses to the word recognition test were coded for whole word accuracy and for the number of correct phonemes per response (0-3). We first report the results for the German-trained listeners, followed by the English-trained listeners.

German-trained Listeners. Separate ANOVAs for good and poor learners were run on the whole word correct data with Talker Group (Group 1 talkers, Group 2 talkers) and SNR (clear, +10, +5, 0 dB SNR) as within-subjects factors and Listener Group (trained on Group 1 talkers, trained on Group 2 talkers) as a between-subjects factor. Figure 2 presents the results for whole words correct. For the good German-trained learners, only the main effect of SNR reached significance ($F(3,42) = 263.6$, $p < .001$), indicating that listeners performed better under more favorable SNRs. No other main effects or interactions reached significance. For the poor learners, main effects of SNR ($F(3,42) = 299.3$, $p < .001$) and talker group ($F(1,14) = 5.932$, $p = 0.029$) were also found. The main effect of SNR again shows the benefit of increased SNR. The main effect of talker group indicates that the poor learners found Group 2 talkers more intelligible than Group 1 talkers (45.1% vs. 42.7% words correct). This difference in average intelligibility for the poor learners likely reflects the inherent intelligibility differences in the two groups of talkers (see Table 1).

Similar results were obtained for the number of phonemes correctly identified during word recognition (Figure 3). For the good German-trained learners, only the main effect of SNR reached significance ($F(3,42) = 363.9$, $p < .001$). For the poor learners, main effects for SNR ($F(3,42) = 332.8$, $p < .001$) and talker group were found ($F(1,14) = 6.104$, $p = 0.027$). As with the whole word correct data, poor German-trained learners perceived more phonemes correctly for Group 2 talkers than Group 1 talkers (70.4% vs. 68.8% phonemes correct).

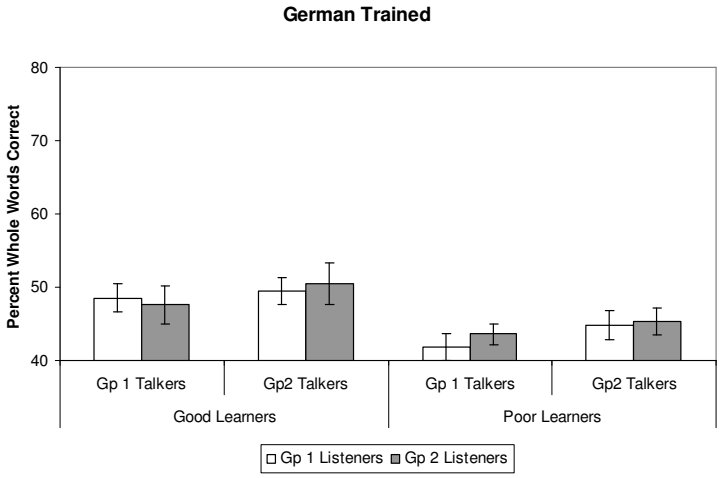


Figure 2. Percent whole words correct for German-trained listeners.

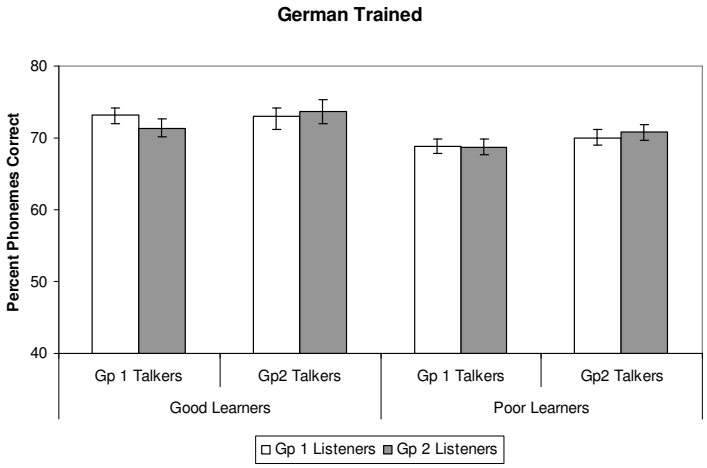


Figure 3. Percent phonemes correct for German-trained listeners.

English-trained Listeners. The pattern of results for the English-trained listeners differs from of the results obtained for the German-trained listeners. As with the German-trained listeners, separate ANOVAs for good and poor learners were conducted on the whole word correct data with Talker Group and SNR as within-subjects factors and Listener Group as a between-subjects factor. For the good English-trained learners, a main effect of SNR was found ($F(3,54) = 2.14.8$, $p < .001$). In addition to this main effect, the talker group by listener group by SNR interaction also reached significance ($F(3,54) = 2.918$, $p = .041$) and the talker group by listener group interaction approached significance ($F(1,18) = 3.632$, $p = .071$). This latter crossover interaction indicates that good English-trained learners perceived more whole words correct for trained talkers than for untrained talkers. This result is displayed in Figure 4 where the outer bars for the good learners (Group 1 talkers matched with Group 1 listeners and Group 2 talkers matched with Group 2 listeners) are higher than the inner bars. The significant three-way

interaction results from different patterns of results at each SNR, driven mostly by a large benefit of familiarity at the +5 dB SNR, and less benefit at the other SNRs. For the poor English-trained learners, only a main effect of SNR was found ($F(3,30) = 339.7$, $p < .001$).

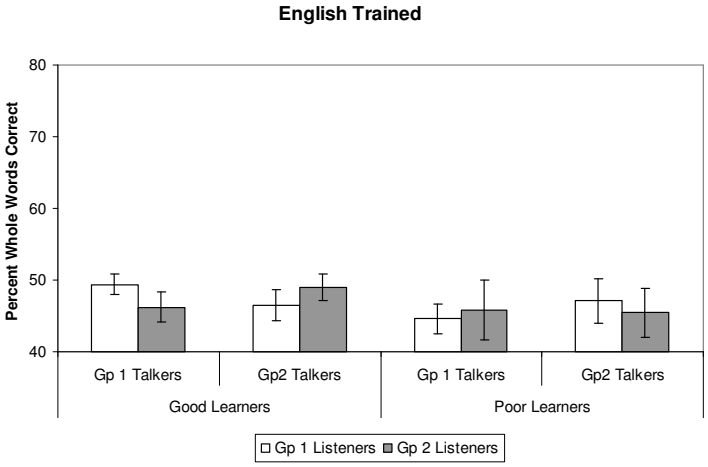


Figure 4. Percent whole words correct for English-trained listeners.

Similar results were found for the number of phonemes correctly identified during word recognition (Figure 5). For the good English-trained learners, a main effect of SNR was found ($F(3,54) = 8.654$, $p < .001$). In addition, the talker group by SNR interaction ($F(3,54) = 3.121$, $p = .032$) and the talker group by listener group interaction ($F(1,18) = 8.674$, $p = 0.008$) were significant. Paired-samples t-tests of the talker group by SNR interaction revealed that in the clear condition, listeners perceived more phonemes correct for the Group 2 talkers than for the Group 1 talkers ($p = .036$), likely reflecting the inherent differences between the two talker groups; no differences in talker intelligibility were found for the other three SNRs. As with the whole word correct data, the talker group by listener group interaction indicated that good learners perceived more phonemes correct when listening to familiar talkers than to unfamiliar talkers. For the poor learners, the main effect of SNR reached significance ($F(3,30) = 5.321$, $p < .001$), as did the talker group by listener group by SNR interaction ($F(3,30) = 3.597$, $p = .025$). Further examination of this three-way interaction revealed that in the clear listening condition, poor learners actually perceived more phonemes correct for untrained talkers than for trained talkers.

*Correlational Data***.** Converging evidence for the differences between English-trained and German-trained listeners and additional support for separating listeners into good and poor learners was obtained from correlations carried out between the degree of talker familiarity and performance on the word recognition task. Bivariate correlations were conducted between each listener's talker identification score (percent of talkers correctly identified on the last day of training) – a measure which indicates degree of familiarity with the talkers – and the observed gain in speech intelligibility, which was computed as the difference between trained talkers and untrained talkers in the word recognition task. Separate correlations were run for the whole word correct data and the phonemes correct data. For German-trained listeners, no significant correlations were found for either whole words correct or phonemes correct. In contrast, significant correlations were found for both whole words ($r = .353$, $p = .041$) and phonemes ($r = .466$, $p = .006$) correct for the English-trained listeners. The correlations found between degree of talker familiarity and intelligibility gain for the English-trained listeners indicates that listeners who were more familiar with the talkers – as measured by higher identification scores –

exhibited greater gains in speech intelligibility for familiar talkers. Thus, the better listeners are at learning a talker's voice, the better they are at recognizing spoken words from that talker. This positive correlation was only found with listeners who were trained with English stimuli. German-trained listeners did not exhibit this correlation, corroborating data in the previous sections which showed no familiar talker advantage for these listeners. Taken together, all of these results indicate that the familiar talker advantage is only found for English-trained listeners and that they show a relationship between degree of talker-familiarity and intelligibility gain.

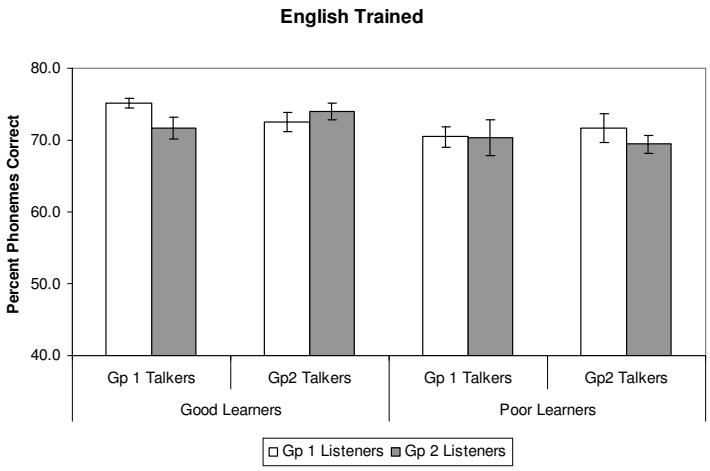


Figure 5. Percent phonemes correct for English-trained listeners.

General Discussion

The present study demonstrated that the familiar talker advantage only occurs if the same language is used during perceptual learning and word recognition, in particular where the talker learning task incorporates a linguistic component; English-speaking listeners learning voices in English do not avoid or suppress word recognition during the talker learning phase and therefore also perceive the linguistic content of the speech. These listeners showed the expected familiar talker advantage by correctly perceiving more words and more phonemes for trained talkers than for untrained talkers. In contrast, listeners who were trained on the same bilingual talkers, but using German stimuli, did not show any benefit of talker familiarity during the linguistic task. Correlational data corroborated these findings, showing that greater familiarity with a set of talkers is associated with a larger familiar talker advantage for English-trained listeners but not for German-trained listeners. Thus, it appears that increased performance in a linguistic task is contingent upon being familiar with a talker's linguistically-relevant productions in that language and upon learning the link between talkers and their linguistic characteristics.

We attribute these findings to the overlap and type of indexical information that are available in these two tasks. Listeners trained on English stimuli acquired both language-independent and English-specific indexical properties about the individual talkers, both of which are also present in the English stimuli used during word recognition. In contrast, listeners trained on German stimuli learned only language-independent indexical properties – and some German-specific indexical properties – but crucially not English-specific indexical properties. When these listeners performed a word recognition task with English stimuli, the only talker information that they had stored was language-independent

information, which contains no linguistic information that could facilitate English word recognition. We conclude that knowledge of language-specific indexical properties is necessary to generate a familiar talker advantage because listeners must know linguistically relevant talker information (i.e., language-specific indexical information) to display gains in linguistic processing.

The absence of a familiar talker advantage for the German-trained listeners cannot be explained by arguing that the talkers are perceived as unfamiliar in English. The existence of language-independent indexical properties which make talkers identifiable across languages has been established in work on cross-language talker identification and discrimination (Winters et al., submitted). Winters et al. found that listeners trained to identify talkers in one language were able to identify them in another language at levels well above chance. Furthermore, Winters et al. showed that untrained listeners can reliably discriminate talkers when speaking different languages. Thus, the lack of a familiar talker advantage by German-trained listeners is not due to the talkers sounding unfamiliar in English, because talkers can be identified from English stimuli by German-trained listeners with no decrease in performance. Rather, the absence of a familiar talker advantage must be attributed to the lack of learned English-specific indexical properties.

In the remainder of this section we explore why knowledge of language-specific indexical properties is necessary to generate the familiar talker advantage by examining data from perceptual learning studies, bilingual speech production, and cross-modal talker familiarity. We then briefly introduce two theories that have been used to account for the effects of talker variability on linguistic processing and discuss how these same perceptual mechanisms also explain the benefit of familiar voices on linguistic processing by English-trained listeners but not German-trained listeners.

Source of the Familiar Talker Advantage

When listeners learn a talker's voice from English stimuli, they also perceive the linguistic content of the utterance and thus acquire valuable information about how a talker articulates specific linguistic contrasts. Data from several perceptual learning studies show that listeners encode and retain talker-specific information about speech production and that this knowledge modifies how listeners perceive linguistic contrasts (Allen & Miller, 2004, Eisner & McQueen, 2005; Kraljic & Samuel, 2005). In one study, Allen and Miller (2004) trained listeners on the voices of two talkers, one with long voice onset times (VOTs) and one with short VOTs. During the test phase, listeners generalized talker-specific VOT differences to novel words, indicating that listeners' sensitivity to subphonemic, acoustic-phonetic differences was retained in memory and used in language processing tasks in a talker-contingent manner. Similarly, Eisner and McQueen (2005) and Kraljic and Samuel (2005) reported talker-specific subphonemic attunement in a fricative perception task. Eisner and McQueen trained listeners with an ambiguous fricative in either an [f]- or [s]-biasing lexical context and then asked them to categorize stimuli along an f/s continuum. Listeners' category boundaries were only shifted for stimuli produced in the same voice as the training stimuli. In another study using ambiguous [s] and [ʃ] stimuli, Kraljic and Samuel showed that perceptual learning of talker-specific characteristics is retained up to at least 25 minutes. All of these earlier studies provide clear and consistent evidence that listeners encode and retain talker-specific, linguistically-relevant production information for different talkers. Furthermore, this talker-contingent knowledge alters how listeners perceive the speech of familiar talkers, showing that listeners' category boundaries are talker-dependent.

Further evidence for why language-specific indexical information is necessary to elicit the familiar talker advantage comes from production studies of bilingual speakers. Languages which contain the "same" phonological contrast do not necessarily use the same cues or the same category boundary to

differentiate segments. For example, VOT values for stop consonants in Canadian French are shorter than for Canadian English (voiceless: 37 ms vs. 88 ms; voiced: -99 ms vs. 20 ms) (MacLeod & Stoel-Gammon, 2005). This difference in monolingual production is largely maintained in the speech of bilinguals, who use language-appropriate VOTs when speaking the different languages (Caramazza, Yeni-Komshian, Zurif, & Carbone, 1973; MacLeod & Stoel-Gammon, 2005). From these findings it is clear that knowledge of how a talker articulates a linguistic contrast in one language will not necessarily help to perceive a linguistic contrast in another language because the location of a talker's category boundary, as well as the types of cues used to distinguish different segments, are language-dependent. Thus, for the German-trained listeners in our study, learned German-specific indexical properties will not help them perceive a familiar talker's speech in English.

Finally, evidence that listeners need exposure to language-specific (i.e., English-specific) indexical properties to exhibit a familiar talker advantage comes a recent study of cross-modal talker facilitation. Rosenblum, Miller, and Sanchez (2007) had participants first transcribe sentences from visual-only stimuli and then transcribe novel sentences from auditory-only stimuli produced by either the same or different talker. Although participants were not explicitly directed to attend to talker characteristics during the initial visual-only transcription task, participants nonetheless perceived more words correctly when the same talker was used in both the visual-only and auditory-only tasks. Crucially, familiarization and testing were conducted in the same language, thus providing participants with language-specific indexical information. Although some learned indexical information in Rosenblum et al.'s study was non-acoustic, the gestural articulations of contrasts were specific to English. This knowledge of English-specific articulations facilitated speech perception of a familiar talker across modalities.

Taken together, these recent studies present converging evidence that listeners must have prior knowledge of language-specific indexical information for the familiar talker advantage to be observed. In addition, the perceptual learning studies and bilingual production studies suggest that this knowledge is necessary to enhance linguistic performance when listeners are asked to recognize words mixed in noise. When listeners are familiar with how a talker produces linguistic contrasts, they make fewer errors in (linguistic) perception. Furthermore, when listeners are only familiar with a talkers production in a different language (e.g., German), there is no facilitation of linguistic processing, because listeners lack the relevant language-specific knowledge about the talker. We now consider two accounts for why familiar voices are processed more quickly and accurately than unfamiliar talkers.

Relationship between Indexical and Linguistic Processing

Two accounts have been proposed to explain the link between indexical and linguistic processing and in particular to explain the adverse effects of talker variability on linguistic processing. Exemplar models (e.g., Goldinger, 1998, Johnson, 1997) assume that the processing cost associated with talker variability exists because both linguistic and indexical information are transmitted through the same stream of information (i.e., the acoustic waveform) and because both of these types of information are simultaneously retained in an exemplar stored in memory. In contrast, normalization theories (e.g., Nusbaum & Magnuson, 1997, Magnuson & Nusbaum, 2007) assume that each change in the talker dimension requires listeners to continually adjust and readjust their perceptual system to map a different talker's utterances onto the correct linguistic target, thus slowing perception and increasing the likelihood of errors.

Although these theories have been primarily used to explain the effects of talker variability on linguistic processing, they can also readily account for the facilitation due to talker familiarity and why

this facilitation is only observed for the English-trained listeners in our study. In exemplar theories, familiar talkers are represented by more exemplars stored in memory. The more exemplars that exist for a particular talker, the more likely it is that an incoming stimulus will match these stored exemplars along various linguistic dimensions facilitating linguistic processing. Thus, listeners trained on English stimuli exhibit fewer errors when listening to familiar talkers. In contrast, listeners trained on German stimuli only store German exemplars and examples of German linguistic contrasts. For these listeners, an incoming English stimulus from a familiar talker is not inherently more similar to existing English exemplars than a stimulus from an unfamiliar talker because no English exemplars exist for either talker. Therefore, German-trained listeners should not exhibit a familiar talker advantage.

In theories of talker normalization, listeners actively adjust their perceptual system to talker differences and learn to process the linguistic content of an utterance in a talker-contingent manner. For familiar talkers in English, the path between a speech stimulus and the linguistic abstractions is well-paved because listeners have abundant experience interpreting a familiar talker's speech from previous experience. If a listener is familiarized with talker in a different language, then the process of recognizing the talker and mapping his/her utterance onto a linguistic representation is never completed. Because German-trained listeners cannot create the mapping between talker and linguistic content, linguistic processing of "familiar" talkers is no different from unfamiliar talkers; both require a new processing strategy. Whichever theory is ultimately shown to best explain the interaction between indexical and linguistic processing, the important point here is that the same mechanisms that account for talker variability effects can also be used to account for the familiar talker advantage observed with English-trained listeners and the lack of this advantage for German-trained listeners.

Conclusions

Using a cross-language voice learning paradigm, we sought to explore the underlying causes for the familiar talker advantage. To this end, we manipulated the type of information available to listeners by familiarizing one group of listeners with all potentially relevant talker characteristics (language-independent and English-specific indexical properties) and a second group with a limited amount of talker characteristics (language-independent indexical properties). The group of listeners trained on English stimuli showed the expected familiar talker advantage during word recognition, whereas the group trained with German stimuli did not. The results of this study provide additional evidence that linguistic processing is performed in a "talker-contingent" manner and that listeners must have knowledge of talker-specific linguistic information to facilitate linguistic processing in a word recognition task. The absence of a familiar talker advantage for the German-trained listeners demonstrates that the familiar talker advantage is not due to knowing a voice *per se* or to being able to identify or discriminate different talkers, but rather to knowing how a voice (talker) produces linguistically significant contrasts in the language. Thus, to show an advantage in linguistic processing, a listener must acquire linguistic knowledge about the talker's speech.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine Publishing Company.

Allen, J.S. & Miller, J.L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 115, 3171-3183.

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Caramazza, A., Yeni-Komshian, G.H., Zurif, E.B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America*, 54, 421-428.

Eisner, F. & McQueen, J.M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67, 224-238.

Goggin, J.P., Thompson, C. P., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory & Cognition,* 19*,* 448-458.

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279.

Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 152-162.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J.W. Mullennix (Eds.) *Talker variability in speech processing* (pp. 145-164). San Diego: Academic Press.

Köster, O., & Schiller, N.O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics,* 4*,* 18-28.

Kraljic, T. & Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51, 141-178.

Levi, S.V., Winters, S.J., & Pisoni, D.B. (2007). Speaker-independent factors affecting the perception of foreign accent in a second language. *Journal of the Acoustical Society of America,* 121, 2327-2338.

MacLeod, A.A.N. & Stoel-Gammon, C. (2005). Are bilinguals different? What VOT tells us about simultaneous bilinguals. *Journal of Multilingual Communication Disorders*, 3, 118-127.

Magnuson, J.S. & Nusbaum, H.C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 391-409.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics,* 47*,* 379-390.

Mullennix, J.W., Pisoni, D.B, & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, 85, 365-378.

Nusbaum, H. & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J.W. Mullennix (Eds.) *Talker variability in speech processing* (pp. 109-132). San Diego: Academic Press.

Nygaard, L.C., & Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60, 355-376.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42-46.

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 19*,* 309-328.

Rosenblum L.D., Miller R.M., & Sanchez, K. (2007). Lip-read me now, hear me better later. *Psychological Science*, 18, 392-396.

Schacter, D.L. & Church, B.A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 915-930.

Schiller, N.O., & Köster, O. (1996) Evaluation of a foreign speaker in forensic phonetic: a report. *Forensic Linguistics,* 3*,* 176-185.

Schlichting, F. & Sullivan, K.P.H. (1997). The imitated voice – a problem for voice line ups? *Forensic Linguistics*, 4, 148-165.

Sullivan, K.P.H., & Schlichting, F. (2000). Speaker discrimination in a foreign language: first language environment, second language learners. *Forensic Linguistics,* 7, 95-111.

Thompson, C.P. (1987). A language effect in voice identification. *Applied Cognitive Psychology,* 1, 121-131.

Winters, S.J., Levi, S.V., & Pisoni, D.B. (submitted). Identification and discrimination of talkers across languages. *Journal of the Acoustical Society of America*.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Developing Coding Schemes for Assessing Errors in Open-Set Speech Recognition and Environmental Sound Identification[1]

Althea N. Bauernschmidt and Jeremy L. Loebach

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This research supported by NIH NIDCD R01 Research Grant DC00111 and NIH NIDCD T32 Training Grant DC00012 to Indiana University.

Developing Coding Schemes for Assessing Errors in Open-Set Speech Recognition and Environmental Sound Identification

Abstract. In the present study, we report on a series of coding schemes to classify errors in open-set recognition of speech and environmental sounds from previous perceptual learning experiments with cochlear implant simulations (Loebach and Pisoni, 2007; in press). Open-set responses to MRT and PB words were coded for place of articulation, manner of articulation and voicing errors in the identification of word initial and word final consonants. Open-set responses to meaningful and semantically anomalous Harvard sentences were coded for phonemic, lexical, and thematic errors in keywords. Open-set responses to environmental sounds were coded for errors in identifying the agent, action, and rhythm of the sounds. Overall, our coding scheme provided a more accurate assessment of performance producing higher percent correct recognition scores for the isolated words and environmental sounds than absolute coding schemes that simply identified entire words as correct or incorrect. Although time intensive, these coding schemes revealed perceptual elements with which subjects were having difficulty that were not apparent from the absolute coding scheme. The utility of open-set coding schemes is discussed for perceptual experiments with cochlear implant users, who often must make verbal responses to stimuli.

Introduction

According to the National Institute on Deafness and Other Communication Disorders, nearly 100,000 people worldwide have received cochlear implants as a treatment for profound hearing loss (National Institutes of Health, 2007). A cochlear implant is an auditory prosthesis that electrically stimulates the auditory nerve directly via electrodes placed in the cochlea. It thus bypasses damaged or missing hair cells that would normally stimulate the auditory nerve leading to the perception of sound. Although cochlear implant technology has been steadily improving over the past three decades, it does not and cannot restore normal hearing. The performance of the implant is influenced by several physiological and technological constraints and the resulting signal is severely spectrally degraded. However, considering the highly degraded signal that they receive, most cochlear implant users perform surprisingly well on speech recognition tasks in the quiet.

A common approach used to compensate for the degraded signal has focused on the optimization of acoustic speech information in the patterns of electric stimulation. The limited number of electrodes in the implant imposes fundamental limitations on the amount and type of information that can be transmitted. The number of electrodes that are available in the array is analogous to the number of channels of information that can be transmitted. In the normal hearing ear, there are approximately 20,000 channels transmitting detailed spectral and temporal information to the brain: in the best cochlear implant, however, there are only 24. One approach commonly used to determine how to best optimize the information in the speech signal has been to reduce the number of channels in the implant and determine the effect it has on speech perception. Generally, the more electrodes that are present in the cochlear implant, the more channels of information can be transmitted. A single channel implant only provides a crude signal that does not carry much fine spectral detail. Increasing the number of channels yields a richer signal and more spectral information. However, due to technological and physiological constraints, the number of electrodes that can be successfully implanted and utilized is limited.

It is possible to simulate the experience of hearing with a cochlear implant by using signal processing strategies similar those used in cochlear implants speech processors. This is done by band pass filtering the speech signal into the same number of channels as the number of electrodes in the array,

thereby limiting the number of channels of information that can be transmitted. The spectral detail is removed by replacing the spectral information in each band with a noise carrier, which is then modulated with the temporal envelope from the original band to simulate the temporal patterns of stimulation by the electrode array. Although there is no way to directly assess whether the simulations reproduce the actual experiences of a CI user, signal processing techniques based on CI speech processors may come close.

Studies using acoustic simulations of cochlear implants have demonstrated that the minimum number of channels that elicits high levels of speech perception can be as low as 4 (Shannon, Zeng, Wygonski & Ekelid, 1995) or as high as 20 (Dorman, Loizou, Fitzke & Tu, 1998) depending on the materials and task. These conflicting findings indicate that the type of information that is needed for good speech perception – whether purely temporal cues (Shannon et al., 1995), frequency cues (Dorman, Loizou & Rainey, 1997), or different listening strategies (Munson, Donaldson, Allen, Collison & Nelson, 2003) – is not known. Moreover, the acoustic environment, whether in quiet (Shannon et al., 1995) or in noise (Dorman et al., 1998), as well as the speech perception task that is being tested (Shannon et al., 1995; Dorman et al., 1997; Friesen, Shannon & Cruz, 2005) can influence the number of channels needed for good speech perception

Although most researchers have focused on the perception of speech under these degraded conditions, there are other important aspects of the world that are experienced through hearing. Another important and expected benefit of a cochlear implant is the improved recognition of environmental sounds. Many CI users report being able to perceive different types of sounds, such as "footsteps, slamming of doors, sounds of engines, ringing of the telephone, barking of dogs, whistling of the tea kettle, rustling of leaves, the sound of a light switch being turned on and off, and so on" (Food and Drug Administration, 2004). Moreover, the awareness of environmental sounds is often cited as an expected benefit from implantation (Clark, 2003). A particularly striking example of the expectations that implantees have for recognition of environmental sounds can be seen in the documentary Sound and Fury, a 2001 Academy Award Nominee for Best Documentary Feature. While the family of Heather, a 6 year old deaf child, is interviewing people about the benefits of their child receiving a cochlear implant the ability to perceive environmental sounds is constantly cited a reason for the surgery. Heather is told that she will be able to hear the birds chirping outside as well as many other environmental sounds. It is stressed that even if she won't be able to achieve full speech perception capabilities she will be able to hear other sounds (Aronson, 2000). Despite being a major motivation for cochlear implantation there has been little research into the perception of environmental sounds by individuals with cochlear implants, or normal hearing subjects listening to acoustic simulations of cochlear implants.

Moreover, few experiments have investigated the perception of environmental sounds in unprocessed auditory stimuli and much less is known about perception under conditions of auditory degradation. For normal hearing listeners, accuracy in open-set identification of 120 unprocessed environmental sounds exceeds 74% correct (Marcell, Borella, Green, Kerr & Rogers, 2000). For CI users, only closed-set experiments have been conducted, and accuracies as high as 79% have been reported when the stimulus set is limited to 37 sounds (Reed & Delhorne, 2005). Moreover, the important information for recognition of environmental sound stimuli has been elusive. While some studies have determined that temporal cues are important (Reed & Delhorne, 2005), others have found that the reliance on spectral or temporal information is conditional based on the specific type of sound (Gygi, Kidd, & Watson, 2004). Because only closed-set experiments on the recognition of environmental sounds have been conducted, it is unknown how well CI users perform in open-set tasks. A CI user's performance on an open-set task would provide information about which cues are important for the perception of environmental sounds, not only for CI users, but also for normal-hearing listeners as well. Performance on closed-set tasks may not be an accurate reflection of how well the CI is performing in relation to the perception of environmental sounds. Results on closed-set tasks may be inflated because the subject has learned the stimulus set or is relying primarily on contextual cues. These results do not allow

generalizations to how well CI users would perform on novel environmental sounds or sounds heard out of context.

By investigating how normal hearing listeners adapt to cochlear implant simulations, we can not only learn more about speech perception under degraded conditions, but may also discover information about perceptual learning that could have implications for rehabilitation of new CI users. This was the motivation for our previous perceptual learning experiment, which examined the effect of training on the recognition of speech and environmental sounds that were processed by an 8-channel sinewave vocoder (Loebach & Pisoni, 2007; in press). Using a diverse set of stimuli we compared whether training on words, sentences (meaningful or semantically anomalous), or environmental stimuli produced different levels of generalization to different materials using a pre-/post-test design. More importantly, we used an open-set task to assess how well subjects could identify the speech as well as environmental stimuli and determine what errors subjects make in doing so. Overall, we showed that training had a significant impact on generalization, with subjects who were trained on environmental stimuli performing as well on the speech perception tasks as subjects trained on speech. In addition, subjects who were trained on words did better when generalizing to new words and subjects who were trained on sentences did better when generalizing to new sentences.

One potential problem with our analysis is that we only scored each word as correct or incorrect, and did not take into account the types of errors subjects were making. Moreover, we did not describe the types of confusions subjects reported during the identification of environmental sounds. For the present study, we developed and implemented coding systems that provided a more detailed analysis of the errors that subjects made. Absolute coding schemes that score the number of keywords correct provide little insight into the errors that subjects make or why they make them. Understanding why subjects make mistakes in these perceptual learning tasks can aid in the evaluation of the strengths and weaknesses of different training paradigms and can lead to the development of better and more effective rehabilitation strategies for new CI users. In addition, knowing why subjects are making errors may provide insight into what information is successfully transmitted via the spectrally reduced signal, and how we can better optimize the perceptual cues available to the listener.

Method

Open-set responses to words, meaningful and anomalous sentences, and environmental sounds processed with an 8-channel sinewave vocoder were obtained from 125 normal hearing subjects (Loebach & Pisoni, 2007; in press). In that study, subjects were assigned to one of five training groups, and were explicitly trained to identify one type of stimulus materials, but were tested on all materials to assess generalization.

Coding Schemes for Single Words

Two of the sets of stimuli that we used in our original study were isolated monosyllabic words, varying in frequency of occurrence in American English and difficulty. The MRT word set is a corpus of 300 words made up of fifty lists of six rhymed variations on a common syllable (House, Williams, Hecker & Kryter, 1965). In each list of six rhymed words the word initial or word final consonant is systematically altered to produce six minimal pairs (e.g., ‘bat’, ‘bad’, ‘back’, ‘bass’, ‘ban’, ‘bath’). Ninety CVC words drawn from the MRT list were used in this experiment. The PB corpus consists of words whose phonemic composition approximates the statistical occurrence in American English (Egan, 1948). The corpus contains twenty lists of fifty monosyllabic worse. Ninety unique words drawn from lists 1-3 of the PB corpus were used in this experiment.

To analyze the words, both PB and MRT, we divided the word into three parts: the word initial consonant or consonant cluster (C1), the word medial vowel (V), and the word final consonant or consonant cluster (C2). We developed a coding system to quantify the types of errors subjects made when listening to a degraded signal as well as how they adjust their perception after experience with the stimuli. Consonantal errors were classified by place of articulation (bilabial, alveolar or velar), manner of articulation (stop, fricative, or nasal), voicing (voiced or unvoiced), cluster insertion (adding a fricative or nasal to a plosive stop, or a plosive stop to a fricative or nasal), cluster deletion (deleting a fricative or a plosive stop), indeterminate (varying on more than two axes), or omissions (word was omitted completely). The response 'plow' for the target 'cloud' is an example of a word initial place error: the phoneme /k/ in 'cloud' is an unvoiced velar stop and the phoneme /p/ in 'plow' is an unvoiced bilabial stop but the two differ only in place of articulation. The response 'bass' for the target 'mass' is an example of a word initial manner error: /b/ is a voiced bilabial stop and /m/ is a voiced nasal bilabial stop and the two differ only in manner of articulation (stop versus nasal stop). The response 'need' for 'neat' is an example of a word final voicing error: /d/ and /t/ are both alveolar stops and differ only in voicing. An example of a cluster insertion error is the response 'faint' for 'fate': the word final consonant is the same in both except for the /n/ that was inserted before the /t/. An example of a cluster deletion error is the response 'lush' for 'blush': the /l/ in the word initial /bl/ sequence has been deleted, making a word initial cluster deletion error.

A response was considered indeterminate if there were more than two errors. For example, the response 'out' for the target word 'earl' was considered to be indeterminate at all places in the word, CI, V, and C2. In addition, a response could have been marked wrong as a combination of place, manner, or voicing. For example, if the word was 'bat' and the subject responded 'tat' then C1 would be wrong in both place and voicing, but not manner. Due to the response mode used in the original study (where subjects typed their responses on a computer keyboard) vowels were scored as correct or incorrect. If the word lacked a C1 or C2 consonant/consonant cluster (e.g., 'earl') the missing consonant/consonant cluster was considered correct if the subject correctly omitted it. When /r/ preceded a vowel, it was classified as part of C1 (ex: 'drop'). When /r/ followed a vowel, it was classified as part of the vowel (ex: 'earl').

Coding Schemes for Sentences

Two different types of sentences were used in our original study: meaningful Harvard sentences, and semantically anomalous Harvard sentences. The meaningful Harvard sentences were drawn from the Harvard Sentence database (Karl & Pisoni, 1994). The database consists of seventy-two lists of ten meaningful sentences (IEEE, 1969). Sentences are phonetically balanced (relative to American English) and contain five keywords within a semantically rich meaningful sentence. Stimuli for the experiment consisted of twenty-five sentences taken from lists 1-10 of the Harvard Sentence database. The Anomalous Harvard sentences are semantically anomalous sentences that preserve the canonical syntactic structure of English, but contain thematically unrelated keywords. They were derived from the Harvard Sentence by taking the keywords from the 100 sentences in lists 11-20 and replacing them with words of equivalent semantic categories from lists 21-70 (Herman & Pisoni, 2000).

To identify the types of errors that were produced, whole keyword errors were classified as phonetic (errors made in a single phoneme or phoneme cluster), lexical (confusion with a lexically related word), thematic (filling in a word that is thematically related to the preceding or proceeding target), and indeterminate or omission. An error was considered a phonetic error if it deviated from the target word by one or two features (e.g., 'bat' for 'back'), similar to the scoring of the MRT and PB words. A lexical error was the substitution of a word that is similar in meaning but not in sound (e.g., 'boards' for 'planks'). A thematic error was a keyword error that was influenced by the surrounding words, or the overall interpretation of the sentence (e.g., 'These days a chicken liver is a rare dish' for 'These days a chicken leg is a rare dish').

Coding Schemes for Environmental Sounds

Environmental stimuli were drawn from the environmental signal database of Marcell and colleagues (2000). The database consists of stimuli recorded from a wide variety of acoustic environments developed for use in neuropsychological evaluation and confrontation naming studies (Marcell et al., 2000). Ninety stimuli from this database were used for the experiment. The environmental sounds were particularly challenging to score. As stated in the introduction, few experiments have investigated the perception of environmental sounds. As we were unable to find a conventional method of classifying environmental sounds that was comparable to the feature classification of speech sounds, we tried to identify the smallest number of subcategories that would differentiate the largest number of sounds.

Each of the environmental sounds was classified according to 3 features: the source or agent of the sound (the physical producer of the sound), the action that causes the sound, and the rhythmic or pitch information that differentiates it from other sounds. For the environmental sounds in the experiment, possible agents could include one or more of the following: animal (e.g., 'wolf howl'), human (e.g., 'child coughing'), wind (e.g., 'boat horn'), glass (e.g., 'glass shattering'), metal (e.g., 'car crash'), wood (e.g., 'door knocking'), liquid (e.g., 'water boiling'), insect (e.g., 'mosquito'), motor (e.g., 'motorcycle'), plastic (e.g., 'ping pong') and string (e.g., 'banjo'). The types of actions that could have caused the sounds presented were: burst (e.g., 'harmonica'), strike (e.g. 'basketball'), slide smooth (e.g., 'sword fight'), slide rough (e.g., 'violin'), tear (e.g., 'paper tearing'), rumble (e.g., 'thunder'), bubble (e.g., 'water boiling'), blow (e.g., 'whistling'), roll (e.g., 'pinball'), crash (e.g., 'glass shattering'), and buzz (e.g., 'mosquito'). The types of rhythmic or pitch information that could distinguish the sounds presented were: pitch high (e.g., 'baby crying'), pitch low (e.g., 'boat horn'), pitch change (e.g., 'airplane'), harmonic (e.g., 'owl'), complex (e.g., 'cash register'), transient (e.g., 'rain'), periodic (e.g., 'clock ticking') and pulse (e.g., 'gun shots').

The difficulty in this method of scoring lies in scoring unique incorrect responses. For example, 'cell phone' was a common response for the sound 'jackhammer'. Though we know that the sound of a jackhammer would have the agents metal and motor, the action of burst, and a periodic rhythm, we do not know along which dimension the sound of a cell phone differs from the sound of a jackhammer. Due to the high number of occurrences it was assumed that this was not a random error and that subjects were mistaking the sound of a cell phone set to vibrate for the sound of a jackhammer. Common responses like this were also classified according to our scheme and added to a large table for cross-reference. A portion of this table is reproduced in Appendix A for reference.

Results

Words

Analysis using the new individual phoneme based coding scheme increased performance on the MRT words across all five training groups (M = 62.8% correct) as compared to the absolute whole word correct coding scheme (M = 34% correct) used previously. When examined individually, performance was equivalent for word initial (M = 58.6%) and word final (M = 60.8%) consonants ($t(222) = -1.553$, $p = 0.123$) (Figure 1). For word initial consonants C1, subjects were in error 41.4 percent of the time. Of the responses that subjects could have made, 29.9% were errors in place of articulation, 17.2% were errors in manner of articulation and .3% were errors in voicing. Of the multiple feature errors, place and manner of articulation occurred 5.7% of the time, and errors on place and voicing, and manner and voicing were rare occurring less than .01% of the time. Cluster insertion and cluster deletion errors were also rare, occurring less than .01% of the time. For word final consonants C2, subjects were in error

39.1% of the time. Of the responses that could have been made, 19.3% were errors in place of articulation, 3.2% were errors in manner of articulation, and 1.5% were errors in voicing. Of the multiple feature errors, place and voicing errors occurred 3% of the time, place and manner errors occurred 2.5% and errors on manner and voicing were rare occurring less than .01% of the time. Cluster insertion errors occurred 5% of the time, while cluster deletion errors occurred less than .01% of the time. Comparing error prevalence across word initial and word final consonants revealed that subjects made significantly more place of articulation errors on C1 and C2 ($t(180) = 9.2356$, $p < 0.001$), but all other errors did not differ by consonantal position (all $p > 0.05$). Subjects made few errors in identifying vowels, scoring 69.1% correct.

Across the five training groups, training had a significant effect on the number of correct responses on MRT word initial ($F(4, 120) = 20.74$, $p < 0.001$) and MRT word final ($F(4, 120) = 6.59$, $p < 0.001$) consonants. Subjects scored significantly better on MRT C1 when trained on MRT words ($M = 73.5\%$) than when trained on environmental sounds ($M = 56\%$), Anomalous sentences ($M = 54.7\%$), Harvard sentences ($M = 48.7\%$), and PB words ($M = 60\%$) (all $p < 0.001$). Subjects also scored significantly better on MRT C1 when trained on PB words ($M = 60\%$) than when trained on Harvard sentences ($M = 48.7\%$; $p < 0.001$). Subjects were also more accurate at identifying MRT C2 correct when trained on MRT stimuli ($M = 64.7\%$) than when trained on Harvard sentences ($M = 54.8\%$; $p < 0.001$), but equally well as when trained on PB words ($M = 65.2\%$), Anomalous sentences ($M = 60.2\%$) or Environmental sounds ($M = 59\%$) (all $p > 0.05$).

Training had a significant effect on the number of place of articulation errors on C1 ($F(4, 120) = 19.618$, $p < 0.001$) and C2 ($F(4, 120) = 2.597$, $p = 0.041$). Subjects were significantly less likely to make place errors on C1 when trained on MRT words ($M = 17.2\%$) than when trained on environmental sounds ($M = 31.7\%$), anomalous sentences ($M = 34.2\%$), Harvard sentences ($M = 38.5\%$), and PB words ($M = 28\%$) (all $p < 0.001$). Although there was a significant effect of training on C2 errors, no specific training stimuli affected the number of place errors made on word-final consonants. Training did have a significant effect on cluster insertions in word-final ($F(4, 120) = 7.292$, $p < 0.001$) but not word-initial ($F(4, 120) = 2.433$, $p = 0.051$) consonantal position. For word final consonants, training on MRT words ($M = 0\%$) leading to significantly fewer cluster insertion errors than training on environmental sounds ($M = .8\%$), Anomalous sentences ($M = 1.4\%$), Harvard sentences ($M = .9\%$), and PB words ($M = .8\%$) (all $p < 0.03$). However, given the low occurrence of these errors across training groups this effect may be more apparent than real.

Analysis using the new individual phoneme-based coding scheme increased performance on the PB words across all five training groups ($M = 69.1\%$ correct) as compared to the absolute whole word correct coding scheme ($M = 44\%$ correct) used previously. When examined individually, subjects performed significantly better on word initial ($M = 70.2\%$) than word final ($M = 66.04\%$) consonants ($t(226) = 4.667$, $p < 0.001$) (Figure 1). For word initial consonants, subjects were in error 30.9 percent of the time. Of the errors that could have been made, 9.2% were errors in place of articulation, 1.9% were errors in manner of articulation and 2.3% were voicing errors. Of the errors that could be made on two features, place and voicing errors occurred 2% of the time, place and manner errors occurred 6.5% of the time, and manner and voicing errors were rare, occurring less than .01% of the time. Cluster insertion errors were also rare, occurring .5% of the time, while cluster deletion errors were more common, occurring 4.3% of the time. For C2, subjects were in error 30% of the time. Of the errors that could have been made, 12.1% were place of articulation errors, 2.0% were manner of articulation errors, and 2.2% were voicing errors. Of errors that could have been made on multiple features, .8% were place and voicing errors, 4% were place and manner errors, and manner and voicing errors were rare, occurring less than .01% of the time. Cluster insertion errors were more common, occurring 4.9% of the time, while cluster deletion errors were comparatively rare, occurring 1.8% of the time. Unlike the MRT words, subjects made significantly more place of articulation errors on the word final consonant ($M = 12.1\%$)

than the word initial consonant (M = 9.2%) ($t(248)$ = -5.98, $p < 0.001$), but all other errors were equivalent across word initial and word final consonants (all $p > 0.05$). Subjects scored 71.1% correct on the PB vowels.

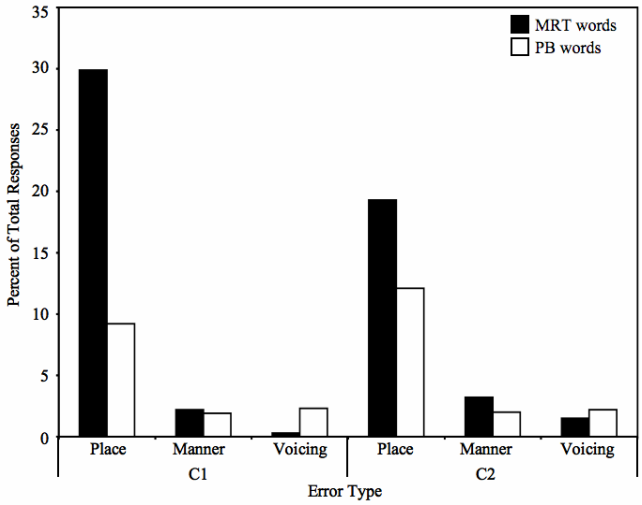


Figure 1. Comparison of errors made on MRT (black) and PB (white) word initial (C1) or word final (C2) consonants.

When comparing the performance on the PB words across the five training groups, training did not have a significant main effect on the percent correct recognition of C1 ($F(4,120\_$ = .78, p = 0.541): all training conditions produced equivalent performance on C1. While there was a significant main effect of training on number of correct C2 responses ($F(4, 120)$ = 2.451, p=.05), no specific training condition was significantly more likely to improve the subject's word final PB consonant score (all $p > 0.074$). Training had significant effect on the number of voicing errors made on C1 ($F(4, 120)$ = 15.590, $p < 0.001$) and C2 ($F(4, 120)$ = 2.805, p = 0.029). Subjects were least likely to make voicing errors on C1 when trained on PB words (M = .000) than on any other type of material: MRT (M = .3%), HS (M = .2%), AS (.3%), ENV (M = .3%). Subjects were also significantly less likely to make voicing errors on C2 when trained on PB words (M = .1%) than when trained on MRT words (M = .2%), Anomalous sentences (M = .3%) Harvard sentences (M = .3%) and Environmental sounds (M = .3%). However, given the low prevalence of errors, this effect may be more apparent than real.). Training also had a significant main effect on the number of manner and voicing errors that subjects made on both C1 ($F(4, 120)$ = 2.674, p = 0.035) and C2 ($F(4, 120)$ = 3.578, p = 0.008), but again, given the low prevalence of occurrence (less than 5% of the time) these differences may be more apparent than real.

Comparing only MRT and PB training groups, subjects performed equally well on the word initial consonants for PB (M = 70.1%) and MRT words (M = 72.5%; $t(48)$ = 1.598, p = 0.117). For the word final consonant, however, subjects performed significantly better on PB words (M = 70.0%) than MRT words (M = 64.7%; $t(48)$ = 2.163, p = 0.036). As reported above, training specificity had a greater effect for MRT consonants than PB consonants: MRT training produced significantly better performance on C1 for MRT words than training on other materials, whereas all forms of training were equally effective for C1 in PB words. Subjects also made significantly fewer place of articulation errors on C1 for PB words (M = 10.0%) than on C1 for MRT words (M = 17.1%; $t(36)$ = 4.58, $p < .001$). Additionally, subjects made significantly fewer place of articulation errors on word final consonants for PB words (M = 10.6%) than for MRT words (M = 21.6%, $t(39)$ = .8.652, $p < 0.001$). The difference in error types as well as overall percent correct recognition scores on MRT and PB words may be due to the difference in the

phonemic composition of the two types of stimuli. The variability in the types of errors made on the PB words may be because that their phonemic composition approximates the statistical occurrence of those phonemes in American English. MRT words are not phonetically balanced relative to American English, and are composed of minimal pairs. This may predispose subjects to making specific types of errors (such as errors in place of articulation) since the stimuli can only differ on one or two dimensions. In addition to having more varied error types, PB words showed less training specificity than MRT words, which could also be due to the differences in phonetic balance.

Sentences

Across all five training groups, subjects scored 70.3% correct on Harvard sentence keywords. Of the responses, 14.2% were phonetic errors, 1.6% were thematic errors, and .4% were lexical errors (Figure 2). There was a significant effect of training on the number of Harvard sentence keywords correctly identified ($F(4, 120) = 3.47$, $p = .01$): subjects' performance improved when trained on Harvard sentences ($M = 76.5\%$) than on MRT ($M = 68\%$) or PB ($M = 68\%$) words (both $p < 0.03$). Subjects trained on Anomalous sentences ($M = 69\%$) and Environmental sounds ($M = 69.8\%$), however, performed as well as subjects trained on Harvard sentences (both $p > 0.07$). Subjects made significantly less phonetic errors when trained on Harvard sentences ($M = 8\%$) than on MRT words ($M = 18\%$), Environmental sounds ($M = 16\%$), Anomalous sentences ($M = 16\%$), or PB words ($M = 14\%$) ($F(4, 120) = 13.8$, $p < .001$). Subjects also made significantly fewer thematic errors ($F(4, 120) = 5.161$, $p = .001$) when trained on Harvard sentences ($M = .8\%$) than on Environmental sounds ($M = 2\%$) Anomalous sentences ($M = 2\%$), MRT words ($M = 1.6\%$) and PB words ($M = 1.3\%$). However, given the low prevalence of thematic errors overall, this effect may be more apparent than real.

Across all five training groups, subjects scored 55% correct on Anomalous sentence keywords. Of the total Anomalous keyword responses, 29.5% were phonetic errors, .6% were thematic errors, and there were no lexical errors (Figure 2). There was a significant effect of training on number of Anomalous sentence keywords correct ($F(4, 120) = 21.05$, $p < .001$), and subjects performed better on Anomalous sentences when trained on Anomalous sentences ($M = 68\%$) than on MRT words ($M = 50\%$), Environmental sounds ($M = 50\%$), Harvard sentences ($M = 60\%$), and PB words ($M = 47\%$). Subjects made significantly fewer phonetic errors ($F(4, 120) = 17.4$, $p < .001$) when trained on Anomalous sentences ($M = 21\%$) than when trained on MRT words ($M = 33\%$), Environmental sounds ($M = 34\%$), Harvard sentences ($M = 28\%$), or PB words ($M = 29.8\%$).

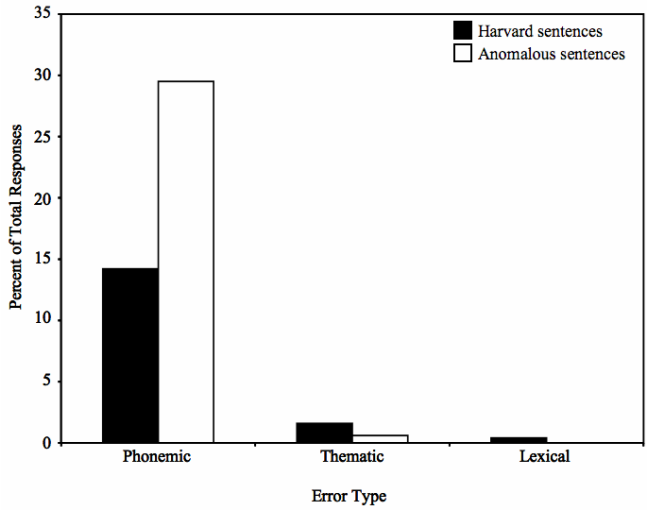


Figure 2. Comparison of errors made in Harvard (black) and Anomalous sentences (white).

Comparing the two training groups, subjects performed significantly better on Harvard sentences than Anomalous sentences ($t(42.9) = -3.36$, $p = .002$). Subjects made significantly more phonetic errors on Anomalous sentences than on Harvard sentences ($t(48) = 9.39$, $p < .001$). Subjects also made significantly more lexical errors when trained on Harvard sentences than when trained on Anomalous sentences ($t(33.6) = -2.08$, $p = .045$). There were no significant differences in thematic errors across the two training groups. Because the Anomalous sentences are derived from the Harvard sentences, any difference in performance between the two is presumably due to sentence context. It is therefore interesting that the only error type not to be significantly affected by training is thematic errors. If subjects were filling in misheard items with thematically related items, it would suggest that they are using a contextually driven strategy for keyword identification. Since the thematic errors were rare overall, and did not differ between the two training groups, it would suggest that subjects were primarily using an acoustic phonetic perceptual strategy for identification. Moreover, no differences in performance were observed between the new coding scheme and the previous coding scheme (Loebach & Pisoni, 2007; in press) due to both coded for whole keywords only. The new coding scheme, however, did provide an additional level upon which to assess subject performance and determine the cognitive strategies that subjects employ when listening to these sentences.

Preliminary Results for Environmental Sounds

Error coding for the environmental sound recognition tasks is ongoing due to the complexity of the coding scheme and the large number of novel responses that subjects report. Coding is complete for one of the five training groups (MRT training group), and these preliminary data are presented below.

Using the new coding scheme, a significant increase in performance was observed for environmental sounds, increasing from 37.6% correct for the absolute coding scheme to 49.6% correct for the 3-tiered coding scheme. For the Agent valence, subjects were correct 46.9% of the time. Of the errors that subjects made, 17.3% were determinate (i.e., could be classified under the new coding scheme) and 38.8% were indeterminable, for a total error rate of 53.13%. Table 1 displays an error matrix of possible responses (columns) to target agents (rows). Cell values represent the total percentage of responses made. Values along the diagonal (grey shaded cells) indicate the percentage of incorrect responses that were

	Animal	Human	Insect	Liquid	Wind	Glass	Metal	Wood	Paper	Rubber	String	Motor	Unk.
Animal	**.09**	.01	.03	0	.02	0	***.13***	0	0	0	0	.01	.43
Human	***.12***	**0**	0	.02	.01	0	.09	.01	0	0	.01	0	.37
Insect	.04	0	***.19***	.02	0	0	.02	0	0	0	0	0	.21
Liquid	0	0	0	**.01**	.01	0	.04	0	0	0	0	0	.27
Wind	.01	.01	0	.01	**0**	0	***.10***	.01	0	0	.01	.07	.40
Glass	0	0	0	0	0	**0**	***.44***	0	0	0	0	0	.56
Metal	.01	0	0	0	.01	0	**.01**	0	0	0	0	0	.30
Wood	0	0	0	.01	.01	.01	0	**.02**	0	0	0	0	.30
Paper	.06	0	0	.06	0	0	***.24***	0	**0**	0	0	0	.59
Rubber	0	0	0	.07	.03	0	***.17***	0	0	**0**	0	.03	.63
String	0	0	0	0	.09	0	***.17***	0	0	0	**.02**	0	.66
Motor	0	0	0	0	.02	0	.03	0	0	0	0	***.10***	.17

Table 1. Error matrix displaying the frequency of responses for each agent. Target agents appear in column 1, and response agents appear in row 1.

within the same agent class as the target (e.g., responding "duck" to a stimulus of a "cow"). Row sums indicate the total percentage of incorrect responses, and can be subtracted from one to obtain the percent

correct recognition scores for each target agent. Of the possible agents that subjects described in their responses, Metal was the most prevalent agent error, with subjects indicating metallic agents 53% of the time. The next most prevalent error was Animal, with subjects indicating animal agents 12% of the time.

For the Action valence, subjects correctly identified the appropriate action 52.1% of the time. Of the errors that subjects made, 8.4% were determinate (i.e., could be classified under the new coding scheme) and 39.5% were indeterminable (for a total error rate of 47.9%). Of the possible actions that subjects described in their responses (Table 2), Strike was the most prevalent action error, with subjects indicating striking actions 52% of the time. The next most prevalent error was Rumble, with subjects indicating rumbling actions 18% of the time. A comparison of the frequency of incorrect actions being selected revealed that there was a significant main effect of action ($F(10,709) = 14.299$, $p < 0.001$). Post hoc Bonferroni tests revealed that this effect was driven entirely by the Strike action (with striking actions being described significantly more often than any other action; all $p < 0.001$). No other action was selected significantly more often than any other.

	Blow	Bubble	Burst	Buzz	Crash	Roll	Rumble	Slide R	Slide S	Strike	Tear	Unk.
Blow	**0**	0	.02	0	.01	0	.02	0	0	***.08***	0	.47
Bubble	0	**0**	***.08***	0	0	0	.04	0	0	0	0	.58
Burst	.01	0	**0**	0	.02	0	0	0	0	***.11***	0	.51
Buzz	0	0	0	**0**	0	0	0	0	0	0	0	.30
Crash	0	0	0	0	**0**	0	0	0	0	.02	0	.23
Roll	0	0	0	0	0	**0**	0	0	0	***.08***	0	.83
Rumble	.01	0	0	0	.01	0	**0**	.02	0	.01	0	.27
Slide R	0	0	0	0	0	0	***.07***	**0**	.02	0	0	.34
Slide S	.02	0	0	0	.01	0	0	0	**0**	.04	0	.41
Strike	0	0	0	0	0	0	0	0	0	**.01**	0	.26
Tear	0	0	0	0	0	0	0	0	0	***.06***	**0**	.82

Table 2. Error matrix displaying the frequency of responses for each action. Target actions appear in column 1, and response actions appear in row 1.

For the Rhythm valence, subjects correctly identified the appropriate rhythm 49.9% of the time. 6.7% of errors were determinate (i.e., could be classified under the new coding scheme) and 43.5% were indeterminable (for a total error rate of 50.1%). Of the possible rhythms that subjects described in their responses (Table 3), Pitch High or Low was the most prevalent rhythm error, with subjects indicating high or low pitches actions 42% of the time. The next most prevalent error was Periodic, with subjects indicating periodic rhythms 25% of the time. A comparison of the frequency of incorrect rhythms being selected revealed that there was a significant main effect of rhythm ($F(6,413) = 2.94$, $p = 0.008$). Post hoc Bonferroni tests failed to differentiate rhythms, and given the small amount of determinate errors, the main effect of rhythm may be more apparent than real.

	Complex	Harmonic	Periodic	Pitch C	Pitch H/L	Pulse	Transient	Unk.
Complex	**0**	.01	.03	.02	.02	0	0	.60
Harmonic	0	**0**	.04	0	.02	0	0	.56
Periodic	0	0	**0**	.02	.02	0	0	.47
Pitch C	0	0	0	**0**	0	.02	0	.24
Pitch H/L	0	.01	0	0	**.03**	0	0	.46
Pulse	0	.01	.02	.05	.01	**0**	0	.38
Transient	.01	.01	***.06***	.01	.03	.01	**.01**	.46

Table 3. Error matrix displaying the frequency of responses for each rhythm. Target rhythms appear in column 1, and response rhythms appear in row 1.

Discussion

Using the new coding schemes, several interesting finding emerged. For single words (MRT and PB), subjects' performance increased using the new 3-tier coding scheme over the absolute whole word correct coding scheme used previously. Performance was comparable on word initial and word final consonants for both MRT and PB words, indicating that subjects were equally likely to make errors in C1 and C2. Place of articulation errors were most common regardless of word type (MRT or PB) or consonantal position (word initial or word final). This is most likely due to the reduced spectral detail in the sinewave vocoded stimuli. These findings are similar to those of Shannon and colleagues (1995), who found that place errors were the most common for vocoded stimuli regardless of the number of channels used in synthesis or low pass filter cutoff frequencies used to derive the amplitude envelopes. The finding that subjects made few voicing errors overall is not surprising because temporal information, which provides many cues for voicing, was preserved through the use of the 400 Hz filter for envelope detection. This result is also similar to those of Shannon and colleagues (1995) who found that varying the amount of temporal information by using higher and lower cutoff frequencies for amplitude envelope detection altered the number of voicing errors made. When higher frequency filter cutoffs were used (preserving more temporal information) voicing errors were less prevalent than when lower frequency filter cutoffs were used (preserving less temporal information and increasing the number of voicing errors).

The difference between the performance on the two word sets and the types of errors that were made on each is interesting as well. Subjects made more varied errors (i.e., more errors of different types) on PB words despite performing better overall (69% correct) than on MRT words (63%). Subjects also showed less training specificity on PB words, which could be due to differences in phonetic balance between MRT (not phonetically balanced) and PB words (phonetically balanced relative to American English). The MRT words were composed of minimal pairs, varying along only one or two dimensions, and are therefore limited to words that have five other words that rhyme with them. By comparison, testing on PB words is probably a better assessment of an individual's open-set word recognition under degraded conditions, whereas testing on MRT words may provide a better assessment of feature discrimination and reception.

That subjects performed better on the contextually rich Harvard sentences (70.3% correct) than on the semantically anomalous sentences (55% correct) is unsurprising because subjects could use semantic context to make informed guesses if they were unsure of identity of the keyword. The coding scheme revealed that the majority of the keyword errors were phonetic errors and very few were lexical or thematic errors as was expected. The lack of thematic errors is particularly intriguing, since both the Harvard and the Anomalous sentences are grammatically licit. We predicted that subjects would make errors based on the semantic context of the surrounding words in the sentences, but this did not appear to be the case. Previous work with SPIN sentences has demonstrated that word predictability significantly influences sentence recognition (Kalikow, Stevens & Elliott, 1977). This predictability, however, was limited to the thematic relationship of the final word in the sentences to the preceding stem, not the thematic relationship between each word in the sentence relative to one another. Thus, it could be the case that when the target item is isolated in the sentence, predictability may constrain the response set, but when all words are thematically unrelated, predictability may not be invoked as a perceptual strategy. Further research will be necessary to determine the extent of the relationship between predictability and sentence semantic structure. In addition to there being few thematic errors overall, there were no differences between in thematic error prevalence between meaningful and anomalous sentences. The high number of phonetic errors on sentences as well as the large number of place errors on single word stimuli suggests that it would be effective to train people on phonetic contrasts when adapting to degraded stimuli.

Using the new 3-tier coding scheme, subjects' performance was much higher on environmental sounds than using the absolute scheme used previously. We initially designed this 3-tier coding scheme to provide more information about the cues that are important for the perception of environmental sounds. However, due to the high number of indeterminate errors, it is unclear how useful the present coding scheme will be to that end. The agent, or object or event that produced the sound, was the only valence that could be differentiated reliably under the new coding scheme (with 17% of the errors made being determinate errors). Both action and rhythm errors could only be reliably differentiated 8 and 6 percent of the time respectively. This is somewhat disappointing, since differentiating errors in rhythm was one of the reasons that we designed this coding scheme. Future work will re-examine the three valences to determine whether more powerful and reliable coding schemes can be devised to differentiate subject errors.

Investigating how normal hearing listeners adapt to cochlear implant simulations not only provides more information about speech perception under degraded conditions, but may have implications for rehabilitation strategies for new CI users. Moreover, many experiments with CI users utilize open-set recognition of stimuli where the subject listen to a target stimulus and verbally report what they perceived. The development of the coding strategies presented here will therefore be extremely useful to open-set studies with CI users. In an ongoing experiment in our lab, CI users are being tested on the same materials as reported on here so that a comparison can be made between the performance of normal hearing subjects on CI simulations and CI users themselves. The same coding strategies implemented here will be used to assess the types of errors that CI users make on these same stimuli to determine if CI users and normal hearing process these materials in similar ways. We hope that such a detailed error analysis will reveal differences in perceptual processing strategies that are used by the two groups of subjects, which would have important implications for training and rehabilitation paradigms for postlingually deafened cochlear implant users.

References

Aronson, J. (Producer and Director)(2000). Sound and Fury [Motion Picture]. United States: Aronson Films.

Clark, G.M. (2003). "Rehabilitation and Habilitation." In *Cochlear Implants: Fundamentals and Applications*. pp 654-706 New York NY: Springer-Verlag.

Dorman, M.F., Loizou, P.C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102,* 2403-2411.

Dorman, M.F., Loizou, P.C., Fitzke, J. & Tu, Z. (1998). The recognition of sentences in noise by normal hearing listeners using simulations of cochlear implant signal processors with 6-20 channels. *Journal of the Acoustical Society of America*, *104(6),* 3583-3585.

Egan, J.P. (1948). Articulation testing methods. *Laryngoscope*, *58,* 955-991.

Food and Drug Administration. (2004). Cochlear Implants: Frequently Asked Questions. Retrieved August 9, 2007, from http://www.fda.gov/cdrh/cochlear/faq.html.

Friesen, L.M., Shannon, R.V., & Cruz, R.J. (2005). Effects of stimulation rate on speech recognition with cochlear implants. *Audiology and Neuro-otology,* 10, 169-184.

Gygi, B., Kidd, R.R., & Watson, C.S. (2004). Spectral-temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America, 115(3),* 1252-1265.

Herman, R., & Pisoni, D.B. (2000). Perception of elliptical speech by an adult hearing-impaired listener with a cochlear implant: some preliminary findings on coarse-coding in speech perception. In *Research on Spoken Language Processing Progress Report No 24* pp 87-112 Bloomington IN: Speech Research Laboratory, Indiana University.

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37,* 158-66.

IEEE. (1969). IEEE recommended practice for speech quality measurements. IEEE No 297 New York: Author.

Kalikow, D.N., Stevens, K.N., & Elliot, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337-1351.

Karl, J.R. & Pisoni, D.B. (1994). Effects of stimulus variability on recall of spoken sentences: A first report. In *Research on Spoken Language Processing Progress Report No 19,* pp 145-193 Bloomington IN: Speech Research Laboratory, Indiana University.

Loebach, J.L., Bent, T., Peterson, N., Hay-McCutcheon, M. & Pisoni, D.B. (2008). The perception of speech and environmental sounds by normal hearing listeners and cochlear implant users. Poster to be presented at the 31st Annual Midwinter Meeting of the Association for Research in Otolaryngology, Phoenix, AZ.

Loebach, J.L. & Pisoni, D.B. (in press). Perceptual learning of spectrally degraded speech and environmental signals. *Journal of the Acoustical Society of America*.

Loebach, J.L & Pisoni, D.B. (2007). Perceptual learning under a cochlear implant simulation. In *Research on Spoken Language Processing Progress Report No. 28*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Marcell, M.M., Borella, D., Greene, M., Kerr, E. & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology, 22(6*), 830-864.

Munson, B.G.S., Donaldson, G.S., Allen, S.L., Collison, E.A., & Nelson, D.A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *Journal of the Acoustical Society of America*, *113(2),* 925-935.

National Institutes of Health (2007). "Cochlear Implants." Retrieved August 1, 2007, from http://www.nidcd.nih.gov/health/hearing/coch.asp.

Reed, C.M., & Delhorne, L.A. (2005). Reception of environmental sounds through cochlear implants. *Ear and Hearing, 26,* 1 48-61.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303-304.

Appendix A: Excerpted coding decisions for the environmental sound coding scheme

	Airplane'		'Baby crying'		'Banjo'		'Basketball'	
Airplane'	***Mo/Ru/PC***	***Mo/Ru/PC***	Mo/Ru/PC	Hu/B/PH	Mo/Ru/PC	St/P/Cx	Mo/Ru/PC	R/S/Pe
'Baby crying'	Hu/B/PH	Mo/Ru/PC	***Hu/B/PH***	***Hu/B/PH***	Hu/B/PH	St/P/Cx	Hu/B/PH	R/S/Pe
'Banjo'	St/P/Cx	Mo/Ru/PC	St/P/Cx	Hu/B/PH	***St/P/Cx***	***St/P/Cx***	St/P/Cx	R/S/Pe
'Basketball'	R/S/Pe	Mo/Ru/PC	R/S/Pe	Hu/B/PH	R/S/Pe	St/P/Cx	***R/S/Pe***	***R/S/Pe***
'Belch'	Hu/Ru/Pu	Mo/Ru/PC	Hu/Ru/Pu	Hu/B/PH	Hu/Ru/Pu	St/P/Cx	Hu/Ru/Pu	R/S/Pe
'Birds'	A/Bl/H	Mo/Ru/PC	A/Bl/H	Hu/B/PH	A/Bl/H	St/P/Cx	A/Bl/H	R/S/Pe
'Blinds closing'	Me,PL/SR,C/T	Mo/Ru/PC	Me,PL/SR,C/T	Hu/B/PH	Me,PL/SR,C/T	St/P/Cx	Me,PL/SR,C/T	R/S/Pe
'Boat horn'	Wi/Bl/Pl,Pu	Mo/Ru/PC	Wi/Bl/Pl,Pu	Hu/B/PH	Wi/Bl/Pl,Pu	St/P/Cx	Wi/Bl/Pl,Pu	R/S/Pe
'Bongos'	Wi/S/Pe	Mo/Ru/PC	Wi/S/Pe	Hu/B/PH	Wi/S/Pe	St/P/Cx	Wi/S/Pe	R/S/Pe
'Bowling'	Wo/R,Cr/T	Mo/Ru/PC	Wo/R,Cr/T	Hu/B/PH	Wo/R,Cr/T	St/P/Cx	Wo/R,Cr/T	R/S/Pe
'Camera'	Mo/Pl/Ru,Cr/PH	Mo/Ru/PC	Mo/Pl/Ru,Cr/PH	Hu/B/PH	Mo/Pl/Ru,Cr/PH	St/P/Cx	Mo/Pl/Ru,Cr/PH	R/S/Pe
'Can opening'	Me/B/PH	Mo/Ru/PC	Me/B/PH	Hu/B/PH	Me/B/PH	St/P/Cx	Me/B/PH	R/S/Pe
'Car crash'	Ru,Me/SS,Cr/Cx	Mo/Ru/PC	Ru,Me/SS,Cr/Cx	Hu/B/PH	Ru,Me/SS,Cr/Cx	St/P/Cx	Ru,Me/SS,Cr/Cx	R/S/Pe
'Car horn'	Wi/Bl/Pu	Mo/Ru/PC	Wi/Bl/Pu	Hu/B/PH	Wi/Bl/Pu	St/P/Cx	Wi/Bl/Pu	R/S/Pe
'Cash register'	Me/S,SR/Cx	Mo/Ru/PC	Me/S,SR/Cx	Hu/B/PH	Me/S,SR/Cx	St/P/Cx	Me/S,SR/Cx	R/S/Pe
'Cat meow'	A/Bl,Ru/PC	Mo/Ru/PC	A/Bl,Ru/PC	Hu/B/PH	A/Bl,Ru/PC	St/P/Cx	A/Bl,Ru/PC	R/S/Pe
'Chickens'	A/B/Cx	Mo/Ru/PC	A/B/Cx	Hu/B/PH	A/B/Cx	St/P/Cx	A/B/Cx	R/S/Pe
'Child coughing'	Hu/B/Pe	Mo/Ru/PC	Hu/B/Pe	Hu/B/Pe	Hu/B/Pe	St/P/Cx	Hu/B/Pe	R/S/Pe
Church bell'	Me/S/PL	Mo/Ru/PC	Me/S/PL	Hu/B/PH	Me/S/PL	St/P/Cx	Me/S/PL	R/S/Pe

Abbreviations used in the above table

A	Agent	Action	Rhythm/Pitch	W	Wood
An	Animal	B	Burst	Cx	Complex
G	Glass	Bl	Blow	H	Harmonic
Hu	Human	Bub	Bubble	PC	Pitch Change
I	Insect	Cr	Crash	Pe	Periodic
Lq	Liquid	P	Pluck	PH	Pitch High
Me	Metal	R	Roll	PL	Pitch Low
Mo	Motor	Ru	Rumble	Pu	Pulse
P	Paper	S	Strike	T	Transient
Pl	Plastic	SR	Slide Rough		
R	Rubber	SS	Slide Smooth		
St	String	T	Tear		
Wi	Wind	Z	Buzz		

RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 28 (2007)

Indiana University

New Directions in Speech Research[1]

Adam Buchwald, Tessa C. Bent, Christopher M. Conway, Susannah V. Levi and Jeremy L. Loebach

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] We would like to acknowledge NIH NIDCD grant number DC00012 for supporting the production of this paper.

New Directions in Speech Research

Abstract. In October 2006, many of the top scholars of speech perception research gathered in Bloomington, Indiana for a conference focused on new directions in speech research. This short paper provides a summary of the talks that were presented at this conference, which discussed the use of methodological innovations, novel theoretical frameworks, and the use of a variety of research populations in speech research.

Introduction

This paper provides a summary of talks presented at "PisoniFest," a conference held in Bloomington, Indiana on October 20-22, 2006 which explored new avenues and topics for research on speech perception. The talks at this conference focused on methodological innovations and concerns, the application of new – and used – theoretical frameworks to the study of speech perception, and the use of a variety of under-examined research populations in speech research. This paper provides synopses of each presented paper, organized by focus.

Methodological Innovations and Concerns in Speech Research

Cynthia Clopper, Ohio State University, and **Janet Pierrehumbert**, Northwestern University, examined the effects of dialect variation on spoken word recognition and lexical access. Previous research has shown a benefit of local over non-local dialect, as well as standard over non-standard dialect in word recognition. Clopper and Pierrehumbert extended this past work by examining potential phonological sources for the interaction between dialect variation and word recognition. Two dialects – Northern Cities (non-standard) and Midland (standard) – which exhibit acoustic-phonetic overlap of different vowel categories were selected for the studies. In the first experiment, listeners performed an open-set word recognition task with monosyllabic stimuli drawn from these two dialects. Systematic lexical confusions based on dialect differences were found, where words with greater acoustic-phonetic overlap between the dialects resulted in an increased number of lexical confusions. In the second experiment, listeners performed a speeded classification task ("bad" vs. "bed") for stimuli from these two dialects. Listeners were faster at categorizing the words from the Midland dialect than the Northern Cities dialect, confirming findings of previous studies which showed better performance on a standard dialect. The two studies reported by Clopper and Pierrehumbert provide an account of certain vowel confusions that listeners make when listening to non-standard dialects by considering acoustic-phonetic similarity of different vowel categories.

Mitchell S. Sommers, Washington University, presented a new measure of listening comprehension, LISN (lectures, interviews, spoken narratives), which can be used to assess listening comprehension in diverse populations including clinical and non-clinical listeners from a variety of age groups. LISN is part of a larger project investigating changes in cognitive abilities across the lifespan including measures of working memory, speech processing, and auditory processing. LISN includes three types of passages: lectures from the BBC, interviews from CSpan, and spoken narratives. Three types of questions were used to assess a listener's comprehension: information, integration, and inference. Thus far Sommers has used the test with normal hearing young and older adults. and hearing-impaired older adults. He presented three studies which used the LISN to assess listening comprehension in these populations. The first study revealed that younger listeners were better than older listeners in the listening comprehension tests but there was a great deal of variability depending on passage and question

type. In a second study, hearing impairment contributed marginally to a decline in listening comprehension. Lastly, an audiovisual (AV) task was used to assess the benefit of lipreading and AV integration in these three populations. In the AV task all participants performed similarly, suggesting that the addition of visual information eliminated the age differences shown in audio-only conditions. In the future, Sommers hopes to use the test with people with hearing loss, Alzheimer's disease, and aphasics.

Kevin Munhall, Queen's University, discussed "hot problems" in audiovisual speech perception. In particular, he focused his discussion on studies examining gaze fixation and duration during audiovisual (AV) speech perception. He also investigated audiovisual perception of animated speech to address the intelligibility "gain" an observer gets from receiving visual speech information in addition to auditory information. With respect to gaze duration and fixation, Munhall reported that these vary depending on the task as well as on viewer-specific biases. Munhall also reported that dynamic realism in animation is a necessary condition to generate AV gain in animated AV speech perception.

Jennifer Pardo, Barnard College, presented some recent research on the nature of phonetic convergence during conversational interaction. Pardo discussed the notion of accommodation as phonetic convergence (e.g., shifting vowel targets) towards an interlocutor, and suggested that this seemingly unconscious process may actually be a choice, as participants are likely to diverge away from an insulting experimenter. Pardo's research investigated phonetic convergence by comparing the pre- and post-interaction utterances of a speaker with that of their interlocutor, and found that a naïve set of participants were more likely to judge the post-interaction utterances to be similar to the interlocutor, with some interesting gender and task effects.

John Sidtis, New York University School of Medicine, discussed the limitations of fMRI research with respect to understanding more about speech processing. The central point in Sidtis' claim was that complex behaviors are not reliably decomposed by contrasting tasks, and imaging research often relies on this method of "cognitive subtraction." In particular, Sidtis warned against the use of "resting states" as controls for complex cognitive functions such as speech production or perception. Further, he presented his own research indicating that more blood flow may not always be a reliable indicator that a particular brain region area performs a specific function.

J.D. Trout, Loyola University – Chicago, spoke about the use of animal models in understanding human cognition. He focused on the dangers of the "possibility proof" methodology, in which scholars argue that something is not unique to humans because other animals can show the same behavior (e.g., Gentner, Fenn, Margoliash, & Nusbaum, 2006). Trout's critique centered on the claim that it is not clear that animal studies tap into the same skills that humans use when they are performing linguistic tasks. Trout cited common discrepancies in findings as evidence that experiments with animal populations and with humans may be tapping into different skills.

Luis Hernandez, Indiana University, gave an illuminating presentation on the reliability of collecting reaction time data on modern computer systems. Systems that rely on multi-tasking can be unreliable because of differences between the onset of execution and when the physical presentation occurs. An external microcontroller that is not system specific and does not rely on computer resource management was proposed as the best, most accurate and cost effective solution for experiments requiring fine temporal resolution.

Theoretical Frameworks for Studying Speech Perception: New and Used

Olaf Sporns, Indiana University, presented research on the connection between information theory and embodiment and demonstrated how such ideas provide a new understanding of how artificial and biological systems interact with the environment. Although there are different varieties of embodiment, Sporns suggested they all have in common at least three core concepts: the rejection of the idea that cognition is the processing of symbols; an emphasis on the dynamic coupling of organism to the environment; and a focus on development and self-organization. Pursuing work in robotics, Sporns' research investigates how an embodied agent interacting with the environment affects perceptual development. Using mathematically-defined information metrics such as entropy, mutual information, integration, and complexity, Sporns shows that embodied interactions affect the statistical structure of the organism's environment. That is, through its actions, an organism can "shape" its own environmental structure. As an example, Sporns demonstrated a simple robotic active vision system, in which a camera samples visual information, actively adjusting the camera to focus on particular salient parts of the scene (e.g., the color red). The coupling between the robot's action and perception systems was manipulated, with the results showing that decoupling produces less information structure. In sum, understanding how embodied systems benefit from environmental interaction and the coupling of perception/action systems can provide important new insights into the nature of speech perception, which has traditionally been dominated by a classic, information-processing view of perception.

Geoff Bingham, Indiana University, gave an interesting presentation on the underlying tenets of Gibson's theory of direct perception, and demonstrated how it can account for many aspects of perception. In this framework, events in the world are conceived of as spatio-temporal objects that are constrained by the environment. Under Gibson's theory, both humans and animals detect events and objects by recognizing patterns of information specified in the dynamics in the environment. Using point light displays, Bingham illustrated that we can recognize a variety of events, both animate and inanimate, based solely on the dynamics of their movement. Moreover, recognition accuracy is disrupted when the dynamics are altered such that the information specified by them becomes inconsistent with our ecological point of view. Bingham argued that the perception of biological motion, therefore, is not special, in that we can recognize the motion of a variety of objects (both animate and inanimate) even though we may not be able to produce the actions ourselves. In addition, referencing others motion to our own motion is inadequate in that we cannot ourselves witness the motions that we produce under normal circumstances. Moreover, theories that specify a motor code in the recognition of events are incomplete because they would apply only to humans (not other animals or inanimate objects), overestimate the role of the motor code in generating movement (such a code is merely correlated with the motions), and severely underestimate the role of perception. Bingham concluded that we perceive information in our environment, not our motor systems, as is argued by proponents of the Motor Theory of speech perception.

Nelson Cowan, University of Missouri, presented work investigating short-term memory (STM) and forgetting, where STM is informally defined as the small amount of information one can hold in mind for a short period of time. Cowan described a seminal paper by Pisoni (1973) that led Cowan to investigate several important questions about the nature of STM, especially for acoustic and phonetic input: What happens to STM codes over time? What is the role of attention in STM? Is STM memory lost through decay or interference? To investigate the first question, Cowan described research examining memory for vowels, which suggests that forgetting results in an expansion of the uncertainty of the sound. That is, the representation of a particular vowel "slides" toward the average vowel sound located in the middle of vowel space. Thus, Cowan argues that forgetting involves a shift of the memory code toward the average or prototypical representation of that class of sounds. To explore the second

question, Cowan presented work showing that attention is necessary in order to get a stable representation of a phonetic code. Finally, to address the third question, Cowan presented evidence arguing that forgetting may involve a combination of proactive interference and a "sudden" loss of the memory code after a particular amount of time (as opposed to a gradual decay). This work investigating STM and forgetting is important because it helps to clarify the role of memory and cognition in the perception and representation of speech sounds.

Robert Port, Indiana University, discussed his new proposal of "phonology with rich memory." Port argued against the traditional notion of language as a symbol system in which we store mental representations corresponding to sound structure units such as phonemes, phones or segmental features. Instead, Port contended that our mental representations of language consist of the exemplars that we have encountered and encoded. Evidence for this assertion comes from a variety of studies demonstrating that we store and are able to use episodic information. Examples of this occur when participants perform better in a recognition memory task when a word is produced by the same speaker during familiarization and testing (e.g., Palmeri, Goldinger, & Pisoni, 1993). Port strongly argued that the existence of an episodic store of phonological events is incompatible with the traditional linguistic view that we store abstractions over those exemplars. He claimed that part of the reason we are drawn to the notions of these abstract units that compose words is our alphabetic training (an argument famously proposed by Ladefoged, 1980), and cited work on non-literate individuals suggesting that their "phonological awareness" (as defined by segmental awareness) is impoverished compared to literate individuals. He ended with the assertion that traditional phonology is necessary to describe socially agreed-upon linguistic conventions ("social phonology") but that understanding the language processing system can only be done by examining episodic memory.

Robert E. Remez, Columbia University and Barnard College, discussed a neglected problem in speech perception research: How do human listeners determine which auditory inputs should be processed as speech? Most theories of speech perception start with a listener's analysis of the speech signal rather than starting with an analysis of the complete auditory input. Remez argued that deciding which inputs to process as speech is not a trivial problem/task. For example, listeners must determine which parts of the auditory signal are relevant speech samples to be analyzed, which are complex non-speech signals, and which are irrelevant speech samples (e.g., background talkers). Certain technologies (e.g., ViaVoice) and models of speech perception (such as TRACE) match all auditory signals to stored speech templates thereby translating non-speech sounds to the closest speech equivalents. However, early perceptual processes must help listeners distinguish auditory signals from known languages, unknown languages, and non-speech sources. This early stage of processing is necessary for listeners to know what to process as speech. Furthermore, Remez pointed out that the perceptual system is highly flexible; even inputs that do not closely match stored speech templates can be perceived as speech if listeners are told that the signal is speech. For example, sinewave speech is often initially perceived as non-speech but listeners can also extract linguistic and extra-linguistic content from signals once they are told that the signal is speech. Remez concluded that theories of speech perception must consider how listeners determine which auditory inputs to process as speech. Listeners must simultaneously exclude complex speech-like auditory signals that are not speech and include signals that differ from naturally produced speech but can be processed as speech.

Examining Under-Examined Research Populations

Robert Shannon, House Ear Institute, described some of his recent work that patients with Auditory Brainstem Implants. Although cochlear implants have been extraordinarily successful, not everyone with severe hearing impairment is a candidate for cochlear implantation. In some cases, severe

head trauma can sever the auditory nerve, making cochlear implantation impossible. In other cases a genetic disorder can lead to the growth of NF2 tumors on the vestibular branch of the auditory nerve, requiring surgical removal of both tumor and nerve. For such individuals, Auditory Brainstem Implants (ABIs) may be an option. Rather than inserting electrodes into the cochlea, electrode arrays are inserted into the peripheral layers of the cochlear nucleus, the first stop in the ascending auditory pathway. ABI recipients, however, show a mixed pattern of results depending on etiology. Individuals with head trauma show levels of speech recognition comparable to many cochlear implant users. NF2 patients, however, show very poor speech recognition abilities (0-20% correct). In both cases, subjects have access to all of the necessary perceptual information (what Shannon calls "bits"), but only the patients without NF2 tumors can correctly assemble such information to provide high levels of speech recognition. Shannon argued that the progressive growth of NF2 tumors destroy neurons that are vital to the organization of the cochlear nucleus. These low spontaneous rate/high threshold neurons respond to sound over a large dynamic range, and provide input to the small cell cap area of the cochlear nucleus, an area that is particularly sensitive to temporal modulations. Both NF2 and head trauma patients show normal frequency mapping along the tonotopic axis in the cochlear nucleus under ABI stimulation, suggesting that high spontaneous rate/low threshold neurons may be more involved in pitch perception and sound localization. Shannon concluded that the gradual destruction of the auditory nerve due to NF2 tumor growth disrupts the organization of the ascending auditory pathway, limiting patients' capacity to develop high levels of speech perception abilities.

Diana Van Lancker Sidtis, New York University, discussed the perception of voice characteristics. She argued that the representation and processing of voice information is more similar to face processing than speech processing. For example, voice and face information both contain keys to personal identity and show hemispheric specialization. Processing prosodic information – crucial for the perception of voice identification – involves the perception of both timing and pitch. Van Lancker Sidtis presented data from two patients who showed differential loss of timing and pitch information. The first patient produced appropriate timing in both singing and speech, but was unable to produce pitch differences in either task. In contrast, the second patient produced accurate timing, and also produced accurate pitch when singing, though not in speech. Van Lancker Sidtis argued that data from these patients confirmed hemispheric specialization for processing these two aspects of prosody; timing information is processed by the left hemisphere, whereas and pitch information is processed by the right hemisphere. She also summarized data from fMRI, ERP, and lesion studies showing a right hemisphere advantage in processing familiar voices. Using these data, Van Lancker Sidtis concluded that voice information and speech information are processed in different hemispheres.

Rosalie Uchanski presented work on the identification and discrimination of emotions in American English speaking cochlear implant users compared to normal hearing adults and children. Cochlear implant (CI) users had more difficulty than the normal hearing adults or children. In particular, the CI-users frequently confused fearful and happy productions.

Mario Svirsky, New York University, discussed his work on frequency mismatch and spectral degradation in cochlear implant simulations with normal hearing adults. Adaptation to spectral shift was facilitated when the shift was gradually introduced over a series of training sessions. This work suggests that CI-users may benefit from self-selected tuning of electrode mapping.

References

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature, 440*, 1204-1207.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language*, 56, 485-502.

Palmeri, T. J., Goldinger, S. R., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 309-328.

Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*, 253-260.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Integrating Auditory and Visual Information in Speech Perception: Audiovisual Phonological Fusion[1]

Joshua L. Radicke, Susannah V. Levi, Jeremy L. Loebach and David B. Pisoni

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This work supported by NIH-NIDCD Training Grant T32-DC00012 and NIH-NIDCD Research Grant R01-DC00111. We would like to Luis Hernandez for providing technical assistance and advice in the design and implementation of the experimental procedures.

Integrating Auditory and Visual Information in Speech Perception: Audiovisual Phonological Fusion

Abstract. Phonological Fusion is a phenomenon in which different phonemes are presented to each ear, prompting the listener to perceive a blend of the two (e.g., /ba/+/la/=/bla/). The present study assessed whether Phonological Fusion has an audiovisual analogue. That is, when listeners are presented with video clips containing visual stop consonants paired with auditory liquids, do they integrate information across the two modalities as they do in unimodal Phonological Fusion (e.g., visual "back" + auditory "lack" = "black")? The three experiments presented here demonstrate that Audiovisual Phonological Fusion does occur, but primarily for visual bilabial stop consonants (e.g., /b/ and /p/) paired with the auditory liquid /l/. Moreover, the overall rate of fusion is determined by the lexicality of the target word. Taken together, the results of the preset study suggest that similar processes underlie both unimodal and multimodal fusion, suggesting a general mechanism for conflict resolution in speech perception.

Introduction

Integration of sensory information, whether within or across modalities, is necessary for successful interaction with the environment. The fusion of two separate pieces of information to form a single object or event is common to perceptual tasks. Fusions frequently occur in speech perception, and can be unimodal or multimodal. Dichotic listening fusions, such as Phonological Fusion, are a form of auditory unimodal fusion, and demonstrate how the integration of information between the two ears can modify the percept of an auditory stimulus. When each ear is presented with a different speech sound (e.g., /ba/ and /la/), the resulting percept is a combination of the two streams of phonetic information (e.g., /bla/)(Cutting, 1976). Similarly, multimodal fusions in speech perception occur when different sources of visual and auditory phonetic information are integrated (e.g., visual /ba/ auditory /ga/), resulting in the perception of an average of the two streams (e.g., /da/)(McGurk & MacDonald, 1976). Although most forms of multimodal fusions in speech perception have a unimodal analogue, suggesting that there is a general set of rules that governs both, not all fusions have been tested in both domains. The present study, therefore, assessed whether auditory Phonological Fusion has a multimodal analogue, and investigated some of the conditions under which it may arise.

Much of the work on unimodal fusions in speech perception comes from the work of Cutting (1976). Utilizing the dichotic listening paradigm in which different sources of auditory information are presented to each ear of a listener, Cutting described six prominent unimodal fusions in speech perception: Sound Localization, Psychoacoustic Fusion, Spectral Fusion, Spectral/temporal Fusion, Phonetic Feature Fusion, and Phonological Fusion. Although many of these unimodal fusions also have multimodal analogues (Sound Localization, and Phonetic Feature Fusion), not all have been investigated. The present study sought to assess whether Phonological Fusion has a multimodal analogue.

Sound Localization Fusions occur when two speech sounds, whose onsets are temporally asynchronous, are presented dichotically, resulting in the percept of a single sound originating from a particular location in space (Cutting, 1976). In this case, the perceived azimuthal location of the speech sound is determined by the temporal synchrony of the two stimuli: if the sounds arrive at each ear at a different time, the resulting percept is of two separate sounds originating in two separate locations (Cutting, 1976). The multimodal analogue of the Sound Localization Fusion is the Ventriloquist Effect

(Bertelson, Vroomen, Gelder & Driver, 2000). The Ventriloquist Effect is a multimodal fusion in which the perceived location of an auditory stimulus is altered by the presence of a salient visual stimulus (Bertelson *et al.,* 2000). In this case, an auditory stimulus is produced by one sound source (e.g., the human ventriloquist) but attributed to originating from a different location due to the salient visual information (e.g., the movements of the inanimate doll's mouth). In both cases, two separate pieces of information are combined to determine the percept of the location of the sound source.

Psychoacoustic Fusion (Cutting, 1976) is a dichotic listening fusion in which two different phonemes are presented to each ear, resulting in the percept of a single phoneme that is the "average" of the other two. For example, /ba/ is presented to one ear, and /ga/ to the other, resulting in the fused percept of /da/ (Cutting, 1976). An analogous multimodal fusion is the well-known McGurk Effect (McGurk & MacDonald, 1976), in which auditory and visual speech information mismatch, eliciting the percept of something that is an average of the two. When subjects were presented simultaneously with a video clip of a talker producing a velar stop (e.g., "gaga") and an audio track of a talker producing a bilabial stop (e.g., "baba"), the most common percept reported is an alveolar stop that is an average of the two (e.g., "dada"). The perceptual fusions under Phonetic Feature Fusion and the McGurk Effect do not simply combine information across the two sources of input: rather subsegmental information is fused to form a single segment that is not completely specified by either source alone.

Another dichotic listening fusion reported by Cutting (1976) is Phonological Fusion, which occurs when two different sounds are presented to each ear resulting in a percept that is a combination of the two. In this case, one consonant is presented to one ear (e.g., /ba/), and a different consonant to the other (e.g., /la/) resulting in the percept that is the combination of the two (e.g., /bla/). Cutting defines Phonological Fusion as "when two inputs, each of n phonemes, yield a response of $n + 1$ phonemes." (Cutting, 1976, p 121). In the case of the example, each of the inputs has 2 phonemes, but the response has 3 (n+1) phonemes. In other words, two consonants presented in two different auditory streams can be fused to form the percept of a consonant cluster.

Although Phonological Fusion has not been experimentally assessed in the multimodal domain, some evidence for its existence comes from McGurk and MacDonald (1976). In the original configuration, visual velar consonants paired with auditory bilabials result in an averaged percept (e.g., an alveolar consonant). However, when the configuration was reversed, and subjects were presented with a visual bilabial (e.g., "baba") and an auditory velar (e.g., "gaga"), a Combination Response occurred (such as "gabga", "bagba", "gaba", or "baga"). Although it is a perceptual fusion, the Combination Response appears to be governed by different rules than the true McGurk Effect. Combination Responses were given much less frequently than true McGurk Effect responses (90% for the McGurk response, versus 49% for the Combination Response), and could evoke four different possible percepts as opposed to just one. The frequency of occurrence and number of possible unique percepts suggests that the level at which the information is integrated is different for the McGurk Effect and the Combination Response. In the McGurk Effect, neither the auditory nor visual information is present in the final response; instead a fusion occurs at the featural level. The Combination Response, on the other hand, contains a sequence of phonetic segments that are specified by both visual and auditory streams, and appears to be more similar to auditory Phonological Fusion rather than Psychoacoustic Fusion.

Although the Combination Response was observed by McGurk and MacDonald (1976), a true multimodal analogue to unimodal Phonological Fusion has not been documented. If other unimodal fusions (e.g., Sound Localization and Psychoacoustic Fusion) have multimodal analogues (e.g., Ventriloquist Effect and McGurk Effect), it appears likely that a multimodal fusion corresponding to unimodal Phonological Fusion also exists. The purpose of the present study was to investigate whether

Phonological Fusion does indeed have a true multimodal analogue (Audio Visual Phonological Fusion or AVPF).

Understanding the conditions under which different fusions occur may provide additional insight into the integration of conflicting information for speech perception. In the case of unimodal Phonological Fusion, conflict between the two ears leads to a blending of the auditory streams, whereas in the McGurk effect, conflict between the auditory and visual modalities leads to an averaging of the auditory and visual streams. Understanding how the brain resolves conflicting information may lead to a better understanding of the operation of the perceptual processes underlying both unimodal and multimodal speech perception, and may suggest that a common domain general substrate.

Here, we report three experiments that pair visual stop consonants (e.g., "back") with auditory liquids (e.g., "lack") to determine whether true phonological fusions are reported (e.g., "black"). The first experiment used an open set recognition task to assess the occurrence of AVPF by presenting listeners with initial stop consonants at three places of articulation (bilabial, alveolar and velar), paired with one of two liquids (/l/ and /r/). In the second experiment, we assessed the effects of lexicality on open set recognition by comparing the rate of AVPF in words and nonwords. In the third experiment, we replicated the findings of the first two using a closed-set of response alternatives to further constrain the possibility of observing AVPF in other contexts.

Experiment 1

Methods

Stimulus Materials

The words and nonwords used as stimuli in all three experiments were modeled after those used by Cutting and Day (1975) and Cutting (1975; 1976) in dichotic listening experiments and by McGurk and MacDonald (1976). These specific stimuli were selected because they had been shown to successfully elicit either unimodal Phonological Fusion or the McGurk Effect. The specific stimulus set used is shown in Table 1. These stimuli were selected to compare fusion rates across three different places of articulation (bilabial, alveolar, velar) and across the two liquids /l/ and /r/.

Place of Articulation	Visual Stop	Auditory liquid		Fused response	
		/l/	/r/	stop + /l/	stop + /r/
Bilabial	back	lack	rack	black	brack
	pay	lay	ray	play	pray
Alveolar	dead	led	red	dled	dread
	tie	lie	rye	tly	try
Velar	go	low	row	glow	grow
	camp	lamp	ramp	clamp	cramp

Table 1: Stimulus set used in Experiment 1

All stimuli were recorded using a Canon GL1 video camera and lapel microphone (Shure mx-100). The talker was a male, native English speaker, who reported no history of speech or hearing disorders at the time of testing. A computer screen located below the camera displayed the words for the talker to read. One researcher operated the camera, while another monitored for pronunciation errors.

Any errors in pronunciation were noted, and the talker asked to repeat the mispronounced words at the end of the session. Two repetitions of the stimulus materials were originally recorded.

Stimuli were edited using Final Cut Pro 5.0.1 on a Macintosh Powerbook G4. The beginning and ending of each stimulus were identified using both visual inspection of the waveform and auditory discretion. To create the experimental files for presentation, an additional 15 frames (approximately 500 ms) was added before and after the target stimulus. When this method resulted in an unusual beginning or ending of the visual display (e.g., blinking) an additional frame was appended or an extra frame was removed.

Congruent stimuli were used as control items, and contained the same auditory and visual target (e.g., auditory "lack" and visual "lack"). Incongruent stimuli were created by pairing the auditory stimulus from one recording (e.g., "lack") with the visual signals from a different recording (e.g., "back"). All permutations of auditory and visual signals were created for each syllable (5 congruent and 20 incongruent stimuli). All incongruent stimuli attempted to splice utterances with the closest duration. When durations did not match exactly, the beginnings of the stimuli were aligned, and overall differences in duration for the two constituent portions of a stimulus never exceeded two frames (approximately 67 ms). Additionally, audiovisual and dichotic listening fusions have been shown to be robust across a relatively large window of asynchrony, from 30 milliseconds auditory lead to 175 milliseconds visual lead (van Wassenhove, Grant & Poeppel, 2006); thus slight temporal asynchronies in the onset and offset of the stimuli should not affect of the experimental results.

Participants

Twenty-five undergraduate students at Indiana University participated in the study. Each received partial course credit for their participation. All were native speakers of English, reported no history of speech or hearing impairment, and had normal or corrected-to-normal vision.

Procedure

The experiment took place in a quiet room with multiple testing stations. The experiment was run using a custom script written for PsyScript on four Macintosh G3 computers. Participants viewed stimuli on fifteen-inch CRT monitors and listened through Beyer Dynamic DT-100 headphones. Stimuli were presented at a comfortable listening level (approximately 65 dBA) for all participants.

Stimulus presentation was blocked in order to eliminate the potential for repetition priming. Stimuli were randomized within each block, but the blocks were always presented in the same order. The first block consisted of 36 Incongruent stimuli in which neither constituent contained a cluster (e.g., visual "back", auditory "lack"). The second block consisted of 84 Incongruent stimuli in which one or both of the constituents contained a cluster (e.g., visual "black", auditory "lack"). The third block served as a control, and consisted of 30 Congruent stimuli (e.g., visual "lack", auditory "lack").

Participants were instructed to both watch and listen during the experiment and great care was taken not to bias their attention to either modality. A fixation cross at the center of the screen preceded each stimulus and was followed by a 500 ms delay. A dialog box appeared directly after each stimulus, prompting subjects to type what they thought the talker said. They were told that their responses could be real words or nonwords, and were encouraged to check that they had typed their intended response to minimize typographic errors. No time limit was imposed for a participant's response, and each trial was separated by a 1000 ms intertrial interval.

Data analysis

For the purposes of data analysis, only incongruent stimuli containing a visual stop (e.g., "back", "tie") and an auditory liquid (e.g., "lack", "rye") were included in the experimental stimuli (n=12). Congruent stimuli containing a liquid (e.g., "lack", "rye", etc.) served as the control stimuli (n=12). These specific stimuli were selected in order to assess the frequency of occurrence of AVPF and to explore the effects of place of articulation and liquid category on fusion rate. Subjects' responses were coded as either containing a consonant cluster or not. Because the perception of stop voicing is virtually imperceptible from visual information alone, both voiced and voiceless clusters were considered acceptable fusions, regardless of whether the actual stimulus was a voiced or voiceless bilabial (e.g., "black" or "plaque" would both be coded as a fused response for visual "back" and auditory "lack"). Other responses that did not contain a consonant cluster such as "back", "lack", or "mack" were coded as containing no fusion.

Results

A repeated measures ANOVA with stop Place of Articulation (bilabial, alveolar, velar), Liquid Type (/l/, /r/), and Experimental Condition (Congruent, Incongruent) as within-subjects factors was conducted on the data. The ANOVA revealed significant main effects of Place of Articulation ($F(2,46) = 36.33$, $p \leq 0.001$), Liquid Type ($F(1,23) = 26.01$, $p \leq 0.001$), and Experimental Condition ($F(1,23) = 30.67$, $p \leq 0.001$). More fusions were reported for bilabial stops than for alveolar or velar places of articulation (Figure 1). Similarly, across all places of articulation, more fusions occurred for /l/ than for /r/. Finally, more fusions were reported for the experimental condition (e.g., Incongruent stimuli) than the control condition (e.g., Congruent stimuli), indicating that listeners do reliably perceive AVPF in incongruent stimuli with visual stops and auditory liquids.

In addition to these main effects, all two-way and three-way interactions reached significance (all *p-values* ≤ 0.001). All of these results, however, were driven by the presence of fusions in stimuli with a visual bilabial and an auditory /l/, as illustrated in Figure 1. Post-hoc paired-sample t-tests revealed that fusions in the bilabial + /l/ experimental condition were significantly greater than the fusions in the control /l/ + /l/ condition (48% vs. 2% respectively) ($p \leq 0.001$). For all other places of articulation and liquids, the difference between experimental and control conditions did not reach significance (all *p-values* ≥ 0.162). Although some fusions were recorded in the bilabial + /r/ experimental stimuli, the fusion rate did not differ significantly from the controls (4% vs. 0%).

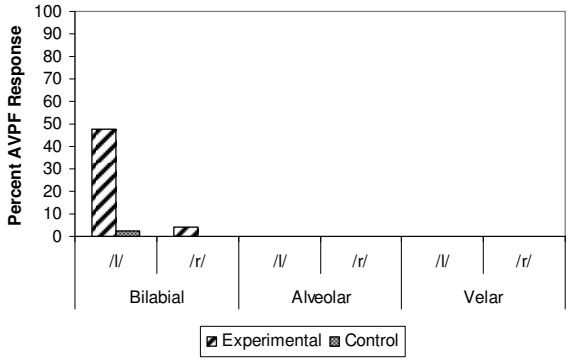


Figure 1. Percent of AVPF responses for visually presented stops at different places of articulation and auditorily presented liquids compared to controls in Experiment 1.

Discussion

Experiment 1 demonstrated the existence of AVPF, but its occurrence was shown to be limited to certain specific conditions. In particular, when a visual bilabial stop (i.e., /b/ or /p/) is presented simultaneously with an auditory /l/, a consonant cluster, /bl/ or /pl/, was perceived. AVPF was not observed when the visual signal contained stop consonants at the other two places of articulation (i.e., /d/, /t/, /g/, or /k/). Additionally, AVPF did not occur with /r/ even when paired with a visual bilabial stop. It is possible that the lack of fused responses with alveolar and velar stops is due to the absence of salient visual cues to these places of articulation. More importantly, stops at these places of articulation are not visually distinctive from the liquid /l/. All of the results found in Experiment 1, and those of other audiovisual fusions, can be explained by the degree and type of conflict between the auditory visual information.

Results from McGurk and MacDonald (1976) and this experiment suggest that AVPF occurs when the information from the two modalities conflict. The resolution of these conflicts yields a robust percept that includes phonetic attributes of both the visual and auditory information. In AVPF, a visual /b/ conflicts with an auditory /l/. The visual information of /b/ strongly specifies the presence of a bilabial, but no bilabial cues are present in the auditory stream of information. This conflict results in the percept of a cluster of phonemes, a serial combination of the phonemes from the visual and auditory streams of information.

Audiovisual conflict, or lack thereof, also explains the absence of AVPF with velars. The visual cues to /k/, /g/, and /l/ are not as perceptually distinct from each other as bilabials and /l/ are. Because the auditory and visual cues do not conflict, AVPF is not observed with visual velar stops paired with auditory /l/. Similarly, a lack of conflict can explain the absence of AVPF with alveolar stops, however, the lack of AVPF with alveolars and /l/ is more likely due to the phonotactics of English which prohibit clusters of this type in word-initial position (i.e., */tl/ and */dl/). This explanation suggests that listener's responses are highly constrained by their knowledge of phonology.

In addition to examining the effect of place of articulation on AVPF, Experiment 1 also assessed differences between the two liquids /r/ and /l/. Whereas AVPF was present for /l/ paired with bilabials, it was not observed at a significant rate for /r/ paired with any stops. The lack of AVPF with the liquid /r/ may have been due to its visual properties. In English, /r/ is produced with lip rounding. The bilabial stops are also produced with a labial gesture, although in this case, the gesture is complete lip closure. Because both bilabial stops and /r/ are produced with a labial articulation, less conflict is present between these segments than there is between the bilabial stop and /l/. We hypothesize that the occurrence of AVPF is dependent on the degree of difference between expected and actual visual information. Both bilabial stops and /r/ are produced with labial gestures, so there is less conflict, and thus no fusion is observed.

Although some degree of conflict exists between auditory /r/, which includes a labial gesture, and velar and alveolar stops, which lack a labial gesture, the type of conflict is quite different than for visual bilabial stops and auditory /l/. In the latter combination, where AVPF occurs reliably, the cues to labiality come from the visual information. In the former, similar to the McGurk Effect, the cues to labiality are specified in the auditory stream. In these cases, an auditory "labial" conflicts with a visual "non-labial". The visual cues provide information for a non-visually salient articulation, which is highly similar to the visual information of other segments (e.g., /d/). In the McGurk Effect, the conflict between auditory and

visual information results in the percept of a phoneme that is not labial, matching the visual information, but is acoustically similar (i.e., /da/).

The results from Experiment 1 and previous audiovisual studies support an interpretation of perceptual fusion that results from perceptual conflict. When conflict is high, observers perceive an utterance that fuses the information either at a subsegmental, featural level or at a segmental, syllable level. In Experiment 2, we sought to further explore the effect of perceptual conflict in AVPF and replicate the results found in Experiment 1. Since some fusions with /r/ were reported, we increased the number of trials in Experiment 2 to examine the frequency of AVPF with /r/. In addition, some of the fused responses in Experiment 1 produced valid English words. In Experiment 2 we examined the effect of lexicality on AVPF by including both word and nonword stimuli. Since no fusions were reported for the alveolar and velar stop consonants, Experiment 2 examined only performance on bilabial stops.

Experiment 2

Methods

Stimulus Materials

Stimuli for Experiment 2 (Table 2) were drawn from the same set of materials used in Experiment 1, but included both words and nonwords in order to test the effects of lexicality. These stimuli were selected to compare the rate of AVPF in the two liquid conditions (/l/ and /r/) and in both words and nonwords.

Lexicality	Visual Stop	Auditory liquid		Fused response	
		/l/	/r/	stop + /l/	stop + /r/
Word	back	lack	rack	black	brack
	pay	lay	ray	play	pray
Nonword	baba	lala	rara	blabla	brabra
	papa	lala	rara	plapla	prapra

Table 2: Stimulus set used in Experiment 2.

Participants

Twenty-six participants took part in this experiment. All participants were undergraduate students at Indiana University and either received partial course credit or were paid $10.00 per hour for their participation. All participants were native speakers of English, reported no history of speech or hearing impairment, and had normal or corrected-to-normal vision.

Procedure

The experiment took place in a quiet room with multiple testing stations. The program for the experiment was written in PsyScript and was run on Macintosh G3 computers. Participants viewed stimuli on CRT monitors and listened through Beyer Dynamic DT-100 headphones. Stimuli were presented at a comfortable listening level (approximately 65 dBA) for all participants.

Experiment 2 was divided into three blocks. Blocks were very similar to those in Experiment 1, except that each stimulus was presented twice within each block. The first block consisted of 48 incongruent stimuli in which neither constituent contained a cluster (e.g., visual "back", auditory "lack"). The second block consisted of 112 incongruent stimuli in which one or both of the constituents contained a cluster (e.g., visual "black", auditory "lack"). The third block consisted of 40 congruent stimuli and served as the control block (e.g., visual "lack", auditory "lack"). Stimuli were randomized within these blocks, but the blocks were always presented in the same order to eliminate the potential for repetition priming by cluster-initial words earlier in the experiment.

Participants were instructed to both watch and listen during the experiment and great care was taken not to bias their attention to either modality. A fixation cross at the center of the screen preceded each stimulus and was followed by a 500 ms delay. A dialog box appeared directly after each stimulus, prompting subjects to type what they thought the talker said. They were told that their responses could be real words or nonwords, and were encouraged to check that they had typed their intended response to minimize typographic errors. No time limit was imposed for a participant's response, and there was a one second intertrial interval.

Data Analysis

For the purposes of data analysis, only incongruent stimuli with a visual stop (e.g., "back") and an auditory liquid (e.g., "lack" or "rack") were included in the experimental stimuli (n=8). Congruent stimuli with a liquid (e.g., "lack", "ray", etc.) served as the control stimuli (n=8). These two sets were selected to determine the degree of AVPF in the two liquid conditions and the two lexical conditions since all fused responses are valid English words. For words, participant responses were coded as either resulting in a consonant cluster (ignoring voicing alternations) or resulting in no consonant cluster. Other responses with no consonant cluster such as "pay" or "lay" were coded as non-fusions. For nonword stimuli, participant responses were coded using both a stringent and a lenient measure. The stringent measure scored a fusion as successful only if fusions were reported at both word-initial and intervocalic positions (e.g., "blabla"), whereas the lenient standard scored a fusion as successful if a fusion occurred at either position (e.g., "blaba", "babla", and "blabla"). Both measures are reported below, although we focus on the lenient measure.

Results

A 2x2x2 repeated measures ANOVA with Lexicality (word, nonword), Liquid Type (/l/, /r/), and Experimental Condition (experimental, control) as within-subjects factors was conducted on the data. The ANOVA revealed significant main effects of Lexicality ($F(1, 25) = 10.155$, $p = 0.004$), Liquid Type ($F(1, 25) = 29.436$, $p < 0.001$), and Experimental Condition ($F(1, 25) = 35.708$, $p < 0.001$). More fusions were reported in response to words than to nonwords. Similarly, more fusions occurred for /l/ than for /r/. Finally, more fusions were reported in the experimental condition than in the control condition, indicating that listeners perceived AVPF when presented with incongruent stimuli consisting of visual bilabial stops and auditory liquids. Results are summarized in Figure 2.

In addition to the main effects, all two-way interactions reached significance (all *p-values* $\leq$ 0.016). The three-way interaction also reached significance ($p = 0.027$). As in Experiment 1, all effects were driven by the large number of reported fusions in the word bilabial plus /l/ experimental condition. Post-hoc paired samples t-tests revealed significant differences between the experimental and control conditions only for the word ($p \leq 0.001$) and nonword ($p = 0.015$) stimuli consisting of visual bilabials and auditory /l/ (36.54% vs. 5.77% for words and 11.54% vs. 0.00% for nonwords). However, using the

stringent coding method described above, the rate of fusion in response to nonwords containing visual bilabials and auditory /l/ was not significantly greater than in the control condition (p = 0.212, 3.85% vs. 0.00%). No other conditions produced fusions that were significantly different from zero (all *p-values* ≥ 0.185). Although some fusions were observed in the word bilabial + /r/ experimental condition, they did not differ significantly from the control condition (2.88% vs. 0%). This result supports the finding from Experiment 1 that auditory /r/ does not induce the perception of AVPF. Additionally, post-hoc comparisons between the four experimental conditions revealed that the rate of AVPF in words with bilabials + /l/ condition was greater than in all other experimental conditions (all *p-values* ≤0.004).

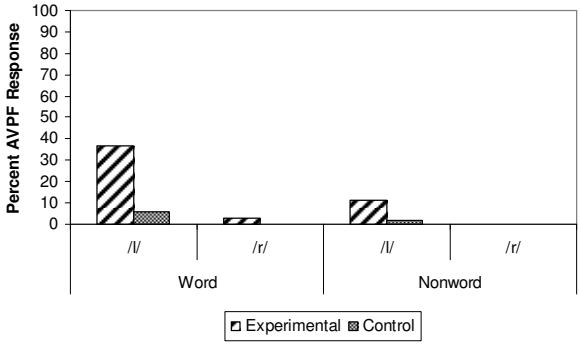


Figure 2. Percent of fused (cluster) responses for visual stops and auditory liquids in monosyllabic words and disyllabic nonwords in Experiment 2.

Discussion

The findings from Experiment 2 replicate and extend the findings from Experiment 1. AVPF was again observed with visual bilabial stops and auditory /l/. Lexicality affected the rate of fusion as AVPF was perceived more often in response to words than to nonwords. AVPF did not occur at a significant rate in response to words or nonwords with the liquid /r/. However, for the trials containing words with the liquid /r/ some fusions were reported, but not enough to reach statistical significance.

Although AVPF did occur in response nonwords, the fusion was much less prevalent than it was in response to words. High-frequency words are perceived more easily than low-frequency words (Broadbent, 1967), so by extension, nonwords, which essentially have a lexical frequency of zero, should produce fewer fusions than words. Interestingly, some of the monosyllabic words resulted in fused percepts that were nonwords (e.g., “blay”) even though their constituents were words (e.g., “pay”, “lay”). This discrepancy suggests that the lower rate of fusion is not caused by the target fusions being nonwords, but rather because the constituents are nonwords.

Since both Experiments 1 and 2 used an open-set response format, it is possible that additional fusions could occur, but that they were not being reported frequently enough to reach significance with the current methodology. Subjects may be less likely to type a non-word response than a word even if their fused percept was closer to a nonword. The third experiment was designed to replicate the findings from the second experiment using a closed-set response paradigm. In this procedure, participants could only select one of six response options (selected from the most common responses indicated in the open set, as well as other possible target fusions) rather than freely typing their response. We hypothesized that

the closed-set response methodology would result in a larger number of fused responses and yield a more stable estimate of AVPF.

Experiment 3

Methods

Stimulus Selection

Stimuli for Experiment 3 were the same as those used in Experiment 2, (word and nonword sets of bilabials stops, Table 3) and were selected to compare fusions of the liquids /l/ and /r/ and words and nonwords.

Lexicality	Visual Stop	Auditory liquid		Fused response	
		/l/	/r/	stop + /l/	stop + /r/
Word	back	lack	rack	black	brack
	pay	lay	ray	play	pray
Nonword	baba	lala	rara	blabla	brabra

Table 3: Stimulus set used in Experiment 3

Participants

Thirty participants took part in this experiment. All participants were undergraduate students at Indiana University and they either received partial course credit or were paid $10.00 for their participation. All participants were native speakers of English and reported no history of speech or hearing impairment and had normal or corrected-to-normal vision.

Procedure

The procedure for Experiment 3 was the same as that for Experiment 2 except for the response method. Whereas Experiments 1 and 2 obtained open-set responses from participants, Experiment 3 required participants to chose from among six response options selected based on the responses that were most common for those conditions in Experiments 1 and 2. Response options included each constituent (i.e., the auditory and visual components) (e.g., “lack” and “back”), the predicted AVPF (e.g., “black”), and the theoretical feature-level fusion that might occur if the auditory and visual components were presented in the other modality (e.g., “dack”). Two additional response options were included. For nonwords, response options included only those with fusions at both locations (e.g., “blabla”). The response options for each stimulus are provided in Table 4.

A fixation cross at the center of the screen preceded each stimulus and was followed by a 500 ms delay. Immediately after the end of the stimulus, six boxes of equal size containing a possible response alternative appeared on the screen. Participants used a mouse to select the response option that was closest to what they thought the talker said. Participants were instructed to both watch and listen during the experiment and that their responses could be either real words or nonwords. Great care was taken not to bias the subjects’ attention to either modality. No time limit was imposed and a I second intertrial interval separated each trial.

Condition	Lexicality	Video/Audio	Fused Responses	Unfused Responses
Experimental	Word	back / lack	black	lack, back, dack, rack, brack
		back / rack	brack	rack, back, dack, lack, black
		pay / lay	blay, play	pay, lay, tay, day
		pay / ray	bray, pray	ray, pay, tay, bay
	Nonword	baba / lala	blabla	lala, baba, dada, rara, brabra
		baba / rara	brabra	rara, baba, dada, lala, blabla
		papa / lala	blabla, plapla	lala, papa, tata, dada
		papa / rara	brabra, prapra	rara, papa, tata, dada
Control	Word	lack / lack	black	lack, back, dack, rack, brack
		rack / rack	brack	rack, back, dack, lack, black
		lay / lay	blay, play	lay, pay, tay, day
		ray /ray	bray, pray	ray, pay, tay, lay
	Nonword	$lala_b$ / $lala_b$	blabla	lala, baba, dada, rara, wawa
		$rara_b$ / $rara_b$	brabra	rara, baba, dada, lala, wawa
		$lala_p$ / $lala_p$	blabla,	lala, papa, tata, rara, wawa
		$rara_p$ / $rara_p$	brabra	rara, papa, tata, lala, wawa

Table 4: Response options for stimuli of interest in Experiment 3.

Data Analysis

As with Experiment 1 and 2, only incongruent stimuli with a bilabial stop (e.g., "back") and a liquid (e.g., "lack") were included in the experimental stimuli (N=8). Congruent stimuli with a liquid in both the visual and auditory domains (e.g., "lack") served as the control stimuli (N=8). These two sets were selected to determine the degree of AVPF. Participant responses were coded as either resulting in a consonant cluster (ignoring voicing alternations). Other responses with no consonant cluster, such as "pay", "lay", or "ray", were coded as no fusion.

Results

A 2x2x2 repeated measures ANOVA using fusion rate as the dependent variable and Lexicality (word, nonword), Liquid Type (/l/, /r/), and Experimental Condition (experimental, control) as within-subjects factors was conducted on the data. The ANOVA revealed main effects of Lexicality ($F(1, 29) = 24.794$, $p \leq 0.001$), Liquid Type ($F(1, 29) = 33.740$, $p \leq 0.001$), and Experimental Condition ($F(1, 29) = 35.263$, $p \leq 0.001$). More fusions were reported for words than for nonwords and for /l/ than for /r/. Additionally, more fusions were reported in the experimental condition than in the control, indicating that listeners perceived AVPF in incongruent stimuli comprised of visual bilabial stops and auditory liquids. Results are presented in Figure 3.

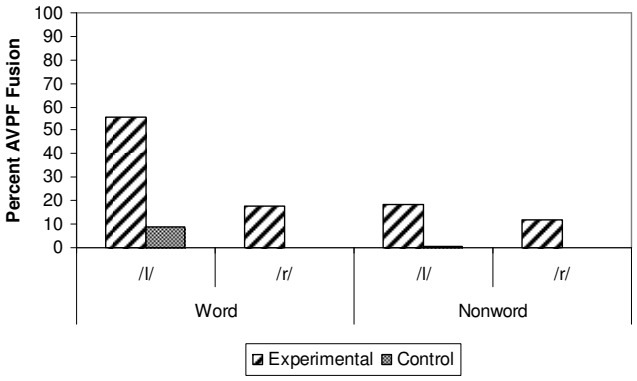


Figure 3: Percent of fused (cluster) responses for visual stops in monosyllabic words and disyllabic nonwords and different auditory liquids in Experiment 3.

In addition to the main effects above, all two-way interactions were significant (all *p-values* ≤ 0.001), as well as the three-way interaction ($p = 0.009$). Post-hoc paired samples t-tests revealed that fusions in the word bilabial + /l/ experimental condition were significantly greater than in its corresponding control condition (55.83% vs. 9.17% respectively) ($p \leq 0.001$). In addition to these expected fusions, AVPF also occurred more in the experimental condition than the control condition for words + /r/ (17.50% vs. 0.00%), for nonwords + /l/ (18.33% vs. 0.83%) and for nonwords + /r/ (11.67% vs. 0.00%) (all $p \leq 0.011$). These findings show that the rate of reported fusions in each individual experimental condition was significantly greater than the corresponding control condition. Thus, AVPF can occur in more conditions than suggested from Experiments 1 and 2. Additional paired-samples t-tests were conducted to examine the rate of AVPF in the four experimental conditions. Results revealed that fusions in the word + /l/ condition were reported significantly more often than in the other conditions ($p < 0.001$) but that fusions in response to the other experimental conditions (e.g., word + /r/, nonword + /l/ or nonword + /r/) did not significantly differ from each other (all *p-values* ≥ 0.147)

Discussion

The results of Experiment 3 replicated and extended the findings obtained in the previous two experiments. The data reported in Experiment 3 revealed that AVPF occurred most frequently in response to real words that contained a visual bilabial and an auditory liquid /l/ as was found in Experiments 1 and 2. However, changing the response format from open-set to closed-set resulted in listeners reporting fusions in other contexts as well; AVPF was also observed in words with the auditory liquid /r/ and in disyllabic nonwords with /l/ and /r/. AVPF was found significantly more frequently for words compared to nonwords and for the liquid /l/ compared to the liquid /r/. These differences are driven by the strength of the effect in the monosyllabic word + /l/ condition.

Stimuli for this experiment were created to emulate the stimuli that were used successfully in previous McGurk effect studies (McGurk & MacDonald 1976) and as such, the nonwords that we used were disyllabic (e.g., “baba”, “lala”). As a result, we cannot determine from out data whether the differences in rate of fusion between the word and nonword conditions were due to the length of the nonwords (disyllable vs. monosyllable) or whether they were due to the lexical frequency of the targets or constituents.

Although the rate of AVPF reported for stimuli with an auditory /r/ reached significance with a closed-set response format, it was still significantly lower than that of stimuli with an auditory /l/. As discussed in Experiment 1, the rate of perceived fusions may be related to the conflict between the visual and auditory stimulus. The most likely context to elicit perceptual fusions is when the conflict between the visual and auditory input is greatest, as with visual bilabials and the auditory liquid /l/. While the conflict between visual bilabials and auditory liquid /r/ is decreased, due to the presence of labial gestures for both segments, it is not eliminated; thus AVPF is found in these conditions as well but to at a lower rate.

General Discussion

In three perceptual experiments, we demonstrated the existence of a novel audiovisual fusion. AVPF is a multimodal fusion in speech perception in which the simultaneous presentation of a viseme and different auditory phoneme results in the perception of a permissible cluster of two phonemes. In addition to revealing the existence of AVPF, the results of our experiments revealed three major findings about the phenomenon. First, the rate at which AVPF is perceived depends critically on the place of articulation of the visual stop consonant, and occurred only with visual bilabial stops that have highly distinctive visual attributes. Second, AVPF occurs more often with the liquid /l/ than with the liquid /r/ when paired with visual bilabial stops. Third, AVPF is affected by the lexical status of the composite stimuli, occurring more frequently with real words than with nonwords.

We believe that an account of this novel multimodal perceptual fusion (AVPF), as well as the McGurk effect and the unimodal, dichotic listening fusions, must appeal to the degree of conflict between the two inputs, whether they are within a single modality or across separate modalities. In Experiment 1, we found that AVPF was limited to visual bilabial stops paired with an auditory /l/. In this case, the auditory input indicated that a non-labial sound had occurred, but the visual display indicated that labial sound had been produced, thus specifying "labial" to the perceiver. Because visual information for labials is the most salient type of visual speech information (Walden, Prosek, Montgomery, Scherr, & Jones, 1977), the degree of conflict between the auditory input ("not labial") and the visual display is high. Perceivers resolve the conflict between two salient perceptual cues by combining both in their responses (e.g., "black"). By similar explanation, AVPF does not occur with visual alveolars and velars paired with /l/. Because both the auditory and visual signals suggest "non-labial" targets, the perceptual system has nothing to resolve and the final percept does not contain a consonant cluster. Generally in this case, subjects' responses are of the auditory signal.

Not only does the notion of conflict account for the differences observed for place of articulation, it can also account for the differences observed between the two liquids. AVPF was found to be less robust for visual labial stops paired with auditory /r/ than auditory /l/. Because /r/ in English is produced with lip rounding, the associated visual gesture is a concomitant labial articulation. Although the labial gestures for /r/ and bilabial stops are not identical, the conflict between the two inputs is reduced, thus decreasing the rate of AVPF. Indeed, in Experiment 3 where AVPF occurred in more contexts than in the previous studies, it still occurred less often with /r/ than with /l/.

Visual alveolar and velar stops paired with /r/ also did not result in AVPF, although the potential consonant clusters (e.g., /dr/, /gr/) are legal sequences in English. In these particular pairings, the auditory input is associated with a labial gesture from /r/, but the visual input is clearly non-labial. In these cases where the visual input is strongly negative for labiality, perceivers frequently report only the auditory signal, but crucially do not report a sequence of two segments.

This final example where the auditory information implies "labial" and the visual information implies "non-labial" resembles the stimulus configuration that results in the McGurk illusion. Similar to AVPF, the McGurk Effect and Combination Responses (e.g., "bagba", see Introduction) reported by McGurk and MacDonald can be explained by the degree of conflict between the auditory and visual information. The McGurk Effect is observed when the visual cue is "not labial". On these types of trials, listeners report a single segment which conforms to the visual input ("not labial") and actually alters the auditory input of /b/ to yield the most anterior non-labial stop, namely /d/. In contrast, AVPF, and the combination responses, occur in the opposite configuration of inputs where the visual cue is "labial". In these cases, the salient visual labial articulation strongly specifies the presence of a bilabial stop even though one is not present in the auditory stream of information. The auditory information is not altered, but the visual information is added to the auditory stream because nothing in the visual domain contradicts the stronger auditory percept as is the case in the McGurk effect. Thus, for both AVPF and the combination response, the only suitable resolution to the mismatch between the auditory and visual inputs is an output response that contains both consonants. In the McGurk effect the visual information "excludes" certain phonemes, whereas in AVPF the visual information strongly indicates the presence of certain phonemes.

The explanation of fusions as resulting from conflict between two sources of phonetic information also accounts for some results of unimodal fusions. Because dichotic listening experiments are conducted within a single perceptual domain (i.e., audition), there is no inherent inequality between the two inputs. Both AVPF and the McGurk effect result from combining inherently unequal types of input. Although visual information can enhance auditory speech perception (Sumby & Pollack, 1954), the auditory stream is the dominant perceptual channel in normal-hearing listeners. It is no surprise that both of these effects show that audiovisual conflict depends on the one cue in which the visual domain may be superior (Summerfield, 1987), namely labial vs. nonlabial.

In contrast, unimodal fusions observed in dichotic listening studies have no inherent dominance of information. Thus, the unimodal analogue to AVPF extends from bilabial stops paired with /l/ to alveolar and velar stops and to /r/, yielding a larger set of clusters (e.g., /gl/, /gr/, /dr/, /br/). In these cases of auditory Phonological Fusion, a conflict exists between the manner of articulation between the two inputs, creating a perception that includes both segments. In the unimodal analogue to the McGurk effect (i.e., Psychoacoustic Fusion), the manner of articulation is the same; all that conflicts is the place of articulation. In these cases, the percept is of the same manner (i.e., a stop) and the perceptual system resolves the conflict by perceiving a stop at an intermediate place of articulation. In addition to stop-stop and stop-liquid conflicts, Cutting (1975) conducted dichotic listening studies pairing /s/ with stops. He found that these combinations also resulted in Phonological Fusions (Cutting, 1975). If perceptual conflict is at work in resolving the previous AVPFs, we would expect that auditory /s/ paired with a visual bilabial stop would also result in the perception of /sp/ clusters.

The other main finding of the current studies was the presence of an interaction between rate of AVPF and lexicality. Experiments 2 and 3 revealed that monosyllabic words fused more readily than disyllabic nonwords. In Experiment 2, the more stringent measure of fusions requiring AVPF in both positions of the two syllable nonwords did not reveal a statistically significant increase in fusions over the control condition. In contrast, the more lenient measure, which counted all responses containing at least one AVPF, did result in a significant number of fusions. Even with this latter measure, however, the rate of AVPF in nonwords was significantly lower than in words. The different findings observed for words vs. nonwords may result from their lexical status, although other differences between these two types of stimuli exist.

The most salient difference between the words and nonwords used in this study was length; the real words were monosyllabic and the nonwords were disyllabic. Because longer words or nonwords have more segments that must be aligned, it is more likely that the timing or synchrony between some segments may not coincide precisely. Although disyllabic nonwords are more likely to be asynchronous, this may not be problematic for perceiving Phonological Fusions. Previous studies have reported a large window of asynchrony over which audiovisual speech stimuli are judged as synchronous (Conrey & Pisoni, 2006; van Wassenhove *et al.,* 2006). If the auditory and visual portions of the stimulus come from the same utterance but are presented with one modality temporally ahead of the other, participants still perceive the stimuli as synchronous: at least 131 ms of asynchrony in monosyllabic words (Conrey & Pisoni, 2006) and at least 74 ms of asynchrony for /da/ (van Wassenhove *et al.,* 2006). Furthermore, the McGurk Effect can be perceived over a similar window of audiovisual asynchrony (30 ms auditory lead to 175 ms visual lead) over which identical stimuli are judged as synchronous (van Wassenhove *et al.,* 2006). Finally, dichotic listening Phonological Fusion is also observed over a large window of asynchrony, up to 150 ms delay between the two auditory presentations (Cutting, 1976). Thus, we conclude that the difference in the rate of AVPF for words and nonwords is not due to differences in overall synchrony of the stimuli, but is due to lexical status.

The difference in length between words and nonwords in this study also yields a difference in the location of the target fusions. Dichotic listening Phonological Fusion occurs not only for consonants in word-initial position, but has also been observed for intervocalic and word-final consonants. In the current study, we demonstrated that AVPF can occur word initially and we have provided some evidence that it can occur intervocalically. Further experiments are needed to fully explore the occurrence of AVPF intervocalically and to test the existence of AVPF word finally. The findings obtained in Experiment 2 suggest that initial position is the most favorable location to perceive AVPF, where it was perceived with /l/ in 10.58% of responses; intervocalic AVPF was only perceived with /l/ in 4.81% of responses.

In addition to lexicality, word frequency could also affect the rate of perceived fusions. In the monosyllabic word conditions, all the constituents were words, while the target fusions were mostly words (e.g., “black”), but included a few nonwords (e.g., “blay”). In the disyllabic nonword condition, the stimuli were nonwords, but were most likely utterances familiar to the participants (e.g., “baba”, “papa”, “lala”, “rara”). In spoken word recognition, high frequency words are perceived more easily and recognition errors tend to be of higher frequency than the frequency of the stimulus (Broadbent, 1967). We would expect the word frequency of the target fusion relative to the frequency of the constituents to have an effect on fusion rate. To our knowledge, no experiments have reported that explore the effect of word frequency on the perception of either multimodal or unimodal fusions. This would be a productive topic of investigation and could potentially provide additional information on whether fusions depend upon the prior linguistic experience of the observer.

Prior linguistic experience has also been demonstrated to play a role in the susceptibility to the McGurk effect. In Finnish, for example, the McGurk effect occurs at a rate similar to English (Sams, Manninen, Surakka, Helin, & Katto, 1998). However, in Japanese, perceptual fusions occur much less frequently (Sekiyama & Tohkura, 1991). A much lower signal-to-noise ratio is required in the auditory stream for native Japanese speakers to exhibit the McGurk effect. In other words, the relative dominance of auditory information must be severely degraded for Japanese listeners to incorporate visual information and perceive a multimodal fusion. Additionally, native speakers of Japanese show no evidence of a combination response when presented with a visual bilabial /b/ simultaneously with an auditory liquid /r/ (Sekiyama & Tohkura, 1991). Since stop-liquid clusters are prohibited in Japanese, the

combination response of /br/ is not a valid percept. Similarly, we would expect that AVPF would be nonexistent in Japanese due to the phonotactics of the language.

The data presented here and in other AV experiments illustrate several parallels between multimodal and unimodal speech perception. Studies of sound localization show that delayed auditory input (unimodal) or a displaced visual input (multimodal, ventriloquist effect) both result in a change in the perception of the location of a sound source. Similarly, fusing the two inputs to yield a single intermediate segment occurs in both unimodal (Psychoacoustic Fusion) and multimodal (McGurk effect) perception. The experiments presented here document another similarity between unimodal and multimodal fusions; the sequential perception of stop-liquid clusters occurs both in unimodal, dichotic listening and in this newly documented AVPF.

Despite these similarities, there is a critical difference that is related to the relative importance or weighting of the two inputs. In dichotic listening, both inputs are auditory and thus equal in terms of the type and amount of information that is conveyed. In AV perception, there is an inherent asymmetry in the relative dominance of the two cues; auditory information carries more robust phonetic information than visual (viz. near ceiling performance of auditory-only speech perception and comparatively low performance – 15.69% of key words in CUNY sentences – for visual-only speech perception; Conrey & Pisoni, 2006).

More specifically, the kind of information that can be contrasted in the visual domain differs, possibly being limited to labial vs. nonlabial articulations. Whether information from the two inputs is equal in importance (unimodal vs. multimodal), this asymmetry between unimodal and multimodal fusions corresponds to differences in the responses. For example, the perception of fused clusters containing alveolars and velars paired with liquids occurs only in unimodal fusions – and not multimodal fusions – where cues to stop place of articulation are retained. Similarly, the combination response versus the McGurk effect depends on which information is presented in each modality. Unimodal, dichotic presentation of two stops does not result in a combination response because the information presented to both ears carries the same perceptual salience (Cutting, 1976).

In summary, we reported three experiments that documented the existence of a novel audiovisual fusion, audiovisual phonological fusion (AVPF). When visual bilabial stops (i.e. /b/ or /p/) and auditory liquids (i.e. /l/ or /r/) are presented simultaneously, observers often perceive a consonant cluster composed of both the stop and liquid (e.g., /bl/ or /br/). Furthermore, we have shown that AVPF depends on both the place of articulation of the visual stop and the identity of the liquid. We argued that the presence of AVPF is directly related to the degree of conflict between the auditory and visual sources of information. Our account of perceptual conflict also explained the earlier results found in other multimodal (e.g., McGurk Effect) and unimodal fusions. Because of the relative importance of the two input streams, what counts as 'conflict' is different for unimodal and multimodal fusions. Here we presented evidence suggesting that fusions occur for a basic reason – conflict – which is resolved by mechanisms that are similar for both unimodal and multimodal perception.

References

Bertelson, P., Vroomen J., Gelder B.D., & Driver J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception and Psychophysics, 62(2),* 321-332.

Broadbent, D.E. (1967). Word-Frequency Effect and Response Bias. *Psychological Review, 74(1),* 1-15.

Conrey, B. & Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America, 119(6),* 4065-4073.

Cutting, J. E. (1975). Aspects of Phonological Fusion. *Journal of Experimental Psychology, 104(2),* 105-120.

Cutting, J. E. (1976). Auditory and Linguistic Processes in Speech Perception: Inferences from Six Fusions in Dichotic Listening. *Psychological Review, 83(2),* 114-140.

Cutting, J. E. & Day, R.S. (1975). The perception of stop-liquid clusters in phonological fusion. *Journal of Phonetics, 3,* 99-113.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Sams, M., Manninen, P., Surakka, V., Helin, P., & Katto, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication, 26,* 75-87.

Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90,* 1797-1805.

Sumby, W. H. & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26(2),* 212-215.

Summerfield, Q. (1987). Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading.* (pp. 241-289). Hillsdale, NJ: Lawrence Earlbaum & Associates.

van Wassenhove, V., Grant K.W., & Poeppel, D. (2006). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45(3),* 598-607.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130-145.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Power Law Degree Distributions Can Fit Averages of Non-Power Law Distributions[1]

Thomas M. Gruenenfelder and Shane T. Mueller[2]

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This research was supported by NIH Grant DC00111.
[2] Now a Senior Scientist at Klein Associates, a Division of ARA Inc., Fairborn, OH.

Power Law Degree Distributions Can Fit Averages of Non-Power Law Distributions

Abstract. Many complex systems have been modeled as networks. Examples of such systems that are of interest to cognitive psychologists include the mental lexicon for spoken word recognition and semantic memory. A frequent finding in such studies is that the frequency distribution of the number of connections for each node in the network follows a power law. This finding has been interpreted to mean that the network grows through a process similar to preferential attachment: when a node is added to the network, it attaches to other nodes with a probability proportional to the number of connections those other nodes already have. Power-law degree distributions, however, may also well describe degree distributions that result when averaging across multiple individual degree distributions, none of which follows a power-law. Further, each of these individual distributions may reflect a random growth process rather than the more systematic process suggested by preferential attachment.

Introduction

There has been a lot of interest over the past 10 years in using the tools of graph theory to model complex networks (see Albert & Barabási, 2002, for a review). In such studies, entities are represented as nodes or vertices, and an edge or link is placed between two nodes if some pre-determined relation exists between them. For example, in a model of the World Wide Web, each site on the web can be represented by a node. A link is placed between two nodes if the site represented by one such node has a (web) link to the site represented by the other node (see, for example, Pastor-Satorras & Vespignani, 2004).

A common finding that has emerged from these studies is that many of these real-world networks have a small-world (Watts & Strogatz, 1998), "scale-free" (Barabási & Albert, 1999) structure. In a small-world network, the mean shortest path between any two nodes in the network is relatively short. The shortest path refers to the smallest number of links that must be traversed to get from one node to the other. In particular, in a small-world structure, the mean shortest path length grows more slowly than the number of nodes. It is similar to the mean shortest path length that occurs in a random graph, where links are placed between nodes randomly. Lattice networks, in contrast, where each node connects to only a very few near neighbors, have a much longer mean shortest path length. Small-world networks are also characterized by a mean clustering coefficient (CC) larger than that expected by chance. A node's CC is the probability that two nodes with a link to that node also have a link to one another. When we say that the CC is higher than the CC expected by chance, we mean that it is higher than what would be expected in a comparably sized network (a network with the same number of nodes and the same total number of links) in which links between nodes have been placed at random. In brief, a small-world network is a network with a relatively small mean shortest path length between any two nodes and a CC that is much higher than that expected in a random graph.

In a scale-free network, the degree distribution follows a power law, $N(k) \sim k^{-\gamma}$, where N(k) is the degree distribution, k is the degree, and γ, the distribution's parameter, is typically between 2 and 3. A node's degree, k, refers to the number of links going into or coming out of that node. The degree distribution is simply the frequency distribution of the degree across all nodes in the network. When plotted on log-log coordinates, the slope of a power-law degree distribution, of course, is a straight line with a slope of $-\gamma$. (For this reason, we sometimes refer to γ as the slope parameter.)

A finding as ubiquitous as that of scale-free degree distributions demands an explanation and Barabási and Albert (1999) have provided one. They showed that if a network grows continuously by adding new nodes, and the new nodes form links with existing nodes with a probability proportional to the number of links that existing node already has, then a power-law degree distribution emerges. They termed this process "preferential attachment." Less formally, it has been referred to as the "rich get richer" principle. In the psychological literature, similar phenomena have been referred to as the Matthew principle (Stanovich, 1986; see also McClelland & Rumelhart, 1981).

Many of these recent studies of complex networks are of direct interest to cognitive psychologists. Schweickert (in press a, b), for example, analyzed the co-occurrence of characters in a given person's dream. Characters were represented as nodes and a link was placed between two nodes if those two characters had appeared in the same dream. Schweickert found that the networks for all three of the dreamers he analyzed showed a small-world structure. Two of the three also showed a power law degree distribution, at least at higher degrees (Schweickert did not use the term scale-free.). Given the similarity of these structures to social networks, he argued that analysis of the characters appearing in a person's dream might be a reliable method for determining that person's social network.

In another recent study, Steyvers and Tenenbaum (2005) analyzed three types of semantic networks, in which the nodes represented words. The first was based on the Nelson, McEvoy, and Scheiber (1999) word association norms. A link was placed between two words if one of those words had been produced as an associate of the other by at least two participants in the Nelson et al. norms. The second network was based on Roget's thesaurus (Roget, 1911). In this network, a link was placed between two words if they shared at least one category in the thesaurus in common. The third network was based on WordNet (Fellbaum, 1998; Miller, 1995) in which word-form nodes are connected to word-meaning nodes and word-meaning nodes, in turn, are connected to one another based upon the relation between those meanings (such as antonymy (LOVE and HATE), hypernymy (A ROBIN is a BIRD), and meronymy (A ROBIN has WINGS)).

All three networks showed a small-world, scale-free structure. Steyvers and Tenenbaum (2005) noted, however, that other aspects of their results were inconsistent with growth through preferential attachment and proposed a somewhat different model to explain their results. In the Steyvers and Tenenbaum model, at each point in time, a node is chosen with a probability proportional to its number of links for differentiation. If node $\underline{i}$ is chosen for differentiation, then a new node is added to the network and connected to $\underline{M}$ randomly chosen nodes that are already neighbors (i.e., have a link to) node $\underline{i}$. This model is not so much a "rich get richer" model as a "rich beget rich" model.

Ferrer and Solé (2001) built a network of words based upon their co-occurrence in sentences. In particular, they placed a link between any two words that occurred within two words of each other in a sentence in their corpus with a probability higher than that expected by chance. Their network showed a small-world structure. Its degree distribution showed two distinct power-law regions, one covering smaller degrees and a second, with a somewhat steeper slope, covering higher degrees. They interpreted this property as being consistent with the notion that the network of words in sentences is scale-free.

Soares, Corso, and Lucena (2001) modeled the syllabic structure of the Portuguese language. Each node in their network represented a word in the Portuguese language. A link was placed between two nodes if those two syllables co-occurred in the same word. Soares et al.

found that their network followed a small-world, scale-free structure. They interpreted this finding to mean that the Portuguese language evolved in a non-random manner, i.e., new words were added to the language through a process akin to preferential attachment.

Finally, Vitevitch (in press) and Gruenenfelder and Pisoni (2005) used the Hoosier Mental Lexicon data base (Nusbaum, Pisoni, & Davis, 1984) to model the mental lexicon of spoken word representations. In both of these studies, spoken word forms were represented as nodes and a link was placed between two nodes if the word represented by one could be changed into the word represented by the second by the deletion, addition, or substitution of a single phoneme (Greenberg & Jenkins, 1964; Landauer & Streeter, 1973; Luce & Pisoni, 1998). (We refer to this rule for defining lexical neighbors as the DAS rule, for Deletion, Addition, or Substitution.) Both studies found that the mental lexicon, modeled in this way, showed a small-world structure. In addition, both studies found that a power law did a good job describing the degree distribution. Vitevitch, for the subset of words that he studied, found that an exponential fit the degree distribution somewhat better than did a power function. Gruenenfelder and Pisoni found that which function provided a better fit was determined by details of how the fit was made (see below).

Gruenenfelder and Pisoni (2005) pointed out that a word's degree is confounded with its length, as measured by the number of phonemes. Shorter words have higher degrees, i.e., have more neighbors, as determined by the DAS rule, than longer words. Further, given that words are constructed from a relatively small set of basic elements or particles, i.e., phonemes, this fact has to be the case. When degree distributions were examined for particular classes of words, as determined by their length, no evidence for power-law degree distributions was evident. The power-law degree distribution emerged only when data were averaged across these multiple, non-power-law distributions. Figure 1 shows the degree distribution, collapsed across word length, reported by Gruenenfelder and Pisoni. The figure is on a log-log plot and hence a power function would be indicated by a straight line. Visual inspection of the figure does in fact suggest that the distribution is well fit by a straight line (with the exception of the last four data points, where the function "falls off the cliff"). A regression analysis showed that in fact the distribution is reasonably well fit by a straight line, $R^2 = .79$. (Incidentally, the fit of an exponential to the overall degree distribution is even better: $R^2 = .89$. If the last four data points to the right, where the function falls off the cliff, and which are inconsistent with both a power law and exponential function, are dropped from the regression analysis, the R^2 for the power law fit increases to .96; the R^2 for the exponential fit increases to only .91. Hence, the better fit of the exponential to the overall distribution seems due to its ability to better describe what are clearly contradictory data points.)

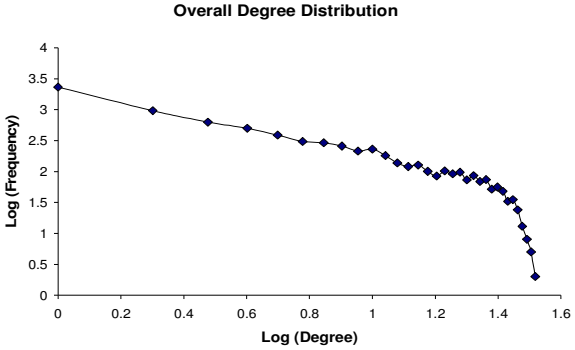


Figure 1. Overall degree distribution of neighbors, collapsed across word length, for the word corpus analyzed by Gruenenfelder and Pisoni (2005).

Figure 2 shows the degree distribution broken down by word length. None of the degree distributions in Figure 2 provides a close match to a (downward trending) power function, particularly at intermediate word lengths (3, 4, 5, and 6), which account for the majority of words. The composite power law degree distribution emerges only when the data are averaged across word lengths.

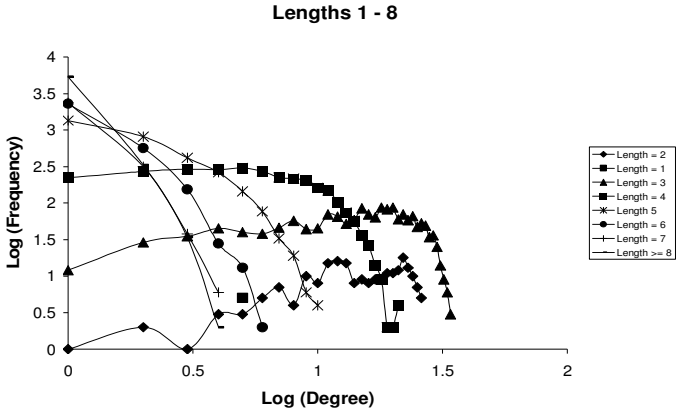


Figure 2. Degree distribution of neighbors, broken down by word length in phonemes, for the word corpus analyzed by Gruenenfelder and Pisoni (2005).

Gruenenfelder and Pisoni's (2005) interpretation is one example of a more general phenomenon that has been observed elsewhere, such as in the memory (Anderson, 2001; see also Brown & Heathcote, 2003a, 2003b) and skill-learning literatures (Newell & Rosenbloom, 1981). A power-law function can frequently well describe the result when data are averaged across a number of underlying distributions, none of which itself is a power law. In fact, the power law fit is sometimes better than a fit of a function of the same form as that of those being averaged to give the composite function.

Anderson's (2001; Anderson & Tweney, 1997)) work is perhaps the most pertinent here. Anderson & Tweney were interested in modeling the memory decay curve—the decline in memory performance as a function of retention interval or time, t. They assumed that memory for an item consisted of multiple memory traces, each decaying independently as an exponential, Ae^{-Bt} (A and B are constants), each with a different decay parameter, B. (Because, when the logarithm of frequency is plotted as a function degree, in our case, or time, in Anderson & Tweney's case, an exponential yields a linear function with a slope of –B, we sometimes refer to B as the slope parameter.) Observed memory performance is the average of these curves. This situation would apply, for example, to the case where a single retention curve is plotted based on average retention across subjects, or where a single retention curve is plotted based on average retention of a single subject for multiple different items. Anderson and Tweney found that in a small number of their simulations (5%) (arithmetic) average performance was better fit by a power law function than by an exponential. In general, power law fits improved as rate parameters varied over a larger range. In addition, when they made the additional assumption that there was also noise in the measurement of each subject's memory performance at each retention interval, power laws fit the resulting retention curves better than exponentials in 97% of the cases. In summary, Anderson and Tweney showed that power law distributions can arise when multiple exponential distributions are averaged together.

Anderson (2001) extended Anderson and Tweney's (1997) work in two important ways. First, he showed that when the component exponential curves were constrained to have a downward slope, even when noise in the measurement of the subject's performance was not assumed, power laws tended to fit the average curve better than did exponentials provided there was sufficient variance in the rate parameters of the component distributions. Second, he showed that the same result—better fit of power laws to the average distribution than of a function of the same form as the component distributions—occurred when the component distributions were range-limited linear or range-limited logarithmic functions. By range-limited, Anderson meant that the function was never permitted to go below 0, certainly a reasonable assumption when discussing memory retention. We can remember nothing but we cannot remember a negative amount of information. Without this restriction, the average of a number of linear components is, of course, another linear function and the average of a number of logarithmic components is another logarithmic function.

Anderson's work with the exponential is especially interesting to us when applying graph theoretic tools to the analysis of complex systems for two reasons. First, Watts and Strogatz (1998) showed that for random graphs (i.e., graphs with a fixed number of nodes in which a fixed number of links is placed between randomly chosen pairs of nodes) the degree distribution follows a Poisson distribution, the right hand tail of which approaches an exponential. It is precisely at higher degrees that the power law degree distribution often becomes most evident. Schweickert (in press a, b), for example, only considered degrees above the median degree. Second, Barabási and Albert (1999) showed that if the assumption that networks grow over time is retained, but the assumption of preferential attachment is replaced with the assumption that new nodes attach randomly to existing nodes (i.e., attach to each existing node with a probability proportional to the number of existing nodes), then the degree distribution is not a power law but an exponential.

Suppose that we have a complex network that is in fact composed of several sub-networks in the following sense. The network is growing over time, but there are several different processes that can result in a new node being added to the network. The existing nodes that a new node connects to are random (i.e., a new node's probability of connecting to an existing node is proportional to the number of nodes in the network), but the constant of proportionality differs from one process to another. In this case, the observed network is the average of the several non-observed "sub-networks," and its degree distribution is the average of the exponential degree distributions of the several underlying "sub-networks." Anderson's work indicates that this aggregate degree distribution is likely to follow a power law even though all the underlying distributions are exponential. There is certainly no guarantee that the observed distribution in the composite network is of the same form as the distribution in each of the underlying sub-networks (cf. Estes, 1956).

To summarize, a power law degree distribution may result because a network grows via preferential attachment. It may also result because the network reflects several component networks, each of which grows randomly and each of which has an exponential (or perhaps even some other) degree distribution.

Are the recent studies we cited above vulnerable to this ambiguous interpretation? Could the power law degree distributions that they observe simply be the result of averaging across "random" processes rather than reflecting some fundamental property of the underlying network, such as growth through preferential attachment? The possibility certainly cannot be ruled out. The power law degree distribution for the spoken word lexicon observed by Gruenenfelder and Pisoni (2005) (and possibly the degree distribution observed by Vitevitch, in press) does seem to be the

result of averaging across words of different lengths. Steyvers and Tenebaum's (2005) network based on association norms averages across, amongst other things, subjects. Their thesaurus based network averages across words of different syntactic classes (verbs, nouns, adjectives). Their WordNet based network averages across different semantic relations.[3] Ferrer and Solé's (2001) network involved averaging over different types of documents. Soares et al. (2001) averaged across words of different lengths, as measured by number of syllables. Schweickert's (in press a; in press b) work seems less affected by this potential ambiguity, as he built separate networks for each of his dreamers. However, he did not report goodness of fit measures to exponential degree distributions.

The present study was carried out to investigate whether the occurrence of power law degree distributions could be the result, in at least some graph theoretic studies, of averaging across multiple exponential distributions, each generated by a random process, rather than the result of a more systematic growth process, such as preferential attachment. We built a simulation that created networks by adding a node to that network at each time step, t. That new node was then connected to each existing node in the network with probability of r/N_t, where N_t is the total number of nodes in the network at time step t, and r, referred to above as the constant of proportionality, is a parameter that we varied across simulations. The simulation continued until some pre-specified maximum number of nodes, N, had been created. We then computed the degree distribution for that network. Based on the work of Barabási and Albert (1999) we expected (and found) that each individual network created in that fashion would yield an exponential degree distribution.

We then "simulated" a composite network by summing the degree distribution of multiples of these individual networks. We then fit the degree distribution of this composite network to linear, exponential, and power functions to determine which function best described the degree distribution of the composite network.

Simulation I

In their simulations, Anderson (2001) and Anderson and Tweney (1997) directly manipulated the slope parameter of the underlying exponential functions that they were simulating. They found that for a power law to best fit the composite function, there needed to be sufficient variability in the slope parameters of the underlying exponentials. The present study does not directly simulate an exponential degree distribution. Instead, it simulates a process that, as it turns out, produces an exponential degree distribution. The slope parameter of that exponential is affected by both N, the size of the network in number of nodes, and r, the probability that a new node will attach to any already existing node in that network. Simulation I investigated more precisely how r affected the slope parameter for individual exponential degree distributions for different sized networks. The purpose was to allow, in the main set of simulations, choosing values of r such that there would be sufficient variation in the slope parameters for good power law fits to the composite degree distributions to emerge.

As already mentioned, the present simulation built a network by adding a node at each time step. That node was then connected to each already existing node in the network with probability r/N_t, where N_t is the number of nodes in the network at time step t. The simulation

[3] To be fair, Steyvers and Tenenbaum (2005) noted discrepancies between their model and the preferential attachment model. They proposed a revised model, sketched above, to account for their results. It is not clear that a model averaging random networks could capture these additional aspects of their data.

continued until N nodes had been added to the network. Both r and N were user supplied parameters to the simulation.[4]

Method

Four different size networks were simulated in Simulation I. The four network sizes were N = 20,000, 40,000, 80,000 and 100,000 total nodes. For each network size, r was varied from .01 to .1 in steps of .01, from .1 to 1.0 in steps of .05, and from 1 to 2 in steps of .1 (a total of 38 r values for each network size). Each combination of N and r was simulated once, for a total of 4 x 38 or 152 simulations. The degree distribution was determined for each simulation and the best fitting exponential to that degree distribution was computed using regression analysis.

Results and Discussion

Recall that for an exponential degree distribution, the logarithm of a degree's frequency is a linear function of the degree. Consequently, linear regression analysis can be used to compute the best fitting exponential to an observed degree distribution and for estimating the slope parameter of that exponential. That method was used in analyzing the present results. It does preclude including hermit nodes—nodes with no neighbors—in the analyses since the logarithm of 0 is undefined. Excluding these nodes has at worst a small effect on the results reported here. It is also the case that the literature on the analyses of degree distributions does not seem to point to a uniformly accepted way of handling these hermit nodes.

The best fitting exponential was determined for each of the 152 degree distributions computed in Simulation I. R^2 values for these fits are shown in Figure 3 as a function of r and N. Overall, exponential functions fit these distributions extremely well. The mean R^2 values for network sizes of 20,000, 40,000, 80,000 and 100,000 nodes were, respectively, .962, .946, .899, and .868. Because it is already known that networks grown in this random fashion produce exponential degree distributions (see, for example, Barabási & Albert, 1997), these results are not new and were not unexpected. They do provide some confirmation that the simulation was in fact simulating the intended process.

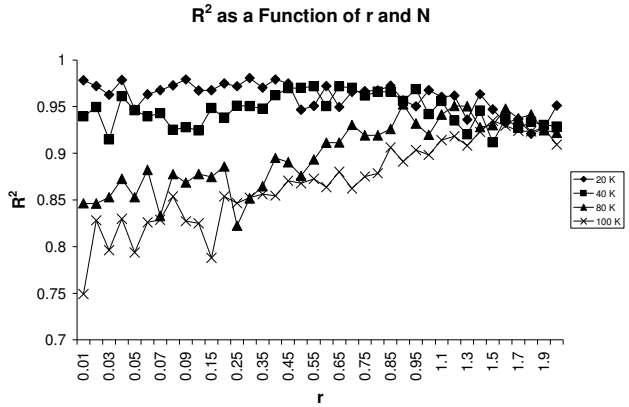


Figure 3. R^2 as a function of r. The curve parameter is N, network size.

[4] There were two additional parameters input to the simulation as well. However, as these were not varied as part of the present study, they are not discussed here.

The fact that the fits became less good as network size grew is perhaps at least in part attributable to the fact that the larger networks had a larger number of degrees. Hence, more data points were being fit for the degree distributions from the larger networks.

The more important results concern how the slope of these best fitting exponential degree distributions changes as a function of r and N. These results are shown in Figure 4. For small values of r, smaller networks show smaller magnitude slopes in the degree distribution than do larger networks. The smaller networks also show more change in the slope as r varies. Above values of r ≈ 1.0, network size does not appear to have much influence on the slope of the degree distribution. Similarly, the effect of r itself on the slope appears to be much larger at values of r less than approximately 1.0. Above that value, the slope changes much more gradually with r. Accordingly, in our main simulation, we worked with relatively small networks (N = 20,000) and with mean values of r under 1.0.

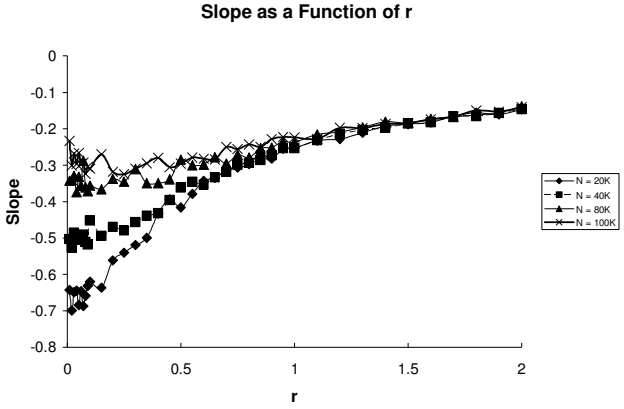


Figure 4. Slope of the predicted degree distributions as a function of r and N.

Simulation II

Simulation II comprised the main simulation of the present study. Its purpose was to determine how well the degree distribution of a composite network, composed of a number of individual random networks, each of which would be expected to have an exponential degree distribution, could be fit with a power law function. Based on the results of Simulation I, all individual networks created in Simulation II were relatively small, consisting of 20,000 nodes. A number of composite networks were built as part of Simulation II. These networks differed from one another in the mean value of the r parameter used to create the individual networks and its standard deviation.

Method

A total of 70 composite networks were simulated by factorially combining 7 mean values of the r parameter of the underlying network (.15 through and including .75 in steps of .1) with 10 values of the standard deviation of r (.1 through and including 1.0 in steps of .1). Each composite network was formed by summing together 20 individual networks, each of which consisted of 20,000 nodes. Each of the individual networks was formed by randomly sampling a value of r from a normal distribution with a mean and standard deviation corresponding to the mean and standard deviation of the corresponding cell in the design. The sampling was done with the restriction that r could not take on negative values. If a negative value of r was sampled it was discarded and the distribution re-sampled. The result of this sampling procedure is that the actual

mean values of r, especially for cells with low means and high standard deviations, was somewhat larger than the nominal means, sometimes greatly so. For example, the actual mean value of r for the individual networks sampled from a distribution with a mean of .15 and standard deviation of .8 was .48.

Results

The degree distribution of each of the 70 composite networks was calculated. The linear, exponential, and power law distributions that best fit these observed distributions were then determined using linear regression analyses. (The linear degree distribution is linear when frequency is plotted as a function of degree. The exponential degree distribution is linear when the logarithm of frequency is plotted as a function of degree. And the power law degree distribution is linear when the logarithm of frequency is plotted as a function of the logarithm of the degree.) Because the logarithm of 0 is undefined, nodes with degree 0 (i.e., with no edges) could not be included when fitting the power law degree distribution. These nodes were also excluded when fitting the linear and exponential distributions so that all three functions would be fitting the same data points. R^2 was used as the measure of goodness of fit.

Tables 1, 2, and 3 show the R^2 values for the best fitting linear, exponential, and power law degree distributions, respectively, as a function of the mean value of r and its standard deviation. As expected, linear degree distributions do not provide a very good fit to the observed degree distributions. Across the 70 composite networks, the mean R^2 for the best fitting linear degree distribution was .446. Exponential distributions, in contrast, provide excellent fits to the observed degree distributions. Across the 70 composite networks, the mean R^2 for the best fitting exponential degree distribution was .989. The goodness of fit of the exponentials did not vary much as r or its standard deviation varied, at least in part because of ceiling effects. Power law degree distributions produced what might be best called good fits to the observed degree distributions. Across the 70 composite networks, the mean R^2 for the best fitting power law degree distribution was .873. The goodness of fit of the power law did not vary much as the standard deviation of r varied. There is, however, a trend to somewhat worse fits as the mean value of r increased. Although the data in Table 3 indicate reasonable power law fits to the composite degree distributions, a comparison of Tables 2 and 3 clearly show that the exponential fits are superior to the power law fits. In none of the 70 composite networks did a power law better fit the degree distribution than did the exponential.

	Standard Deviation									
Mean	**.1**	**.2**	**.3**	**.4**	**.5**	**.6**	**.7**	**8**	**.9**	**1.0**
.15	.517	.436	.410	.434	.456	.458	.428	.354	.323	.454
.25	.555	.436	.493	.414	.453	.438	.419	.397	.442	.329
.35	.552	.490	.484	.430	.482	.491	.456	.325	.365	.376
.45	.556	.476	.455	.427	.441	.414	.409	.439	.423	.470
.55	.508	.524	.433	.419	.462	.475	.378	.428	.409	.342
.65	.555	.511	.516	.480	.424	.440	.441	.434	.481	.371
.75	.517	.451	.468	.530	.438	.454	.429	.488	.438	.384

Table 1. Goodness of fit (R^2) measures for linear fits to the observed degree distributions for the 70 composite networks, as a function of r and its standard deviation.

	Standard Deviation									
Mean	**.1**	**.2**	**.3**	**.4**	**.5**	**.6**	**.7**	**8**	**.9**	**1.0**
.15	.996	.992	.972	.988	.995	.995	.997	.987	.983	.991
.25	.996	.992	.998	.993	.994	.994	.996	.977	.997	.973
.35	.995	.999	.990	.996	.992	.995	.997	.961	.991	.976
.45	.997	.986	.994	.969	.991	.987	.994	.988	.979	.993
.55	.994	.995	.993	.976	.992	.993	.957	.988	.989	.984
.65	.991	.997	.994	.992	.980	.984	.988	.988	.992	.979
.75	.992	.987	.993	.996	.988	.984	.991	.995	.988	.976

Table 2. Goodness of fit (R^2) measures for exponential fits to the observed degree distributions for the 70 composite networks, as a function of r and its standard deviation.

	Standard Deviation									
Mean	**.1**	**.2**	**.3**	**.4**	**.5**	**.6**	**.7**	**8**	**.9**	**1.0**
.15	.901	.913	.898	.902	.877	.869	.880	.897	.897	.855
.25	.874	.897	.883	.886	.894	.866	.868	.892	.871	.902
.35	.865	.873	.861	.891	.870	.854	.865	.906	.882	.875
.45	.863	.870	.879	.866	.871	.884	.884	.859	.878	.852
.55	.875	.853	.894	.881	.872	.849	.895	.855	.880	.881
.65	.831	.860	.859	.858	.880	.864	.873	.857	.848	.856
.75	.850	.873	.871	.868	.863	.855	.869	.857	.860	.882

Table 3. Goodness of fit (R^2) measures for power law fits to the observed degree distributions for the 70 composite networks, as a function of r and its standard deviation.

Discussion

The present study grew networks by simulating a process known to produce exponential degree distributions. It then formed composite networks by adding together several of these individual networks and examined the degree distributions of those composite networks. Based on the earlier work of others (e.g., Anderson, 2001; Anderson & Tweney, 1997), we hypothesized that the degree distributions of those composite networks might be best fit by a power law function. Contrary to that hypothesis, the degree distribution of every composite network was better fit with an exponential than with a power law. Hence, at least for the parameter space explored here, it is tempting to conclude that when evaluating the degree distributions of networks, researchers do not need to be overly concerned with the possibility of power law mimicry. Furthermore, because we used network sizes and values of r intended to maximize the variability in the slopes of the individual networks' degree distributions, a condition Anderson and Tweney (1997) found necessary to produce power law mimicry, it may also be tempting to conclude that it is unlikely that power law mimicry will turn out to be an extensive problem when evaluating the degree distributions of complex systems.

However, power laws did provide relatively good fits to the composite networks, even though our individual networks were grown in a completely random manner, with no process akin to preferential attachment operating. Consequently, observing a good fit of a power law function to a degree distribution is not sufficient evidence that a process such as preferential attachment is operating. Minimally, the degree distribution also needs to be fit to an exponential distribution. If the exponential fits better than the power law, there is little evidence that anything

other than random processes are operating. To the extent that the power law provides a better fit than the exponential, given the results of the present simulations, increased confidence can be put into a claim that the observed network grew via a process such as preferential attachment. However, that observation alone cannot exclude the possibility that such a study had the misfortune of stumbling into some area of the parameter space where power law mimicry was operating.

Regrettably, in those studies using graph-theoretic analyses in areas of most interest to cognitive science, it does not seem to be routine practice to report fits of exponential functions to observed degree distributions. Steyvers and Tenenbaum (2005) reported that for 3 or their 4 semantic networks, a power law "almost perfectly" (p. 53) fit the observed degree distribution. The degree distribution of the fourth network "show[ed] a slight deviation from the power law form" (p. 54). (Although Steyvers and Tenenbaum seemed to conclude from this finding (along with the observed slopes of those distribution) that their networks generally showed a scale-free structure, they did reject the hypothesis of growth through preferential attachment due to other observed characteristics of their network.) They reported no quantitative measure of goodness of fit, and did not report on how well those degree distributions were fit by an exponential function. Ferrer and Solé (2001) similarly fit their degree distributions of the co-occurrence of English words with a power law but did not compare that fit to the fit of an exponential. Soares et al. (2005), in their study of the syllabic structure of Portuguese reported good fits of their degree distributions to power functions. They did not, however, report a measure of the goodness of fit, nor did they compare that fit to the fit of an exponential. They concluded that language evolves following a process similar to the preferential attachment rule.

Schweickert (in press b), in his study of the networks of characters appearing in individuals' dreams, did report goodness of fit measures for his fits to a power law degree distribution. The R^2 values for the networks of his three dreamers were .84, .85, and .82. These values are slightly lower than most of the R^2 values we observed for the random composite networks grown in the present study. Unfortunately, Schweickert did not report corresponding goodness of fit measures for exponential fits to his degree distributions.

Vitevitch (in press) did report on the fit of both power law and exponential functions to his degree distribution. He found a better fit for the exponential than for the power function. Interestingly, he then speculated on what developmental processes might produce such a degree distribution without mentioning the simplest such process: a network that grows over time, in which newly added nodes randomly attach to existing nodes with a probability proportional to the current size of the network (Barabási & Albert, 1999). That is, he did not consider the type of random graph simulated in the present study.

The degree distribution, of course, is not the only characteristic of graphs that can differentiate different growth processes that give rise to the observable network. Steyvers and Tenenbaum (2005), for example, rejected preferential attachment as the growth mechanism underlying their observed semantic networks because the clustering coefficient in their graphs was much larger than that predicted by preferential attachment. Although we have not yet done a systematic examination, we did look at the clustering coefficient in a few of the graphs generated via our random growth process and found them to be extremely small.[5] Soares et al. (2005), in their study of the syllabic structure of Portuguese words, and Vitevitch (in press), in his study of 6508 English words, similarly observed larger clustering coefficients than occur in a certain form

[5] Note that these clustering coefficients were calculated from the individual sub-networks, not from the composite networks.

of random graph, different from that studied here. At the present time, however, it is unclear whether such an observation actually does rule out growth through random processes in the domains studied by these researchers. Both studies considered a domain that has what Abler (1989) refers to as a particulate structure. In such a domain, an infinite or near infinite set of structures is constructed from a much smaller set of particulate units. The Portuguese words studied by Soares et al. were created out of syllable particulates. The English words studied by Vitevitch were created out of phoneme particulates. In each case, a relatively small number of particles is used to create a relatively large number of words. In von Humboldt's words (cited in Abler), words are constructed by "mak[ing] infinite use of finite media." Words are then linked by virtue of what amounts to having particles in common. Intuitively—and a stronger statement awaits further investigation—such a structure seems likely to produce dense neighborhoods with relatively high clustering coefficients. In brief, it has not been shown that a relatively high clustering coefficient necessarily rules out random growth processes.

To summarize, power law degree distributions are frequently observed when modeling a complex system as a graph. Such a ubiquitous finding inspires a search for a common cause. A possible common cause in this case is growth through preferential attachment. However, the random growth process described in the present simulations produced degree distributions that, though better fit by an exponential, were also well fit by a power law. As a science, we need to rule out simple explanations based on random processes before moving on to the more complex explanations.

References

Abler, W.L. (1989). On the particulate principle of self-diversifying systems. *Journal of Social & Biological Structures*, *12*, 1-13.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47–97.

Anderson, R.B. (2001). The power law as an emergent property. *Memory & Cognition, 29*, 1061-1068.

Anderson, R.B. & Tweney, R. (1997). Artifactual power curves in forgetting. *Memory & Cognition, 25*, 724 – 730.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.

Brown, S. & Heathcote, A. (2003a). Bias in exponential and power function fits due to noise: Comment on Myung, Kim and Pitt. *Memory and Cognition, 31*, 656-661.

Brown, S. & Heathcote, A. (2003b). Averaging learning curves across and within participants. *Behaviour Research Methods, Instruments & Computers, 35*, 11-21.

Estes, W.K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134 – 140.

Fellbaum, C. (Ed.) (1998). *WordNet, an electronic lexical database*. Cambridge, MA: MIT Press.

Ferrer, R. & Solé, R.V. (2001). The small world of human language. *Proceedings of the Royal Society of London B, 268*, 2261–2265.

Greenberg, J.H. & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word, 20*, 157–177.

Gruenenfelder, T.M. & Pisoni, D.B. (2005). Modeling the mental lexicon as a complex system: Some preliminary results using graph theoretic measures. In *Research on Spoken Language Processing Progress Report No. 27* (pp. 27-27). Bloomington, IN: Speech Research Laboratory, Indiana University.

Landauer, T.K. & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119–131.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1–36.

McClelland, J.L. & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375 – 407.

Miller, G. A. (1995). WordNet: An on-line lexical database[Special issue]. *International Journal of Lexicography, 3*(4).

Nelson, D.L., McEvoy, C.L., & Schreiber, T.A. (1999). *The University of South Florida word association norms.* Available at http://w3.usf.edu/FreeAssociation

Newell, A. & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Roget, P. M. (1911). *Roget's thesaurus of English words and phrases* (1911 ed.). Retrieved (by Steyvers & Tenebaum) October 28, 2004 from http://www.gutenberg.org/etext/10681

Schweickert, R. (in press a). The structure of semantic and phonological networks and the structure of a social network in dreams. In J.S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III*. New York: Psychology Press.

Schweickert, R. (in press b). Properties of the organization of memory for people: Evidence from dream reports. *Psychonomic Bulletin & Review*.

Soares, M.M., Corso, G., & Lucena, L.S. (2005). The network of syllables in Portuguese. *Physica A: Statistical Mechanics and its Applications, 355*, 678- 684.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Quarterly Journal of Reading, 21*, 360 – 407.

Steyvers, M. & Tenenbaum, J.B. (2005). The large-scale structure of semantic networks: Statistical analysis and a model of semantic growth. *Cognitive Science, 29*, 41-78.

Vitevitch, M.S. (in press). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech-Language-Hearing Research*.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Reduced Cluster Switching in Category Fluency Reveals Cognitive Decline: A Longitudinal Study[1]

Vanessa Taler, David B. Pisoni, Martin Farlow,[2] Ann Marie Hake,[2]
David Kareken[2] and Frederick Unverzagt[2]

Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405

[1] This research supported by NIH NIDCD R01 Research Grant DC00111 and NIH NIDCD T32 Training Grant DC00012 to Indiana University. V.T. is supported by Fonds de la Recherche en Santé de Québec.
[2] Currently at the Indiana University School of Medicine, Indianapolis.

Reduced Cluster Switching in Category Fluency Reveals Cognitive Decline: A Longitudinal Study

Abstract. Impairments in semantic fluency tasks are well-established in Alzheimer's disease (AD). These are apparent both in quantitative measures, namely total number of items produced, and qualititative measures, namely the frequency with which AD patients switch between semantic clusters (e.g., from farm animals to African animals). Similar deficits have been seen in quantitative output of individuals who will go on to develop AD or who have been diagnosed with mild cognitive impairment (MCI). However, less research has examined qualitative aspects of fluency performance in these populations. We assessed the fluency performance over time of twelve healthy elderly who went on to be diagnosed with MCI. Over a seven-year period, declines were seen in qualitative measures, specifically the number of cluster switches, but not in total output. The finding that switching between clusters on a semantic fluency task begins to decline up to seven years before diagnosis with MCI indicates that performance on this task may be an important predictor of future cognitive decline in healthy elderly adults.

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease affecting a number of cognitive domains, including memory, language and executive function (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; American Psychiatric Association, 1994). It has become clear in recent years that individuals with AD manifest cognitive deficits across a number of tasks even prior to meeting diagnostic criteria for probable AD (e.g., Flicker, Ferris, & Reisberg, 1991; Hodges & Patterson, 1995; Jacobs et al., 1995), and interest in early identification of individuals who will develop AD has surged.

Recently, Petersen et al. (1999) developed a set of criteria to identify individuals with mild cognitive impairment: subjective and objective memory impairment, generally preserved other cognitive function and absence of dementia, and no other neurological or psychiatric explanations for the memory impairment. MCI was found by Petersen et al. and a number of subsequent studies to constitute a significant risk factor for dementia, with around 15% of individuals meeting criteria for MCI developing probable AD per annum, versus 1-2% in the healthy elderly population (Chertkow, 2002). However, a significant proportion of MCI individuals remain undemented. Recent recommendations for revised MCI criteria include the criterion that the patient should show either impaired performance on cognitive tasks in the context of self and/or informant report of decline, or decline over time on objective cognitive tasks (Winblad et al., 2004). These criteria should assist in identifying those individuals who show cognitive decline from a high baseline, thus appearing unimpaired on cognitive testing.

One area in which deficits are often reported in pre-clinical AD is on lexical-semantic tasks such as verbal fluency (for a review of language performance in MCI and pre-clinical AD, see Taler & Phillips, in press). In verbal fluency tasks, participants must name as many items as possible that conform to a given criterion within a certain time limit (typically one minute). The criterion may either be semantic (e.g., animals) or phonemic (e.g., words beginning with the letter F). Both semantic and letter fluency tasks impose significant demands on executive function, since participants must organize verbal retrieval and recall, initiate responses, and monitor prior responses, as well as inhibiting inappropriate responses (Henry et al., 2004). However, unlike letter fluency, semantic fluency requires that participants

retrieve semantic extensions of a superordinate term. This task requires intact semantic associations within the mental lexicon (Rohrer, Salmon, Wixted, & Paulsen, 1999).

There exists a great deal of research demonstrating impairments in verbal fluency tasks in AD, and a recent meta-analysis (Henry, Crawford, & Phillips, 2004) indicated that, while both letter and semantic fluency are impaired in AD, the impairment is more severe in semantic than in letter fluency. This disparity is in part due to the degradation in semantic knowledge that occurs in AD; however, impairments in object naming are also less severe than those in semantic fluency, suggesting that deficits in executive function that affect semantic search may also play a role. The research to date indicates that declines in verbal fluency performance are also seen in MCI, particularly in category fluency (for a review, see Taler & Phillips, 2007). Similarly, deficits in semantic fluency have been observed in individuals at risk for AD, either because they are carrying the APOE-4 allele or due to a family history of the disease (Miller, Rogers, Siddartha, & Small, 2005).

The vast majority of research to date has analyzed category fluency scores without examining more closely the qualitative aspects of category fluency performance in these populations. Troyer, Moscovitch, Winocur, Leach and Freedman (1998) were the first to report qualitative alterations in verbal fluency performance in AD. They focused on aspects of semantic search, analyzing the number of semantic clusters (groups of semantically or phonemically related items) and the number of switches from one category to another that occur in AD patients' output on this task. They found that AD patients produced smaller clusters on both letter and semantic fluency, and fewer switches on semantic fluency than healthy control participants. Subsequent research has confirmed this finding and indicated that that clustering and switching variables can discriminate between very mild AD and healthy elderly (Gomez & White, 2006).

The present study extends previous research on semantic fluency performance in pre-clinical AD, examining alterations in performance over time. We report on quantitative and qualitative aspects of semantic fluency performance in healthy control participants who remain cognitively intact as well as a group who were subsequently diagnosed with MCI. Participants were assessed annually and their performance over time was analyzed.

Methods

Participants

A total of 29 participants were included in the present study: 17 healthy elderly who remained cognitively intact and 12 healthy elderly who went on to be diagnosed with MCI. All participants were native speakers of English with no neurological or psychiatric history, other than MCI, and were right-handed. The diagnosis of MCI was established according to criteria similar to those proposed by Petersen et al. (1999; 2001). For the healthy elderly who remained unimpaired, average follow-up time was 4.47 years, with an average of 4.12 assessments. In the group who were eventually diagnosed with MCI, the average follow-up time was 4.83 years, with an average of 3.67 assessments. For the MCI group, year 0 was defined as time of diagnosis, and performance in the seven years prior was entered into the analysis. Further details about the participants are provided in Table 1.

	CN Participants – Mean (SD)	CN Participants – Range	CN to MCI Participants – Mean (SD)	CN to MCI Participants – Range
n	17		12	
age	67.59 (8.13)	55-79	72.58 (6.67)	61-81
sex	10 women/7 men		5 women/7 men	
education	15.53 (2.55)	12-20	15.17 (3.49)	8-19
MMSE (/30)*	29.59 (0.62)	28-30	28.33 (1.56)	25-30
BNT (/15)*	14.88 (0.33)	14-15	14.08 (1.24)	11-15
COWA (letter)	39.53 (11.03)	26-66	42.75 (10.81)	26-62

Table 1. Participant characteristics, baseline assessment.
BNT=Boston Naming Test. CN=cognitively normal. COWA=Controlled Oral Word Association. MCI=mild cognitive impairment. MMSE=Mini-Mental State Examination.
*groups differ at $p < 0.05$

Procedures and Scoring

As part of a larger neuropsychological battery, participants completed a semantic fluency task in which they were asked to name as many animals as they could in one minute. This task was completed at each neuropsychological assessment for a period of up to seven years. For each participant at each assessment, total number of responses, excluding errors and repetitions, was recorded.

In addition to total scores, semantic fluency performance was coded according to switching and clustering. Following the guidelines set out by Troyer, Moscovitch and Winocur (1997), participants' output was scored for total number of times that the participant moved from one semantic subcategory to another (switches) and mean number of items produced in each subcategory (clusters). Clusters included animals that were similar in terms of living environment (e.g., water animals, African animals); zoological categories (e.g., birds, rodents); or human use (e.g., pets, beasts of burden)[3]. Following Troyer et al., number of switches was coded as [total number of clusters – 1], and cluster size was coded as [total items in cluster – 1].

Results

Figure 1 presents the mean total items generated over time by each group. Those participants who remain cognitively intact appear to increase in total number of items produced, while those who go on to be diagnosed with MCI do not. One-tailed Pearson correlations reveal a borderline positive

[3] For a more comprehensive list of categories, see Troyer et al. In our analysis, three additional categories were included: South American animals (e.g., llama, alpaca); equids (e.g., horse, zebra, mule); and nocturnal pests (raccoon, possum, skunk).

correlation between year and total items produced for the healthy CN group (r=0.17, p<0.08) but no such correlation for the group who went on to be diagnosed with MCI (r=-0.16, p>0.15).

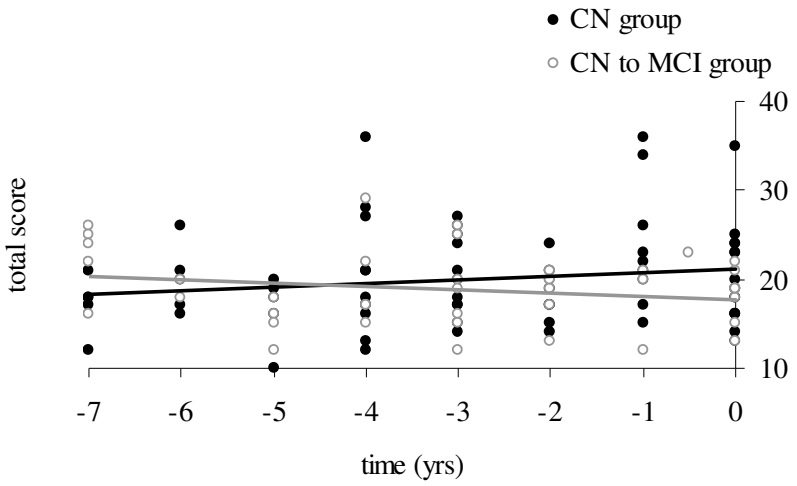


Figure 1. Total number of items generated at each assessment by each participant group. Year 0 = time of diagnosis for converter group. Trend lines represent best linear fit.

In Figure 2, the correlation between number of switches between semantic clusters and year of assessment is shown for each group. One-tailed Pearson correlations reveal a significant negative correlation between year of assessment and number of switches for the CN to MCI group (r=-0.26, p<0.05) but not for the group who remained cognitively intact (r=0.019, p>0.44).

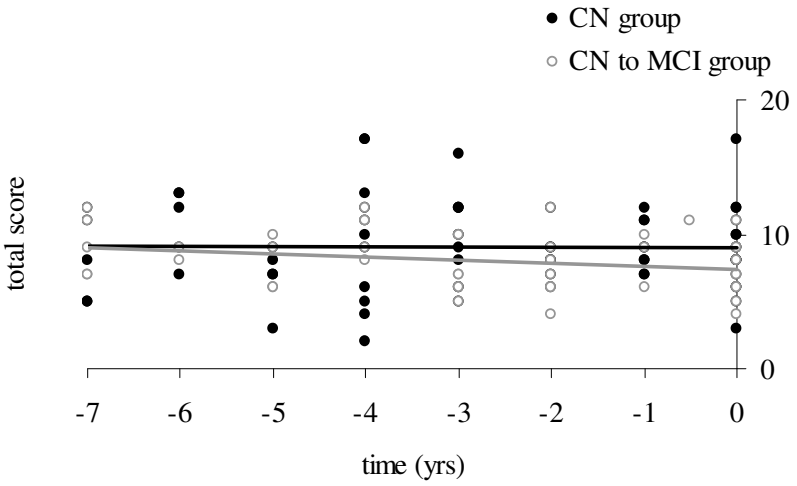


Figure 2. Total number of switches generated at each assessment by each participant group. Year 0 = time of diagnosis for converter group. Trend lines represent best linear fit.

Finally, Figure 3 shows the correlations between average cluster size and year of assessment for each group. No significant correlation was seen between year of assessment and cluster size in either group (CN: r=-0.13, p>0.14; CN to MCI: r=0.03, p>0.43).

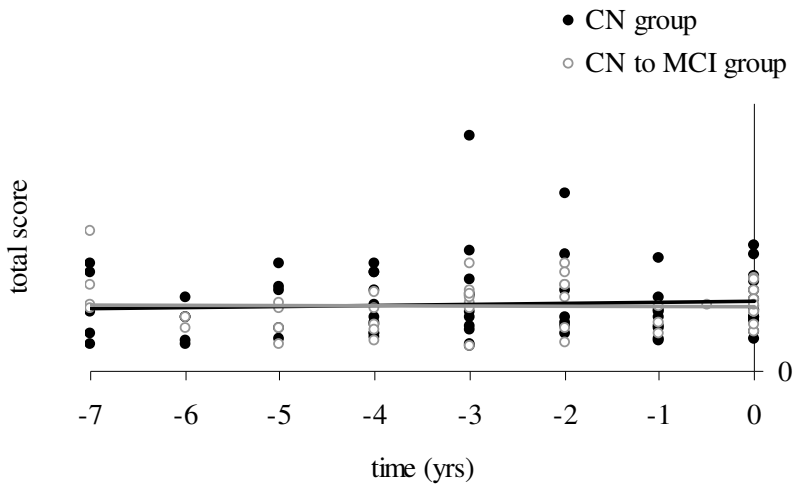


Figure 3. Average size of cluster generated at each assessment by each participant group. Year 0 = time of diagnosis for converter group. Trend lines represent best linear fit.

Discussion

Overall, healthy control participants showed a marginally significant increase in the number of items generated, and no change in the number of cluster switches or cluster size over the course of seven years. In contrast, in the seven years prior to diagnosis, those participants who were eventually diagnosed with MCI showed no change in the number of items generated and a decrease in the number of cluster switches, while the cluster size remained the same.

The results reported here are consistent with previous reports indicating declines in number of cluster switches and stability in cluster size in individuals with a diagnosis of AD (e.g., Troyer et al., 1998). However, to our knowledge, this is the first report to examine the qualitative aspects of category fluency performance longitudinally in a pre-clinical population. The present results indicate that these changes in cluster switching begin much earlier than previously reported, many years prior to the appearance of any objective memory impairment.

That cluster size remains the same suggests that the declines in performance observed here are driven not by impairment in semantic representation per se, but likely by a deficit in executive search within semantic memory. Previous studies of semantic memory in MCI have pointed to similar conclusions (e.g., Duong, Whitehead, Hanratty, & Chertkow, 2006), and the present research indicates that these executive search impairments are a very early marker of cognitive decline in MCI.

It is also of interest that the total number of items produced remains relatively stable in individuals who will be diagnosed MCI, while increasing in healthy control participants. It seems likely that healthy participants are able to recall the tasks included in previous testing sessions, particularly when the same tasks are used over several consecutive annual neuropsychological assessments. These individuals are thus able to benefit from developing strategies over multiple testing sessions. This finding

emphasizes the importance of using alternate versions of tasks such as semantic fluency (for a discussion of the validity of alternate versions of fluency tasks, see Cunje, Molloy, Standish, & Lewis, 2007). Participants who will go on to be diagnosed with MCI, in contrast, are not able to benefit from these practice effects, suggesting deficits in episodic memory. This finding has been reported previously in clinically diagnosed MCI (Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004), but to our knowledge this is the first report indicating that declines in practice effects over annual sessions are seen in healthy participants who will go on to develop MCI.

In conclusion, the finding that switching between clusters begins to decline up to seven years prior to diagnosis with MCI has important ramifications for early detection of AD. Semantic fluency is a task that is included in routine neuropsychological evaluations, and as such these data are readily available to the clinician in providing a prognosis for the elderly patient. We thus believe that qualitative as well as quantitative analysis of semantic fluency performance in elderly participants is of great potential value in clinical practice.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Chertkow, H. (2002). Mild cognitive impairment. *Current Opinion in Neurobiology, 15,* 401-407.

Cooper, D.B., Lacritz, L.H., Weiner, M.F., Rosenberg, R.N., & Cullum, C.M. (2004). Category fluency in mild cognitive impairment: reduced effect of practice in test-retest conditions. *Alzheimer Disease and Associated Disorders, 18,* 120-122.

Cunje, A., Molloy, D.W., Standish, T.I., & Lewis, D.L. (2007). Alternate forms of logical memory and verbal fluency tasks for repeated testing in early cognitive changes. *International Psychogeriatrics, 19*, 65-75.

Duong, A., Whitehead, V., Hanratty, K., & Chertkow, H. (2006). The nature of lexico-semantic processing deficits in mild cognitive impairment. *Neuropsychologia, 44,* 1928-1935.

Flicker, C., Ferris, S.H., & Reisberg, B. (1991). Mild cognitive impairment in the elderly: predictors of dementia. *Neurology, 41,* 1006-1009.

Gomez, R.G., & White, D.A. (2006). Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology, 21,* 771-775.

Henry, J.D., Crawford, J.R., & Phillips, L.H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia, 42,* 1212-1222.

Hodges, J.R., & Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia, 33,* 441-459.

Jacobs, D.M., Sano, M., Dooneief, G., Marder, K., Bell, K.L., & Stern, Y. (1995). Neuropsychological detection and characterization of preclinical Alzheimer's disease. *Neurology, 45,* 317-324.

Miller, K.J., Rogers, S.A., Siddartha, P., & Small, G.W. (2005). Object naming and semantic fluency among individuals with genetic risk for Alzheimer's disease. *International Journal of Geriatric Psychiatry, 20,* 128-136.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology, 56,* 303-308.

Petersen, R.C., Stevens, J.C., Ganguli, M., Tangalos, E.G., Cummings, J.L., & DeKosky, S.T. (2001). Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review). *Neurology, 56,* 133-142.

Rohrer, D., Salmon, D.P., Wixted, J.T., & Paulsen, J.S. (1999). The disparate effects of Alzheimer's disease and Huntington's disease on semantic memory. *Neuropsychology, 13,* 381-388.

Taler, V., & Phillips, N.A. (in press). Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology.*

Troyer, A.K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology, 11,* 138-146.

Troyer, A.K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society, 4*, 137-143.

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., et al. (2004). Mild cognitive impairment - beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine, 256,* 240-246.

III. Publications

ARTICLES PUBLISHED:

Burkholder-Juhasz, R.A., Levi, S.V., Dillon, C.M. & Pisoni, D.B. (2007). Nonword repetition with spectrally reduced speech: Some developmental and clinical findings from pediatric cochlear implantation. *Journal of Deaf Studies and Deaf Education, 12,* 472-485.

Clopper, C.G. & Pisoni, D.B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics, 35,* 421-438.

Clopper, C.G., Pisoni, D.B. & Tierney, A.T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology, 17*, 331-349.

Conway, C.M., Karpicke, J. & Pisoni, D.B. (2007). Contribution of implicit sequence learning to spoken language processing: Some preliminary findings with normal-hearing adults. *Journal of Deaf Studies and Deaf Education, 12,* 317-334.

Dillon, C.M. & Pisoni, D.B. (2006). Nonword repetition and reading skills in children who are deaf and have cochlear implants. *Volta Review, 106*, 121-145.

Fagan, M.K., Pisoni, D.B., Horn, D.L., & Dillon, C.M. (2007). Neuropsychological processes associated with vocabulary, reading, and working memory in deaf children with cochlear implants. *Journal of Deaf Studies and Deaf Education, 12,* 461-471.

Horn, D.L., Fagan, M.K., Dillon, C.M., Pisoni, D.B. & Miyamoto, R.T. (2007). Visual-motor skills of pre-lingually deaf children: Implications for pediatric cochlear implantation. *Laryngoscope, 117,* 2017-2025.

Levi, S.V., Winters, S.J. & Pisoni, D.B. (2007). Speaker-independent factors affecting the degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 121,* 2327-2338.

BOOK CHAPTERS PUBLISHED:

Bent, T. & Pisoni, D.B. (2007). Some comparisons in perception between speech and nonspeech signals. In M. Ball (Ed.), *Handbook of Clinical Linguistics.* Pp. 400-411. Blackwell Publishers.

Burkholder, R.A., & Pisoni, D.B. (2006). Working memory capacity, verbal rehearsal speed, and scanning in deaf children with cochlear implants. In P.E. Spencer & M. Marschark (Eds.), *Advances in the Spoken Language Development of Deaf and Hard-of-Hearing Children*. Pp. 328-357. Oxford University Press.

Levi, S.V. & Pisoni, D.B. (2007). Indexical and linguistic channels in speech perception: Some effects of voiceovers on advertising outcomes. In T. Lowery (Ed.), *Psycholinguistic phenomena in marketing communications.* Pp. 203-219. Mahwah, NJ: Lawrence Erlbaum.

Pisoni, D.B. & Levi, S.V. (2007). Representations and representational specificity in speech perception and spoken word recognition. In M.G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics.* Pp. 3-18. Oxford University Press: UK.

PROCEEDINGS PUBLISHED:

Christiansen, M.H., Conway, C.M., & Onnis, L. (2007). Neural responses to structural incongruencies in language and statistical learning point to similar underlying mechanisms. In D.S. McNamara & J.G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 173-178). Austin, TX: Cognitive Science Society.

Conway, C.M., Goldstone, R.L., & Christiansen, M.H. (2007). Spatial constraints on visual statistical learning of multi-element scenes. In D.S. McNamara & J.G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 185-190). Austin, TX: Cognitive Science Society.

Conway, C.M. & Pisoni, D.B. (2007). Links between implicit learning of sequential patterns and spoken language processing. In D.S. McNamara & J.G. Trafton (Eds), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 191-196). Austin, TX: Cognitive Science Society.

Felty, R. (2007). Confusion patterns and response bias in spoken word recognition of German disyllabic words and nonwords. In *Proceedings of the 16th International Congress of the Phonetic Sciences*, 1957-1960.

Kapatsinski, V.M. (2007). Frequency, neighborhood density, age-of-acquisition, lexicon size, neighborhood density and speed of processing: Towards a domain-general, single-mechanism account. In S. Buescher, K. Holley, E. Ashworth, C. Beckner, B. Jones, & C. Shank (eds.), *Proceedings of the 6th Annual High Desert Linguistics Society Conference*, 121-40. High Desert Linguistics Society: Albuquerque, NM.

Kapatsinski, V. (2006). Towards a single-mechanism account of frequency effects. In S.J.J. Hwang, W.J. Sullivan & A.R. Lommel (eds.), Lacus *Forum XXXII: Networks*, 325-36. The Linguistic Association of Canada and the United States: Houston, TX.

Ronquest, R. & Díaz-Campos, M. (2007). A perceptual discrimination study of pitch-accent alignment in Spanish. Presented at the *Linguistic Symposium on Romance Languages XXXVII (LSRL 37).* University of Pittsburgh, March 15-18.

Ronquest, R. & Díaz-Campos, M. (2007) Discriminating pitch-accent alignment in Spanish: A pilot investigation. Presented at the *2007 Kentucky Foreign Language Conference*. University of Kentucky, April 19-21.

MANUSCRIPTS ACCEPTED FOR PUBLICATION (IN PRESS):

Bent, T., Bradlow, A.R., & Smith, BL. (In press). Production and perception of temporal patterns in native and non-native speech. *Phonetica*.

Bradlow, A.R. & Bent, T. (In press). Perceptual adaptation to non- native speech. *Cognition*.

Clopper, C.G. & Paoillo, J.C. (In press). North American English vowels: A factor analytic perspective. *Literary and Linguistic Computing.*

Conway, C.M. & Pisoni, D.B. (In press). Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences.*

Díaz-Campos, M., & Ronquest, R. (In Press). La percepción de acentos tonales en enuciados afirmativos. *Estudios de Fontéica Experimental, XVI*, pp. 81-98.

Harnsberger, J.D., Pisoni, D.B. & Wright, R. (In press). A new method for eliciting three speaking styles in the laboratory. *Speech Communication*.

Harnsberger, J.D., Wright, R. & Pisoni, D.B. (In press). Effects of speaking style on the perceptual learning of novel voices: A first report. *Phonetica.*

Kapatsinski, V. & Radicke, J. (In press). Frequency and the emergence of prefabs: Evidence from monitoring. In B. Corrigan, E. Moravcsik, H. Ouali, & K. Wheatley (eds.), *Formulaic Language*. Amsterdam, Philadelphia: John Benjamins.

Levi, S.V. (In press). Reconsidering the variable status of glottals in Nasal Harmony. *Chicago Linguistic Society 41*.

Loebach, J.L. & Pisoni, D.B. (In press). Perceptual learning of spectrally degraded speech and environmental sounds. *Journal of the Acoustical Society of America.*

Pisoni, D.B., Kronenberger, W, Conway, C.M., Horn, D.L., Karpicke, J. & Henning, S. (In press). Efficacy and effectiveness of cochlear implants in deaf children. In M. Marschark & P. Hauser (Eds.), *Deaf Cognition: Foundations and Outcomes*. New York: Oxford University Press.

MANUSCRIPTS SUBMITTED:

Buchwald, A.B., Winters, S.J. & Pisoni, D.B. (Submitted). Multimodal speech perception: Evidence form cross-modal priming.

Conway, C.M. & Christiansen, M.H. (submitted). Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *European Journal of Cognitive Psychology.*

Dillon, C.M. & Pisoni, D.B. (Submitted). Phonological awareness, reading skills and vocabulary knowledge in deaf children who use cochlear implants. *Journal of Deaf Studies and Deaf Education.*

Poletiek, F.H., Conway, C.M., Ellefson, M.R., & Christiansen, M.H. (submitted). Effects of starting small in artificial grammar learning of recursive structure. *Journal of Experimental Psychology: General*.

Radicke, J.L., Levi, S.V., Loebach, J.L. & Pisoni, D.B. (Submitted). Audiovisual phonological fusion. *Perception & Psychophysics.*

Ronquest, R.E., Levi, S.V. & Pisoni, D.B. (Submitted). Language identification from visual-only speech. *Perception & Psychophysics*.

Tierney, A.T., Pisoni, D.B. & Bergeson-Dana, T. (Submitted). Some effects of early musical experience on auditory sequence memory. *Psychology of Music.*

Winters, S.J., Levi, S.V. & Pisoni, D.B. (Submitted). When and why feedback matters in the perceptual learning of visual displays of speech. *Journal of the Acoustical Society of America.*

Winters, S.J., Levi, S.V. & Pisoni, D.B. (Submitted). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America*.