RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 23
(1999)

David B. Pisoni, Ph.D.
Principal Investigator

Speech Research Laboratory
Department of Psychology
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

1999
Indiana University

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)

Table of Contents

INTRODUCTION

This is the twenty-third annual progress report summarizing research activities on speech perception and spoken language processing carried out in the Speech Research Laboratory, Department of Psychology, 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.

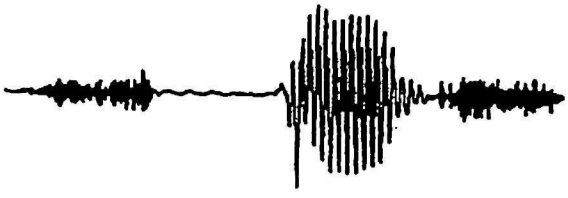
We are distributing progress reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of spoken language processing. We are, of course, very interested in following 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 Psychology
Indiana University
Bloomington, Indiana 47405-1301
United States of America

Telephone: (812) 855-1155, 855-1768
Facsimile: (812) 855-4691
E-mail: pisoni@indiana.edu
Web: http://www.indiana.edu/~srlweb

Copies of this progress report are being sent primarily to libraries and specific 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, 1999–December 31, 1999)

RESEARCH PERSONNEL

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

Karen I. Kirk, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3,4]
Mario A. Svirsky, Ph.D. ..... Associate Professor of Otolaryngology–Head and Neck Surgery[3,5]
Steven B. Chin, Ph.D. ..... Assistant Scientist in Otolaryngology–Head and Neck Surgery[3]

Allyson K. Carter, Ph.D. ..... NIH Postdoctoral Trainee
James D. Harnsberger, Ph.D. ..... NIH Postdoctoral Trainee
Rebecca Herman, Ph.D. ..... NIH Postdoctoral Trainee
Adam R. Kaiser, M.D., Ph.D. ..... NIH Postdoctoral Trainee[3]
Holly L. Storkel, Ph.D. ..... NIH Postdoctoral Trainee
Michael S. Vitevitch, Ph.D. ..... NIH Postdoctoral Trainee

Sarah H. Ferguson, M.A. ..... NIH Predoctoral Trainee
Laura W. McGarrity, M.A. ..... NIH Predoctoral Trainee
Miranda Cleary, M.A. ..... Predoctoral Trainee
Elizabeth J. Cole, B.A. ..... Predoctoral Trainee
Cynthia G. Clopper, B.A. ..... Predoctoral Trainee
Winston D. Goh, M.Soc.Sci. ..... Predoctoral Trainee[6]
Lorin Lachs, B.A. ..... Predoctoral Trainee
Kipp H. McMichael, B.S. ..... Predoctoral Trainee

Brian Chung, B.A. ..... NIH Medical Student Trainee
Ashesh Shah, B.S. ..... NIH Medical Student Trainee
Peter Simmons, B.S. ..... NIH Medical Student Trainee
Corbett M. Smith, B.A. ..... NIH Medical Student Trainee
Woo J. Yi, B.A. ..... NIH Medical Student Trainee

---

[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] Also Adjunct Assistant Professor of Speech and Hearing Sciences, Indiana University, Bloomington, IN.

[5] Also Adjunct Associate Professor of Electrical Engineering, Purdue School of Engineering and Technology, Indianapolis, IN.

[6] Also Senior Tutor, Department of Social Work and Psychology, National University of Singapore.

TECHNICAL PERSONNEL

Luis R. Hernández, B.A. ..... Research Associate in Psychology/Systems Administrator
Darla J. Sallee ..... Administrative Assistant
Carlos Colon ..... Programmer

Jaime Brumfield ..... Undergraduate Research Assistant
Patrick Kelley ..... Undergraduate Research Assistant

E-MAIL ADDRESSES

Jaime Brumfield ..... jbrumfie@indiana.edu
Allyson K. Carter ..... allcarte@indiana.edu
Steven B. Chin ..... schin@iupui.edu
Miranda Cleary ..... micleary@indiana.edu
Elizabeth J. Cole ..... elcole@indiana.edu
Carlos Colon ..... ccolon@indiana.edu
Cynthia G. Clopper ..... cclopper@indiana.edu
Sarah H. Ferguson ..... safergus@indiana.edu
Winston D. Goh ..... wigoh@indiana.edu
James D. Harnsberger ..... jharnsbe@indiana.edu
Rebecca Herman ..... rherman@indiana.edu
Luis R. Hernández ..... hernande@indiana.edu
Adam R. Kaiser ..... arkaiser@iupui.edu
Patrick Kelley ..... pdkelley@indiana.edu
Karen I. Kirk ..... kkirk@iupui.edu
Lorin Lachs ..... llachs@indiana.edu
Laura W. McGarrity ..... lmcgarr@indiana.edu
Kipp H. McMichael ..... kimcmich@indiana.edu
David B. Pisoni ..... pisoni@indiana.edu
Darla J. Sallee ..... dsallee@indiana.edu
Holly L. Storkel ..... hstorkel@indiana.edu
Mario A. Svirsky ..... msvirsky@iupui.edu
Michael S. Vitevitch ..... mvitevit@indiana.edu

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Individual Differences in Effectiveness of Cochlear Implants in Prelingually Deaf Children: Some New Process Measures of Performance[1]

David B. Pisoni[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This research was supported by NIH-NIDCD Research Grants DC00111, DC00012, DC00064 and DC00423 to Indiana University. This is a revised and substantially expanded version of an invited talk given at the Acoustical Society of America meetings held in Columbus, Ohio, November 2-5, 1999. I wish to extend special thanks to Miranda Cleary, Ann Geers and Emily Tobey for their help in various aspects of this project. Without their contribution, this research would not have been possible.

[2] Also, DeVault Otologic Research Laboratory, Department of Otolaryngology, Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Individual Differences in Effectiveness of Cochlear Implants in Prelingually Deaf Children: Some New Process Measures of Performance

Abstract. The "efficacy" of cochlear implants in deaf children has been firmly established in the literature. However, the "effectiveness" of cochlear implants varies widely and is influenced by demographic and experiential factors. Several "key" findings suggest new directions for research on central auditory factors that underlie the effectiveness of cochlear implants. First, there are enormous individual differences in both adults and children on audiological outcome measures. Some patients show large increases in speech perception scores after implantation whereas others display only modest gains on standardized tests. Second, age of implantation and length of deafness affect all outcome measures. Children implanted at younger ages do better than children implanted at older ages and children who have been deaf for shorter periods of time do better than children who have been deaf for longer periods of time. Third, communication mode affects outcome measures. Children from "oral-only" environments do much better on standardized tests that assess phonological processing skills than children who use "total communication." Fourth, there are no preimplant predictors of outcome performance in young children. The underlying perceptual, cognitive and linguistic abilities and skills "emerge" after implantation and improve over time. Finally, there are no significant differences in audiological outcome measures among current implant devices or processing strategies. Taken together, this overall pattern of results suggests that higher-level central cognitive processes such as perception, attention, learning and memory may play important roles in explaining the enormous individual differences observed among users of cochlear implants. Investigations of the content and flow of information in the central nervous system and interactions between sensory input and stored knowledge may provide new insights into individual differences. Knowledge about the underlying basis of individual differences may also help in developing new intervention strategies to improve the effectiveness of cochlear implants in children who show relatively poor development of oral/aural language skills.

Introduction

When I first started to work on cochlear implants, I was immediately struck by the enormous variation in outcome and the large individual differences in performance. I wondered why some children do so well with their cochlear implant and why do other children do so poorly. I thought about this problem on and off for a couple of years and asked a lot of questions at our weekly clinical meetings. The clinicians in our group—the audiologists and speech pathologists who hook up and adjust the implants and test the children day in and day out could always count on me to come up with some unusual questions whenever an interesting child was discussed.

The problem of individual differences in the "effectiveness" of cochlear implants has been in the back of mind for a long time and has always intrigued me because of the challenge it presents for researchers interested in speech perception and spoken language processing. The NIDCD also considers the problem of individual differences to be an important area of research. The 1995 Consensus Statement on

Cochlear Implants in Adults and Children identified this topic as one of the major new directions for research. And, the study of individual differences is also a goal of our research program at the IU School of Medicine. We have moved into a number of new directions in order to understand the sensory, perceptual and cognitive basis for these differences. In the sections below, I give a summary of some of our most recent findings and some implications for future research on this issue.

At the present time, there are very few questions about the "efficacy" of cochlear implants in profoundly deaf children. Cochlear implants work and for some children they work well enough to permit them to develop spoken language through the auditory modality (Waltzman & Cohen, 2000). One of the most difficult problems with cochlear implants in deaf children, however, concerns the "clinical effectiveness" of these devices. Cochlear implants work well in some children but not others and no one seems to have come up with a very good explanation of why this happens. If we eliminate differences due to the number of active electrodes that provides the initial sensory information, there are not a lot of additional factors to investigate other than demographics and device characteristics. Psychophysical differences in frequency and intensity resolution may play an important role in setting initial constraints on how the sensory information is encoded at the auditory periphery but this is not the whole story. Something else is going on at more central levels of processing beyond the auditory nerve. We think individual variation in performance on outcome measures may be related to processing information at more central levels of analysis that are strongly affected by cognitive processes such as perception, attention, learning and memory. But there is very little, if any, research on these factors yet.

Almost all of the past research on cochlear implants has focused on demographic variables and traditional outcome measures using assessment tools developed by clinical audiologists and speech pathologists. Outcome measures of performance are the final product of a large number of complex sensory, perceptual, cognitive and linguistic processes that may be responsible for the observed variation among cochlear implant users. Until our recent studies reported below, no research has focused on "process" or examined the underlying mechanisms used to perceive and produce spoken language. Understanding these intermediate processes may provide new insights into the basis for these individual differences.

In addition to the enormous individual differences in outcome measures, several other findings have been consistently reported in the literature on cochlear implants in children. Age of Implantation has also been shown to affect outcome measures. Children who receive an implant at a young age after short periods of auditory deprivation do much better on a whole range of performance measures than children who are implanted at an older age after longer periods of sensory deprivation. Length of deprivation or length of deafness also predicts outcome. Children who have been deaf for shorter periods of time do much better on a variety of outcome measures than children who have been deaf for longer periods of time. Both findings demonstrate the important contribution of sensitive periods in development and the close links between neural development and behavior, especially, hearing, speech and language development (Ball & Hulse, 1998; Konishi, 1985; Konishi & Nottebohm, 1969; Marler & Peters, 1988).

Communication Mode also affects performance on a wide range of outcome measures. Implanted children who are immersed in "Oral-only" communication environments do much better on standardized tests than implanted children who are in "Total Communication" programs. The differences in performance between these two groups of children are seen most prominently in both receptive and expressive language tasks that involve phonological coding and phonological processing skills such as open-set spoken word recognition, comprehension and speech production.

Until just recently, researchers have been unable to identify reliable preimplant predictors of outcome and success with a cochlear implant. This is a critical theoretical finding because it demonstrates the existence of complex interactions between the newly acquired sensory capabilities of a child after a period of sensory deprivation, attributes of the language-learning environment and the interactions that the child is exposed to early on after receiving a cochlear implant. The lack of preimplant predictors also makes it difficult for both researchers and clinicians to identify those children who are doing poorly at a time in development when changes can be made to improve their language processing skills.

Finally, when all of the outcome and demographic measures are considered together, the evidence strongly suggests that the underlying abilities for speech and language "emerge" after implantation and that performance with a cochlear implant improves over time. Success with a cochlear implant therefore appears to be due to several different kinds of "learning" and exposure to the target language in the environment. Because success with a cochlear implant cannot be predicted reliably from traditional behavioral measures obtained before implantation, any improvement in performance observed after implantation must be due to learning processes that are correlated with maturational changes in neural and perceptual development.

Taken together, these "five key" findings suggest several general conclusions about the way cochlear implants work to facilitate the acquisition and development of spoken language. These findings also point to several underlying factors that affect performance on various outcome measures. Our current hypothesis about the source of individual differences is that while some proportion of the total variance in performance is clearly due to peripheral factors related to audibility and the initial sensory encoding of the speech signal into "information-bearing" sensory channels in the auditory nerve, an additional source of variance may also come from more central "cognitive" factors that are related to processes such as perception, attention, learning and memory. This source of variance is related to information processing operations and cognitive demands—that is, how the child uses the initial sensory input he/she receives from the cochlear implant and how the environment modulates, shapes and facilitates this learning process. These processes are, of course, topics that are the "meat and potatoes" of what cognitive psychologists and cognitive scientists study, namely, the encoding, rehearsal, storage and retrieval of information and the transformation and manipulation of memory codes and neural representations of the initial sensory input in a wide range of language processing tasks.

About three years ago, my colleagues and I began analyzing a set of data from our longitudinal project on cochlear implants in children to get a better handle on the issue of individual differences and variation in outcome. We began by looking at the "exceptionally" good users of cochlear implants—the so-called "Stars." These are the children who did extraordinarily well with their cochlear implants two years after implantation. They were able to acquire spoken language relatively quickly and easily and seemed to be on a developmental trajectory that parallels normal-hearing children. In many ways, at first glance, they look like normal hearing and normally developing children who simply have language delays.

Our interest and motivation in studying the "Stars" came, in part, from an extensive body of research in cognitive psychology over the last twenty-five years on "expertise" and "expert systems" theory (Ericsson & Pennington, 1993; Ericsson & Smith, 1991). Many important insights have come from studying expert chess players, radiologists and other people who have highly developed skills in specific knowledge domains like computer programming, spectrogram readings and even chicken sexing!! The rationale underlying our approach to the problem of individual differences was that if we could learn something about the "Stars" and the reasons why they do so well with their cochlear implants by adopting the orientation of expert systems theory, perhaps we could use this knowledge to develop new intervention

techniques with children who are not doing very well with their implants. Knowledge and understanding of the "Stars" would also be very useful in developing new pre-implant predictors of performance, in modifying current criteria for candidacy and in creating better and more precise methods of assessing performance and measuring outcome over time.

An initial report describing our findings on the "Stars" was presented in NYC in 1997 at the Xth International Cochlear Implant conference (Waltzman & Cohen, 2000). At that time, I presented longitudinal data collected over a period three years after implantation. We now have additional data on these children over six years that I will present below. Since that time, our research on individual differences has continued and expanded into several new directions as we try to understand the nature of these underlying factors. I was also very fortunate to begin collaborating with Ann Geers and her colleagues at CID who obtained some new data on working memory from forty-five 8 and 9 year olds who had used their cochlear implant for five years. New data on digit spans provided an opportunity to test a critical hypothesis about differences in information processing in children with cochlear implants. The work on digit spans then led to other analyses, development of new methodologies to measure working memory and several additional experiments on coding and rehearsal strategies that will be summarized below. We have developed a new methodology to study verbal and spatial coding in children with cochlear implants. Our initial findings were very encouraging and the results have provided some new insights into the underlying basis for the large individual differences observed in these children.

Theoretical Approach

Before I present the results of these studies, it is appropriate to say a few words about the theoretical motivation that underlies our research program on individual differences. Previous research on cochlear implants has relied very heavily on traditional outcome measures of performance that were developed within the field of clinical audiology. Historically, this research orientation focused on static assessment measures based on accuracy, device characteristics and demographic variables. In the past, there has been little if any concern or interest in "process" or a description of the underlying perceptual, cognitive or linguistic mechanisms that mediate performance. Researchers working on cochlear implants are interested in measuring change in performance over time but they have not studied change with an interest in describing the underling neural or behavioral mechanisms or the flow and contents of information in the nervous system.

In contrast, our research program on individual differences is motivated by several general theoretical principles that come from the field of human information processing and cognitive science. We are interested in describing and understanding the kinds of sensory and perceptual information that a child gets from his/her implant. We investigate and try to understand the nature of the phonetic, phonological and lexical representations that the child creates and how these are used in various language processing tasks. In adopting the information processing approach, our goal is to describe the "stages of processing" and to trace out the "time-course" of the various transformations this information takes from stimulus input to an observer's overt response in a specific task. This theoretical perspective is very different from the traditional approach used in clinical research on patients with cochlear implants that has focused almost exclusively on assessment and outcome measures. We hope that our approach will provide new knowledge about the underlying source of the individual differences observed among children and some of the factors that affect performance on traditional outcome measures.

Analysis of The "Stars"

Now let me turn to a summary of the major findings we obtained in our analyses of the "Stars." We have analyzed data obtained from several different outcome measures over a period of six years from the time of implantation in order to examine changes in performance over time. Before I present these results, however, I will describe how we originally identified and selected the "Stars" and the comparison group for our analyses.

The criterion used to identify the "Stars," the exceptionally good users of their cochlear implant, was based on performance on one particular perceptual test, the Phonetically Balanced Kindergarten (PBK) Words test (Haskins, 1949), which is an open-set test of spoken word recognition (also see Meyer & Pisoni, 1999). Among clinicians, this particular test is considered to be very difficult for prelingually deaf children compared to other, closed-set perceptual tests that are routinely included in a standard assessment battery (Zwolan, Zimmerman-Phillips, Asbaugh, Hieber, Kileny & Telian, 1997). The children who do moderately well on the PBK test frequently display ceiling levels of performance on all of the other closed-set speech perception tests that measure speech pattern discrimination. In contrast, open-set tests like the PBK measure word recognition and lexical discrimination and require the child to search and retrieve the phonological representations of the test words from lexical memory. Open-set tests of word recognition are extremely difficult for hearing-impaired children and adults with cochlear implants because the procedure and task demands require the listener to perceive and encode fine phonetic differences based entirely on information present in the speech signal without the aid of any external context or retrieval cues. The listener must then discriminate and select a unique pattern, a phonological representation, from a large number of equivalence classes in lexical memory (see Luce & Pisoni, 1998). This may seem like a simple task at first glance but it is very difficult for a child who has a cochlear implant. Children with normal hearing have very little difficulty with open-set tests like the PBK and they routinely display ceiling levels of performance in recognizing words under these presentation conditions (Kluck et al., 1997).

To learn more about the "Stars," we analyzed outcome data from pediatric cochlear implant users who scored exceptionally well on the PBK test two years after implantation. The PBK score was used as the "criterial variable" to identify and select two groups of subjects for subsequent analysis using an extreme groups design. After these subjects were selected and sorted into groups, we examined their performance on a variety of outcome measures already obtained from these children as part of our large-scale longitudinal study. The measures included tests of speech perception, comprehension, word recognition, receptive vocabulary knowledge, receptive and expressive language development and speech intelligibility.

Methods

Subjects

Scores for the two groups of pediatric cochlear implant users were obtained from a large longitudinal database containing a variety of demographic and outcome measures from 160 deaf children. Subjects in both groups were all prelingually deafened (mean =0.4 years onset). Each child received a cochlear implant because he/she was profoundly deaf and was unable to derive any benefit from conventional hearing aids. The criterion used to identify the "Stars" was based entirely on word recognition scores from the PBK test. This group consisted of 27 children who scored in the upper 20% on the PBK test two years post-implant. A "comparison" group of subjects consisting of 23 children who scored in the bottom 20% on the PBK test two years post-implant was also created for the analysis. The mean percentage of words correctly recognized on the PBK test was 25.64 for the "Stars" and 0.0 for the "comparison" group. A summary of the demographic characteristics of the two groups is shown in Table I.

No attempt was made to match the subjects on any other demographic variable other than length of implant use that was fixed at two years post implantation at the time these analyses were carried out. As a result of this selection procedure, the two groups turned out to be roughly comparable in terms of age of onset of deafness and length of implant use. However, as shown in Table I, the two groups did differ in terms of age at implantation, length of deprivation and communication mode.

For ease of exposition, the results of these analyses will be presented in three sections, one for the receptive scores and one for the expressive scores. The interrelations among these various measures using correlational methods are summarized in the last section.

Table I

Summary of Demographic Information

	Stars **(N = 27)**	***Controls*** **(N = 23)**
Mean Age at Onset (Years)	.3	.8
Mean Age at Implantation (FIT) (Years)	5.8	4.4
Mean Length of Deprivation (Years)	5.5	3.6
Mean Length of Implant Use (Years)	.9	1.0
Communication Mode:		
Oral Communication	N=19	N=8
Total Communication	N=8	N=15

Outcome Measures of Performance

Receptive Measures: Speech Perception and Spoken Word Recognition

Minimal Pairs Test. Measures of speech feature discrimination for both consonants and vowels were obtained for both groups of subjects with the Minimal Pairs Test (Robbins et al., 1988). This test uses a two-alternative forced-choice picture pointing procedure. The child hears a single word spoken in isolation on each trial by the examiner using live voice presentation and is required to select one of the pictures that correspond to the test item.

A summary of the consonant discrimination results for both groups of subjects is shown in Figure 1. Percent correct discrimination is displayed separately for the distinctive features of manner, voicing and place of articulation as a function of implant use in years. Data for the “Stars” are shown by the filled bars; data for the “Controls” are shown by the open bars in this figure. Chance performance on this task is 50 percent correct as shown by a dotted horizontal line. A second dotted horizontal line is also shown in this

figure at 70 percent correct corresponding to scores that are significantly above chance using the binominal distribution.

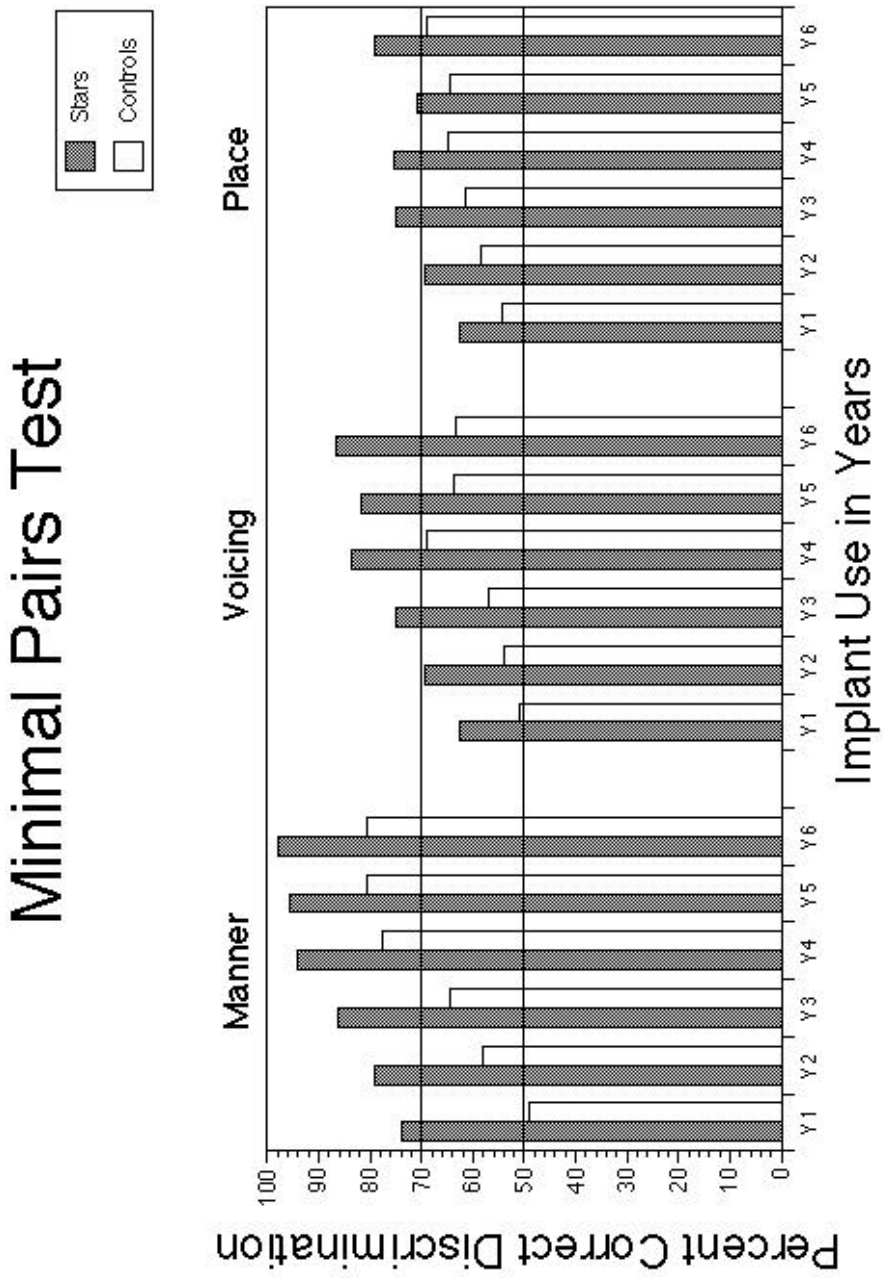


Figure 1. Percent correct discrimination on the minimal pairs (MPT) test for manner, voicing and place as a function of implant use. The "Stars" are shown by filled bars, the "Controls" are shown by shaded bars.

Inspection of the results for the Minimal Pairs Test obtained over a period of six years of implant use reveals several finding. First, performance of the "Stars" was consistently better than the control subjects for every comparison across all three consonant features. Second, discrimination performance

improved over time with implant use for both groups, although the increases were primarily due to improvements in discrimination of manner and voicing by the "Stars." All of the datapoints for the control subjects were at or below chance expectation for discrimination of the voicing and place features. Although there were increases in performance over time for the control subjects, their discrimination scores never reached the levels observed with the "Stars," even for the manner contrasts that eventually exceeded chance performance in years 4, 5 and 6.

The results of the Minimal Pairs Test indicate that both groups of children have difficulty perceiving, encoding and discriminating fine phonetic details of isolated spoken words even in a closed-set testing format. The "Stars" were able to discriminate differences in manner of articulation after one year of implant use and they showed consistent improvements in performance over time for both manner and voicing contrasts but they still had difficulty reliably discriminating differences in place of articulation, even after five years of experience with their implants. In contrast, the "Control" subjects were just barely able to discriminate differences in manner of articulation after four years of implant use and they still had serious problems with voicing and place of articulation even after five or six years of use.

The pattern of speech feature discrimination results shown here suggests that both groups of children are encoding spoken words using "coarse" phonetic representations that contain much less fine-grained acoustic-phonetic detail than normal hearing children typically do. The "Stars" are able to reliably discriminate manner and to some extent voicing much sooner after implantation than the "Controls" and the "Stars" display consistent improvements in speech feature discrimination over time. These speech feature discrimination skills are assumed to place initial constraints on the basic sensory information that can be used for subsequent word learning and lexical development. It is very likely that if a child cannot reliably discriminate differences between pairs of spoken words that are acoustically similar under these relatively easy forced-choice test conditions, they will subsequently have great difficulty recognizing words in isolation with no context or retrieving the phonological representations of these sound patterns from memory for use in simple speech production tasks such as imitation or immediate repetition.

Common Phrases Test. Spoken language comprehension performance was measured using the Common Phrases Test (Osberger et al., 1991). This is an open-set test that employs three presentation formats: auditory-only (CPA), visual-only (CPV) and combined auditory plus visual (CPAV). Children are asked questions or given commands to follow with instructions under these conditions. The results of this test are shown in Figure 2 for both groups of subjects, "Stars" and "Controls" as a function of implant use for the three different presentation formats. Inspection of this figure shows that the "Stars" performed consistently better than the "Controls" in all three presentation conditions and across all six years of implant use although performance begins to approach ceiling levels for both groups in the combined auditory plus visual conditions (CPAV) after five years of implant use. The multi-modal presentation conditions (CPAV) were always better than either the auditory-only or visual-only conditions. This pattern was observed for both groups of subjects. In addition, both groups displayed improvements in performance over time in all three-presentation conditions. Not surprisingly, the largest differences in performance between the two groups occurred in the auditory-only conditions. Even after three years of implant use, the "Control" subjects were barely able to perform this comprehension task above 25 percent correct when they had to rely entirely on auditory cues in the speech signal to carry out the task.

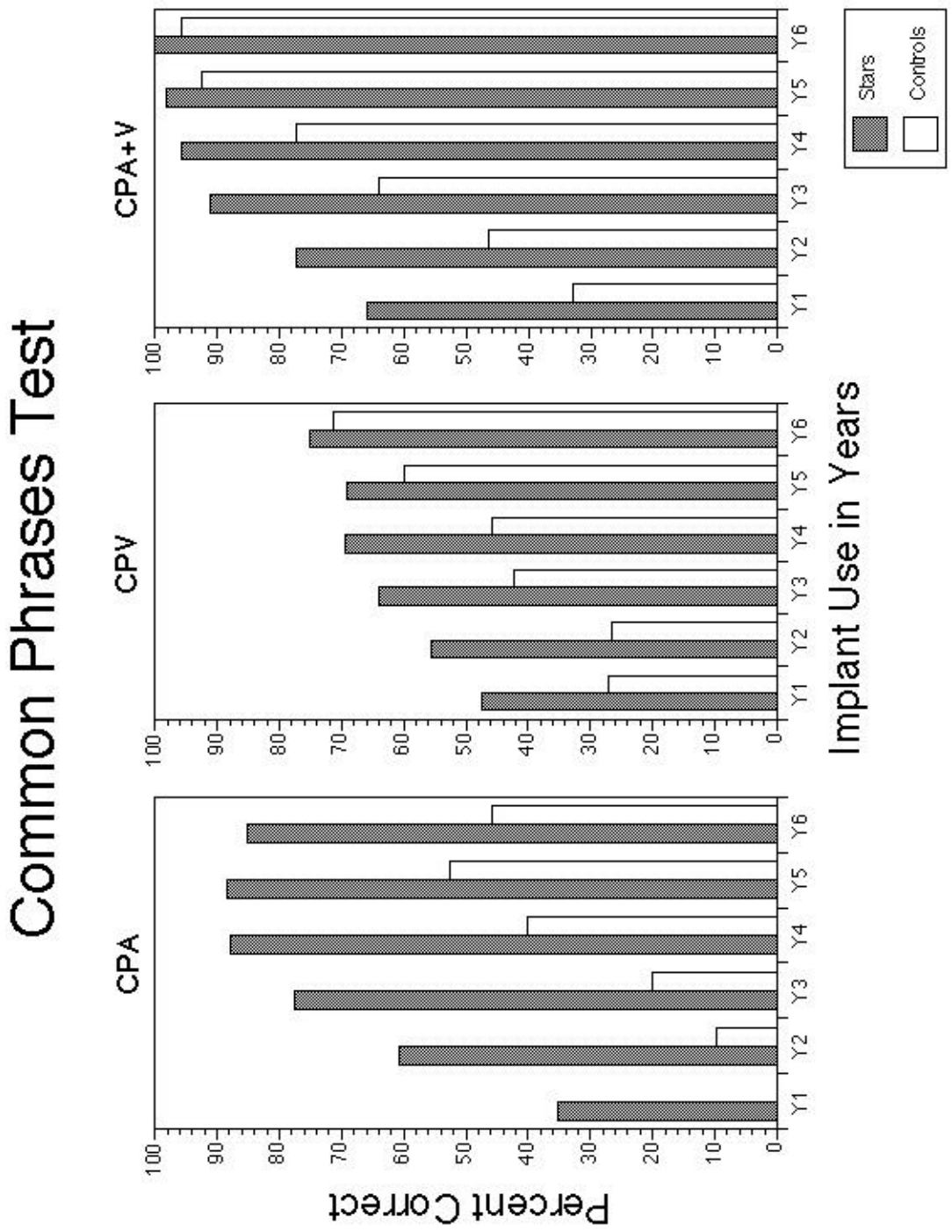


Figure 2. Percent correct performance on the common phrases (CPT) test for auditory-only (CPA), visual-only (CPV) and combined auditory plus visual presentation modes (CPA+V) as a function of implant use. The "Stars" are shown by filled bars, the "Controls" are shown by shaded bars.

Word Recognition Tests. Two new word recognition tests, the Lexical Neighborhood Test (LNT) and the Multi-syllabic Lexical Neighborhood test (MLNT), were used to measure open-set word recognition skills in both groups of subjects (Kirk, Pisoni & Osberger, 1995). Both tests use words that are familiar to preschool age children. The LNT contains short monosyllabic words, the MLNT contains longer polysyllabic words. Both of these tests use two different sets of items in order to measure lexical discrimination and provide details about how the lexical selection process is being carried out. Half of the items in each test consist of lexically "easy" words and half consist of lexically "hard" words. The differences in performance on these two sets of items in each test provide an index of how well a listener is able to make fine phonetic discriminations among acoustically similar words in their lexicons. Differences in performance between the LNT and the MLNT provide a measure of the extent to which the listener is able to make use of word length cues to recognize and access words from the mental lexicon. The items on both tests are presented in isolation one at a time by the examiner using auditory-only format. The child is required to imitate and immediately repeat back the test item after it is presented on each trial.

Figures 3 and 4 show the results, expressed as percent correct word recognition, obtained on the LNT and the MLNT for both groups of subjects as a function of implant use. The data for the "Stars" are shown in the top panel of each figure; the data for the "Controls" are shown in the bottom panels. Scores for the "easy" and "hard" words are shown within each panel. Several important differences in performance are shown in these two figures that provide some insights into the task demands and processing operations used in open-set tests. First, the "Stars" consistently demonstrate higher levels of word recognition performance on both the LNT and the MLNT than the "Controls." These differences are present across all six years but they are most prominent during the first three years after implantation. Word recognition scores for the "Controls" on both the LNT and the MLNT are very low and close to the floor compared to the performance observed for the "Stars" who are doing moderately well on this test although they never reached ceiling levels of performance on either the LNT or MLNT even after six years of implant use. Normal-hearing children typically display ceiling levels of performance on these same tests by age 4 (Kluck et al., 1997).

Another theoretically important finding is also shown in these figures. The "Stars" displayed a word length effect at each testing interval. Recognition was always better for the long words on the MLNT than the short words on the LNT. This pattern is obscured by a floor effect for the "Controls" who were unable to do this open-set task at all during the first three years. The presence of a word length effect for the "Stars" suggests that they are recognizing words "relationally" in the context of other words that they have in their lexicon (Luce & Pisoni, 1998). If these listeners were just recognizing words in isolation, feature by feature or segment-by-segment, without reference to words they already know and can access from lexical memory, one would predict that performance should be worse for longer words than shorter words because longer words simply contain more information. The pattern of findings observed here is exactly the opposite of this prediction and parallels earlier results obtained with normal-hearing adults and children (Luce & Pisoni, 1998; Kirk et al., 1995; Kluck et al., 1997). Longer words are easier to recognize than shorter words because they are more distinctive and less confusable with other phonetically similar words. The present findings suggest that the "Stars" are recognizing words based on their knowledge of other words in the language using processing strategies that are similar to those used by normal-hearing listeners.

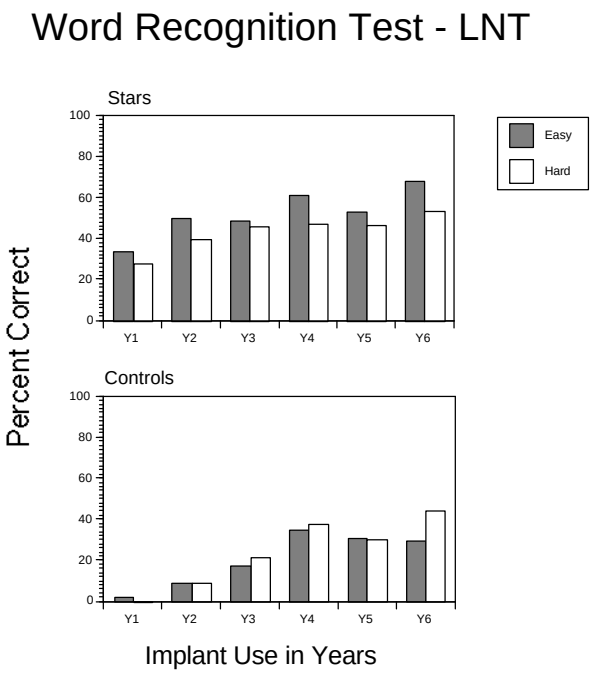


Figure 3. Percent correct word recognition performance for the LNT monosyllabic word lists as a function of implant use and lexical difficulty. "Easy Words" are shown by filled bars, "Hard Words" are shown by shaded bars. Data for the "Stars" are displayed in the top panel; "Controls" are displayed in the bottom panel.

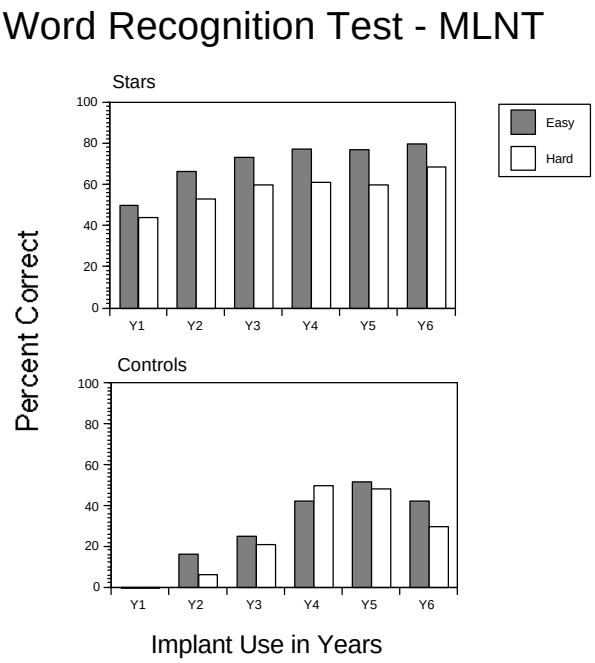


Figure 4. Percent correct word recognition performance for the MLNT multi-syllabic word lists as a function of implant use and lexical difficulty. "Easy Words" are shown by filled bars, "Hard Words" are shown by shaded bars. Data for the "Stars" are displayed in the top panel; "Controls" are displayed in the bottom panel.

Additional support for role of the lexicon and the use of lexical knowledge in open-set word recognition is also provided by another finding shown in both figures. The "Stars" also showed a consistent effect of "lexical discrimination" for both sets of words on the LNT and the MLNT tests. Examination of Figures 3 and 4 reveals that the "Stars" recognize lexically "easy" words better than lexically "hard" words. The difference in performance between "easy" words and "hard" words is present for both the LNT and the MLNT vocabularies but it is larger and more consistent over time for the words on the MLNT test. Because of floor effects, the "Controls" did not display the same consistent pattern of performance or sensitivity to lexical competition among the test words.

The differences observed between these two groups of children on both open-set word recognition tests are not at all surprising and were expected because these two extreme groups were initially created based on their PBK scores. But the overall pattern of the results is theoretically important at this time because the findings obtained with these two new open-set word recognition tests, the LNT and the MLNT, demonstrate that the skills and abilities used to recognize isolated spoken words are not specific to the test items used on the PBK test or the experimental procedures used in open set tests of spoken word recognition. The original differences between the two groups readily generalized to other open-set word recognition tests using different words. This pattern of results strongly suggests the operation and use of some common underlying set of cognitive and linguistic processes that are employed in recognizing, imitating and immediately repeating back spoken words presented in isolation. As suggested below, identifying and understanding the processing mechanisms that are used in these kinds of tasks may provide some new insights into the underlying basis of the large individual differences observed in outcome measures in children with cochlear implants. It is probably no accident that the PBK test has had some important diagnostic utility in identifying the exceptionally good users of cochlear implants over the years (see Kirk et al., 1995; Meyer & Pisoni, 1999). The PBK test is clearly measuring several important language processing skills that may generalize well beyond the specific repetition task used in open-set tests. The most important conceptual issue now is to explain why this happens to be the case and to begin to identify the underlying cognitive and linguistic mechanisms that are being used in open-set word recognition tasks as well as other tasks that draw on the same set of processing resources and operations.

Receptive Vocabulary Knowledge. Vocabulary knowledge was assessed using the Peabody Picture Vocabulary Test (PPVT), a standardized test that provides a measure of receptive language development based on word knowledge (Dunn & Dunn, 1997). Test items were presented using the child's preferred mode of communication, either speech or sign, depending on whether the child is immersed in an Oral-only (OC) or Total-Communication (TC) environment. The scores on the PPVT are shown in Figure 5 for both groups of subjects as a function of implant use. The top panel of this figure shows the raw scores; the bottom panel shows the same data expressed as language quotients that were obtained by dividing the child's language age by his/her chronological age. Language age is based on norms from normal-hearing children. Normal-hearing children with typical age-appropriate language skills would be expected to achieve scores of 1.0 on this scale. Inspection of the top panel shows that when expressed in terms of raw scores, both groups improve over time with implant use. However, the "Stars" score consistently better than the "Controls," although the differences are not as large as those observed on the previous word recognition tests. This pattern may be due to the fact that this test as well as the other standardized language tests are routinely administered in the child's preferred communication mode. As

displayed in Table I, most of the "Controls" included in this group were enrolled in TC programs whereas most of the "Stars" were enrolled in OC programs. Examination of the language quotients shown in the bottom panel indicates that both groups of children display comparable scores that remain the same over time. This is not surprising because chronological age was used to normalize the language scores.

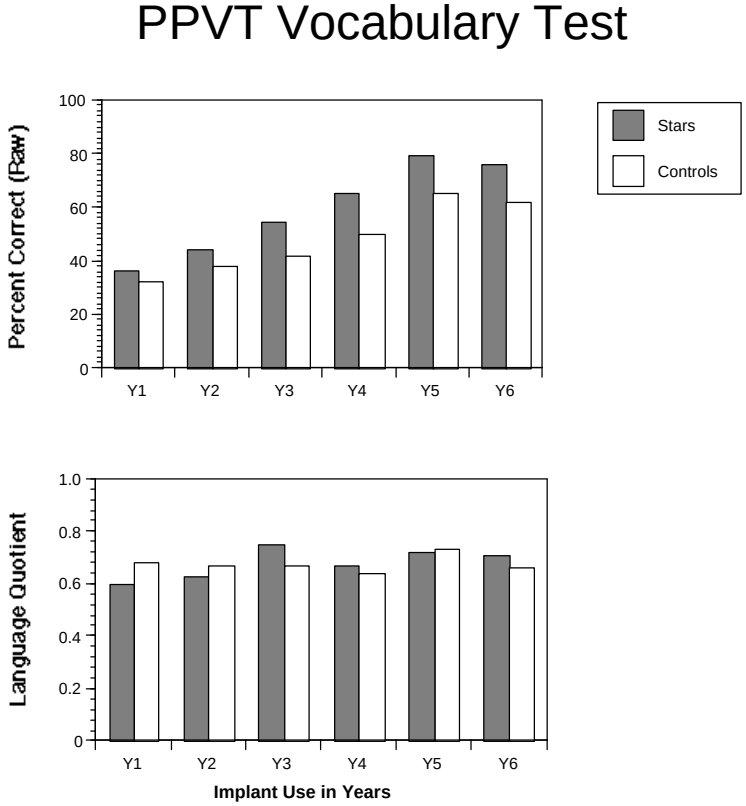


Figure 5. Raw scores (top panel) and language quotients (bottom panel) for the "Stars" and "Controls" on the Peabody Picture Vocabulary Test (PPVT) as a function of implant use in years.

Measures of Language Development

Measures of receptive and expressive language development were obtained for both groups of subjects using the Reynell Developmental Language Scales (Reynell & Huntley, 1985). These scales assess receptive and expressive language skills independently using tasks involving object manipulation and description based on questions that vary in length and linguistic complexity. The Reynell tests have been used extensively with deaf children and are appropriate for a broad age range of children from one to eight years old. Normative data have also been collected on normal-hearing children so appropriate comparisons can be drawn (see Svirsky et al., in press).

Figures 6 and 7 show scores for the Reynell receptive and expressive scales for both groups of children as a function of implant use. The top panel in each figure shows the raw scores for each measure, the bottom panel shows the corresponding language quotients. Scores for the "Stars" in Y6 in each figure are based on projected estimates because these children had reached ceiling levels of performance for their age and the tests were no longer routinely administered at the yearly assessments. Both sets of data for the Reynell show gradual improvement in language over time. Once again, the differences in performance between the two groups were not very large, although the "Stars" achieved higher scores on both receptive

and expressive scales than the controls. In our earlier analyses of the performance of the "Stars," after the first three years of implantation (Pisoni et al., 1997), we found a main effect for communication mode and an interaction of communication mode with group. Overall, TC children scored higher than OC children but this was observed only for the "Control" subjects and not the "Stars."

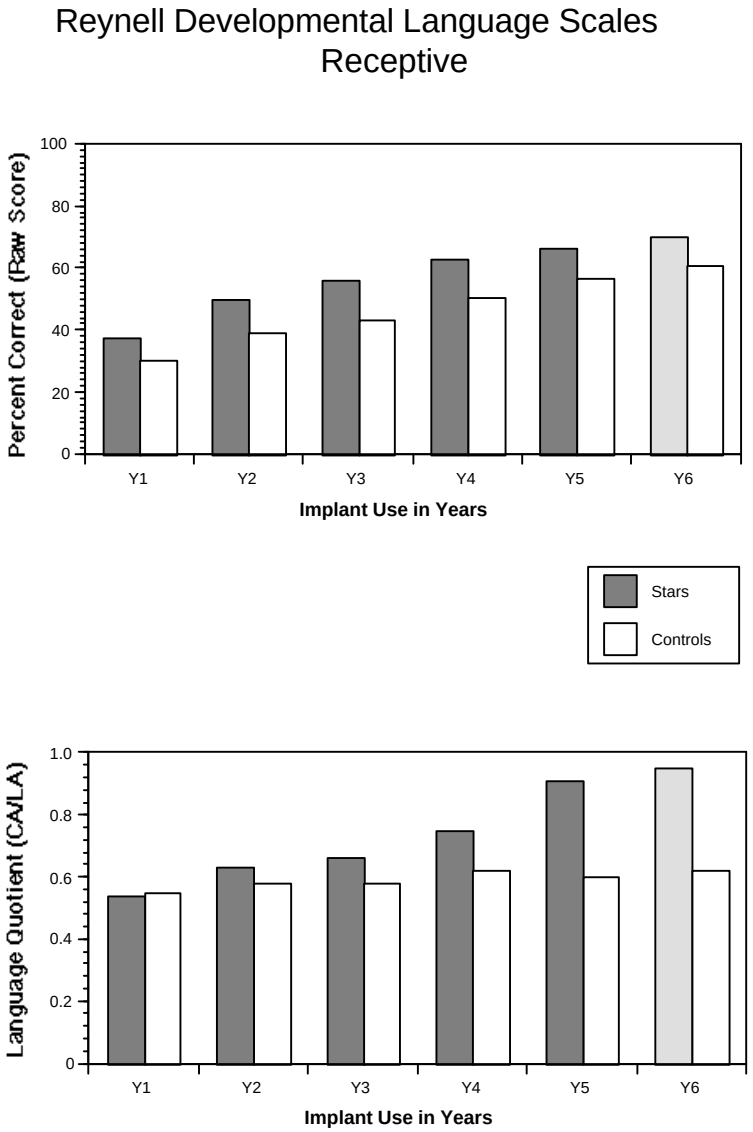


Figure 6. Reynell receptive language scores for "Stars" and "Controls" as a function of implant use. Top panel shows raw scores, bottom panel shows language quotients.

Taken together with the earlier PPVT scores, the present results suggest that communication mode does influence outcome measures on standardized tests that assess language and language-related abilities such as vocabulary knowledge and language use. It is clear that the specific types of social and linguistic interactions that take place in the child's language learning environment after implantation play an important role in promoting and facilitating language development, vocabulary acquisition and overall success with a cochlear implant. Children with cochlear implants who are placed in OC environments consistently show large gains in oral language skills on tasks that specifically require the use of

phonological representations and phonological processing strategies in speech perception and speech production tasks

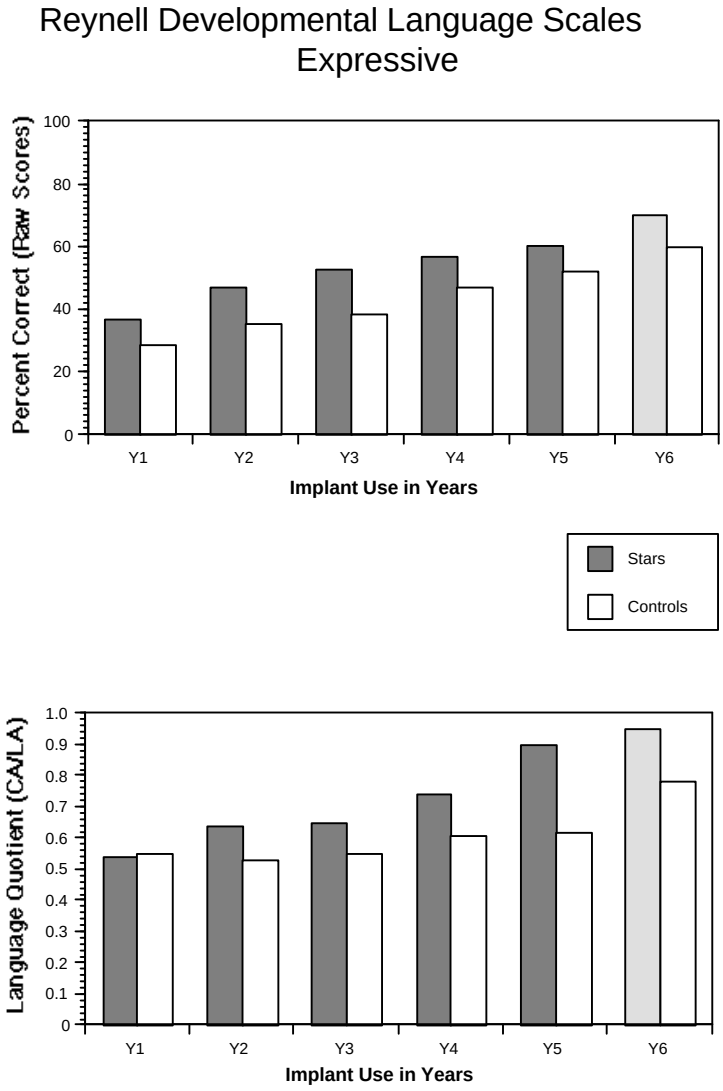


Figure 7. Reynell expressive language scores for "Stars" and "Controls" as a function of implant use. Top panel shows raw scores, bottom panel shows language quotients.

Speech Intelligibility

Measures of speech intelligibility were also obtained for both groups of subjects using a transcription task. Speech samples were first obtained from each child using standardized elicitation materials. Each child produced 10 sentences that were repeated after an examiner's spoken model. One list from the Beginners Intelligibility Test (BIT) was administered to obtain the speech samples from each child. This test uses objects and pictures to convey the target sentence (Osberger et al., 1994). The speech

samples were then played back to small groups of normal-hearing adult listeners who were asked to listen and transcribe what the child had said. A composite score based on the number of words correctly transcribed for each child was obtained from the responses provided by three listeners who heard each child's utterance.

Figure 8 shows the percent correct transcription for the "Stars" and "Controls" as a function of implant use. Examination of this figure shows that the "Stars" display much better speech intelligibility than the "Controls." Although both groups show improvements in speech intelligibility over time, the difference in performance between the "Stars" and "Controls" remains roughly constant even after six years of implant use. The differences in speech intelligibility found here demonstrate that variation in performance between the "Stars" and "Controls" is not restricted to only receptive measures of language processing such as speech perception, spoken word recognition, receptive vocabulary knowledge or comprehension. The present findings on speech intelligibility provide evidence for transfer of knowledge from one linguistic domain to another and suggest an overlap and commonality between perception and production (see O'Donoghue et al., 1999). This overlap of receptive and expressive language function reflects a knowledge of the sound/meaning contrasts in the language and a common underlying linguistic system, a grammar, that the child constructs from the linguistic input he/she is exposed to in the ambient environment. As we observed earlier, the "Stars" showed large and consistent improvements in both receptive and expressive measures of language including speech feature perception, spoken word recognition, vocabulary knowledge, comprehension and speech intelligibility. In contrast, the "Control" subjects not only showed much lower levels of performance overall on these tests but the rate of their improvement in performance was much slower over time.

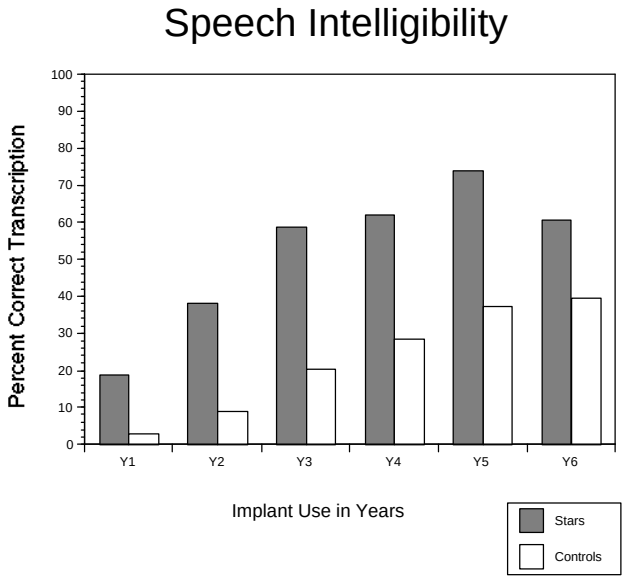


Figure 8. Percent correct transcription scores for "Stars" and "Controls" as a function of implant use in years.

Correlations Among Test Measures

Examination of these descriptive results show that the exceptionally good children, the "Stars," appear to do well on a wide variety of outcome measures including speech feature perception,

comprehension, spoken word recognition, receptive and expressive language as well as measures of speech intelligibility. This pattern of findings was very encouraging because it suggested that there may be some common source of variance that underlies the exceptionally good performance of these children on many different outcome measures. Our working hypothesis is that this particular source of variance reflects "modality-specific" fundamental information processing operations that are involved in the phonological coding of sensory inputs and the construction of phonological representations of speech (see Pisoni et al., 1997).

Until our analyses of these scores from the "Stars," very little previous research was directed specifically at the study of individual differences among pediatric cochlear implant users or an examination of the perceptual, cognitive and linguistic abilities of the exceptionally good subjects. Our analyses of speech perception, word recognition, spoken language comprehension, vocabulary knowledge and language development demonstrate that a child who displays exceptionally good performance on the PBK test also shows very good scores on a variety of other speech and language measures as well. This is a theoretically important finding. The differences in performance observed here between the "Stars" and "Controls" are of substantial interest because it may now be possible to determine precisely how and why the "Stars" differ from other less successful cochlear implant users. If we have knowledge of the factors that are responsible for individual differences in performance among deaf children who receive cochlear implants, particularly the variables that underlie the extraordinarily good performance of the "Stars," we may be able to help those children who are not doing as well with their implant at an early point in development. Moreover, our findings on individual differences may have direct clinical relevance in terms of recommending specific changes in the child's language-learning environment and in modifying the nature of the sensory inputs and linguistic interactions a child has with his/her parents, teachers and speech therapists who provide the primary language model for the child. Our findings on individual differences may also help in providing clinicians and parents with a principled basis for generating realistic expectations about outcome measures, particularly measures of speech perception, comprehension, language development and speech intelligibility in deaf children with cochlear implants.

One of the most interesting and informative analyses that we carried out on these data was a series of simple correlations among the different dependent measures summarized above. We were interested in the following questions: Does a child who performs exceptionally well on the PBK test also perform exceptionally well on other tests of speech feature discrimination, word recognition and language? What is the relationship between performance on the PBK test and speech intelligibility? Is the extraordinarily good performance of the "Stars" restricted to only open-set word recognition tests like the PBK or is it possible to identify a common underlying variable or process that can account for the relationships observed among several different dependent measures? In order to answer these kinds of questions, we examined the intercorrelations for each of the dependent variables described earlier. Simple bivariate correlations were carried out separately for the "Stars" and "Controls" using the test scores obtained after only one year of implant use. A detailed summary of these findings was reported in Pisoni et al., 1997. In this section, we present the correlations for three of the dependent measures, open-set word recognition using the LNT, receptive and expressive language based on the Reynell and speech intelligibility scores using measures of transcription in order to illustrate the general pattern that was found across the other dependent measures. More details are provided in the earlier report.

Open-set Word Recognition. Table II shows the correlations of the LNT word recognition scores with each of the other dependent measures for the "Stars." The correlations for the lexically "easy" words are shown in the left-hand column; the correlations for the lexically "hard" words are shown in the right-hand column. Because the "Control" subjects were unable to recognize any of the words on the LNT after

one year of implant use, it was not possible to compute any correlations with the other test measures. An examination of this table shows that performance on the LNT is highly correlated with comprehension scores, receptive vocabulary knowledge, both receptive and expressive measures of language development and speech intelligibility scores. The pattern of intercorrelations among these dependent measures strongly suggests a common underlying source of variance that is shared by all these different tasks. The extremely high correlations of the LNT word recognition scores with the Common Phrases-Auditory-only scores and both language measures on the Reynell suggests that this common source of variance may be related in some way to the encoding, storage, retrieval and rehearsal of spoken words, specifically, the phonological representations of spoken words in lexical memory. The fundamental cognitive and linguistic processes used to recognize spoken words in an open-set format like the PBK or LNT test where there is no context other than the acoustic-phonetic information in the signal are probably also used in other language processing tasks as well such as comprehension and speech production which draw on various sources of information about spoken words in the lexicon.

Table II

CORRELATIONS: WORD RECOGNITION - YEAR 1

	Lexical Neighborhood Test *(LNT)*			
	Easy Words		***Hard Words***	
	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	*r* =	*r* =	*r* =	*r* =
Minimal Pairs-Manner	.34	----	.51	----
Minimal Pairs-Voicing	.20	----	.58	----
Minimal Pairs-Place	.16	----	-.06	----
COMPREHENSION:				
Common Phrases-Auditory only	.81***	----	.85***	----
Common Phrases-Visual-only	.41	----	.57	----
Common Phrases-Auditory+Visual	.42	----	.55	----
VOCABULARY:				
PPVT-R	.62*	----	.63*	----
LANGUAGE:				
Reynell Receptive Language Quotient	.86***	----	.81**	----
Reynell Expressive Language Quotient	.83***	----	.82**	----
SPEECH INTELLIGIBILITY:				
Transcription	.89**	----	.80**	----

* $p < .05$; ** $p < .01$; *** $p < .001$

Reynell Language Scales. The correlations obtained for the receptive and expressive scales of the Reynell and the other dependent measures are shown in Table III for the "Stars" and the "Controls." Once

again, a systematic pattern of intercorrelations can be observed among almost all of the test scores for the "Stars." These correlations are extremely high and statistically significant given the relatively small sample sizes used here. The strong correlations of both the Reynell receptive and expressive scores with the open-set word recognition scores on the LNT suggest a common underlying factor that is related, in some way, to spoken word recognition and lexical access. The correlations between the language scores and speech intelligibility may reflect a common or shared representational system and a set of phonological processing skills that are used in both receptive and expressive language processing tasks.

TABLE III

CORRELATIONS: LANGUAGE - YEAR 1

	Reynell Language Scales (Language Quotient)			
	Receptive		***Expressive***	
	Stars	Controls	Stars	Controls
SPEECH PERCEPTION:	*r* =	*r* =	*r* =	*r* =
Minimal Pairs-Manner	.77**	.08	.78**	-.28
Minimal Pairs-Voicing	.69*	-.63	.61*	-.49
Minimal Pairs-Place	.20	-.01	.31	.33
COMPREHENSION:				
Common Phrases-Auditory	.82**	----	.85***	----
Common Phrases-Visual-only	.64*	----	.79**	----
Common Phrases-Auditory+Visual	.64*	.33	.67*	.36
WORD RECOGNITION:				
LNT-Easy words	.86***	----	.83***	----
LNT-Hard words	.81**	----	.82**	----
MLNT-Easy words	.84**	----	.87***	----
MLNT-Hard words	.66*	----	.76	----
VOCABULARY:				
PPVT-R	.81***	.69**	.68**	.56*
SPEECH INTELLIGIBILITY:				
Transcription	.80**	-.39	.85**	-.13

* $p < .05$; ** $p < .01$; *** $p < .001$

Speech Intelligibility. The correlations between the speech intelligibility scores and the other dependent measures are shown in Table IV separately for the "Stars" and "Controls." Examination of this table also shows once again a pattern of correlations that is very similar to those observed in the previous two tables. Speech intelligibility is highly correlated with language comprehension, spoken word recognition and language development suggesting a common underlying source of variance (see also

O'Donoghue et al., 1999 for recent findings on the relationship between speech perception and production in young children with cochlear implants).

The results of the present set of analyses suggest several hypotheses about the source of the differences in performance between the "Stars" and the "Controls." We believe these accounts are worth pursuing and evaluating in much greater depth because they suggest new and unexplored areas of basic and clinical research on pediatric cochlear implant users. Our working hypothesis places the locus of the differences in performance between the "Stars" and "Controls" at central rather than peripheral processes. This account of the source of the individual differences focuses on how the initial sensory information is encoded, stored, retrieved and manipulated in various kinds of information processing tasks such as speech feature discrimination, spoken word recognition, language comprehension and speech production. The emphasis here is on higher-level perceptual and cognitive factors that play a critical role in how the sensory, perceptual and linguistic information input is processed, organized and used in various psychological tasks. One of the key components that link these various processes and operations together and serves as the "interface" between the initial sensory input and stored knowledge in memory is the working memory system. The properties of this particular memory system may provide further insights into the nature and locus of the individual differences observed among users of cochlear implants (see Carpenter, Miyake & Just, 1994; Baddeley, Gathercole, & Papagno, 1998; Gupta & MacWhinney, 1997). Unfortunately, at the time these analyses were carried out, we did not have any memory data from the "Stars" and "Controls" to test this proposal, but several new studies have been carried out recently using new measures of performance and these results are reported in the sections below.

TABLE IV

CORRELATIONS: SPEECH INTELLIGIBILITY - YEAR 1

	Transcription Scores	
	Stars	**Controls**
SPEECH PERCEPTION:	*r* =	*r* =
Minimal Pairs-Manner	.55	.19
Minimal Pairs-Voicing	.53	-.11
Minimal Pairs-Place	.41	-.09
COMPREHENSION:		
Common Phrases-Auditory	.65**	.04
Common Phrases-Visual-only	.87**	.25
Common Phrases-Auditory+Visual	.43	.07
WORD RECOGNITION:		
LNT-Easy Words	.89**	----
LNT-Hard Words	.80*	----
MLNT-Easy Words	.87**	----
MLNT-Hard Words	.72	----
VOCABULARY:		
PPVT-R	.45	-.01
LANGUAGE:		
Reynell Receptive Language Quotient	.80**	-.39
Reynell Expressive Language Quotient	.85**	-.13

* p <.05; ** p <.01; *** p <.001

Some New Process Measures of Performance

It is very easy to say that children who "hear" better through their cochlear implant just learn language better and subsequently recognize words better. But it is much more difficult to explain the observed differences in speech intelligibility on the basis of better hearing and language skills without a more detailed description of exactly what these underlying skills and abilities are and what specific cognitive processes they draw on. To account for the differences in speech intelligibility performance and expressive language, it is necessary to assume some underlying linguistic structure and process that mediates between speech perception and speech production. Without access to and use of a common underlying linguistic system—a "grammar," separate receptive and expressive language abilities and skills such as these would not be so closely coordinated and mutually dependent. Reciprocal links exist between speech perception, production and a whole range of language-related abilities and these links reflect the child's linguistic knowledge of phonology, morphology and syntax. Speech perception, spoken word recognition and language comprehension are not isolated autonomous perceptual abilities or skills that are independent of language and the child's developing linguistic system. The same observation is true for speech production, reading and lip-reading. An account framed in terms of hearing, audibility or sensory discrimination abilities cannot provide a satisfactory explanation of all of the results or an adequate description of the "process" of how early auditory experience affects speech perception and language development in these children. Something else underlies the commonalities observed across these diverse tasks.

In order to provide a unified account of these findings, it is necessary to obtain additional performance measures that assess how deaf children with cochlear implants actually "process" and "code" the sensory, perceptual and linguistic information they receive through their implants and how they store, retrieve and use this information in a variety of information processing tasks. The outcome measures in our database were scores on traditional standardized tests that were used for assessment of specific speech and language skills thought to be important for measuring change and success after implantation. The battery of these tests was designed and constructed many years ago when theoretical issues about individual differences and underlying processing strategies were not an important research priority. As a result, there are no data available on psychological/cognitive processes such as memory, learning, attention, automaticity or modes of processing. These are new topics that need to be studied in greater detail. We also need to learn more about the role of early auditory experience on perceptual and cognitive development, especially spoken word recognition, lexical development, language comprehension and speech intelligibility. These have also not been an important priority in earlier research on cochlear implants in children.

If we were going to look at process measures, that is, measures of what a child does with the sensory information he/she receives through the CI, where would we look first? There are several different areas we could explore: perception, attention, learning and memory. And, there are many different techniques and procedures we could use. For a variety of theoretical reasons, we selected "working memory" because this is known to be a very important component of the human information processing system and serves as the interface between sensory input and stored knowledge in long-term memory. Working memory has also been shown to be the source of individual differences observed across a wide range of domains from perception to memory to language (Ackerman, Kyllonen & Roberts, 1999). To obtain some initial measures of working memory from children with cochlear implants, we began

collaborating with Ann Geers and her research group at the Central Institute for the Deaf (CID) in St. Louis. The children in the CID study described in the next section are older than the children I reported on earlier. The children at CID were 8 and 9 years old and they all used their cochlear implant for at least 5 years. Thus, chronological age and implant use were controlled in this study.

Working Memory Span

Methods

Subjects

Forty-three 8 and 9 year old cochlear implant users were recruited for this study from a much larger on-going project conducted at Central Institute for the Deaf (CID) in St. Louis, Missouri. All of these children had used their implant for at least five years before testing was carried out.

Procedure

In addition to the auditory digit span measures that were collected specifically for this study, the children also received an extensive battery of speech, language and reading tests that were part of the original large-scale project. The forward and backward digit spans were obtained using the digit-span subtests of the WISC-III (Weschler, 1991). The forward span task requires the child to repeat back a list of digits in the order in which the sequence was presented. In the backward span task, the child is instructed to say the list of digits backwards. Digit spans were obtained using live voice presentation with lip-reading cues available. In both parts of the WISC digit span task, the lists began with two items and increased in length until the child recalled two lists at a given length incorrectly at which point the procedure was terminated. Items were not repeated within any list and each list of digits was unique. The child was required to recall all of the digits in a given list consecutively in the correct temporal order in order to receive full credit for a given list length. Each child was run individually.

Results

Several measures of memory span performance were obtained from the response protocols. For the present analysis, digit span was defined as the longest length sequence of digits that the child could recall correctly two times in a row preserving both item and order information. This dependent measure was used in all of the analyses reported below. The forward digit spans ranged from zero to eight items correct with a mean span length of 5.3 items averaged over 43 subjects. Only one child failed to carry out the digit span task and his data were not included in any of the final analyses.

TABLE V

CORRELATIONS: SPEECH PERCEPTION
(from Pisoni & Geers, 1998)

	Forward Auditory Digit Span (N=43)
	r
SPOKEN WORD RECOGNITION:	

WIPI	**+.71**
LNT	**+.64**
BKB	**+.59**
AUDITORY+VISUAL:	
Chive V (lip-reading)	**+.52**
Chive VE (visual enhancement)	**+.66**
SPEECH FEATURE DISCRIMINATION:	
VIDSPAC	**+.59**

Table V shows the correlations of the forward auditory digit spans with several measures of speech perception performance that were obtained from these children as part of the larger project at CID. These measures included scores on both closed-set (WIPI) and open-set (LNT) word recognition tests, a sentence perception test (BKB), tests of auditory-visual integration (Chive) as well as speech feature discrimination (VIDSPAC). The correlations are all positive and generally moderate to quite strong, suggesting a common underlying source of variance. In interpreting these simple first-order correlations, it is possible to account for the memory span results in terms of purely sensory factors like audibility and basic speech discrimination skills that propagate and cascade up the processing system. According to this account, children who display longer digit spans simply perceive speech better and have more detailed and robust sensory representations of the speech waveforms than the children who have shorter digit spans.

In order to assess this explanation, a series of partial correlations were also computed using performance on the VIDSPAC, a test of speech feature discrimination, as a measure of overall speech discrimination performance. When the variance due to speech feature discrimination was partialed out, the correlations were reduced in size but they were still statistically significant suggesting that the results were not due to audibility or basic sensory discrimination skills but were related to the way in which the initial sensory information is processed, encoded and retrieved from memory. Processing differences among these children may reflect fundamental limitations on the capacity of working memory in terms of the speed and efficiency that sensory information can be encoded using a cochlear implant. These differences in information processing may affect the initial encoding, rehearsal and the scanning of information in working memory. The pattern of correlations also suggests the presence of a source of variance in these tests that is associated in some way with processing operations—that is, what the child does with the initial sensory information he/she received through the cochlear implant.

Two other theoretically important findings were also obtained in this study of working memory using the WISC digit span task. As part of the larger project at CID, measures of speech intelligibility were also obtained from these children using the elicitation materials developed by McGarr (1981). Using methods that were similar to the intelligibility study described earlier, utterances were obtained from each child and played back to normal-hearing adults who were asked to transcribe the sentences. The transcription scores based on words correctly recognized from three listeners were pooled for each child and a composite measure of speech intelligibility was obtained. Through the kind cooperation of Professor Emily Tobey, we were able to obtain these intelligibility scores along with measures of the sentence durations. We then computed a correlation between the auditory digit spans from the WISC and the speech intelligibility scores for these children. A scatterplot of the individual subjects is shown in the top panel of Figure 9. The WISC digit span is represented on the ordinate and the McGarr Intelligibility score is represented on the abscissa. Examination of this figure shows a very orderly relationship between these two

measures. Subjects with longer digit spans tend to display higher levels of speech intelligibility. The correlation was r =+.69, (p< .001), suggesting a strong association between working memory span and processes used in speech production. This particular correlation is especially important because it suggests a reciprocal relationship between speech perception and speech production and implies that the two processes are closely linked and draw on a common set of processing resources that are related to the retrieval and maintenance of phonological representations of spoken words in working memory.

In addition to the speech intelligibility scores, duration measurements were obtained for the McGarr sentences from each child and these data were also analyzed. Correlations were carried out between the WISC digit span measures and the average sentence durations for these materials produced by each child. A scatterplot of the individual subject data is shown in the bottom panel of Figure 9. The WISC digit span is shown on the ordinate; the average sentence duration is shown on the abscissa. Once again, we see a very orderly and systematic relationship between working memory span and sentence duration. Subjects who display longer digit spans tend to produce sentences with shorter overall durations. The correlation between these two variables was r = -.64, (p < .001). This finding suggests that children who speak faster may have a faster rehearsal speeds in working memory and this may be the reason why these subjects are able to recall longer sequences of digits.

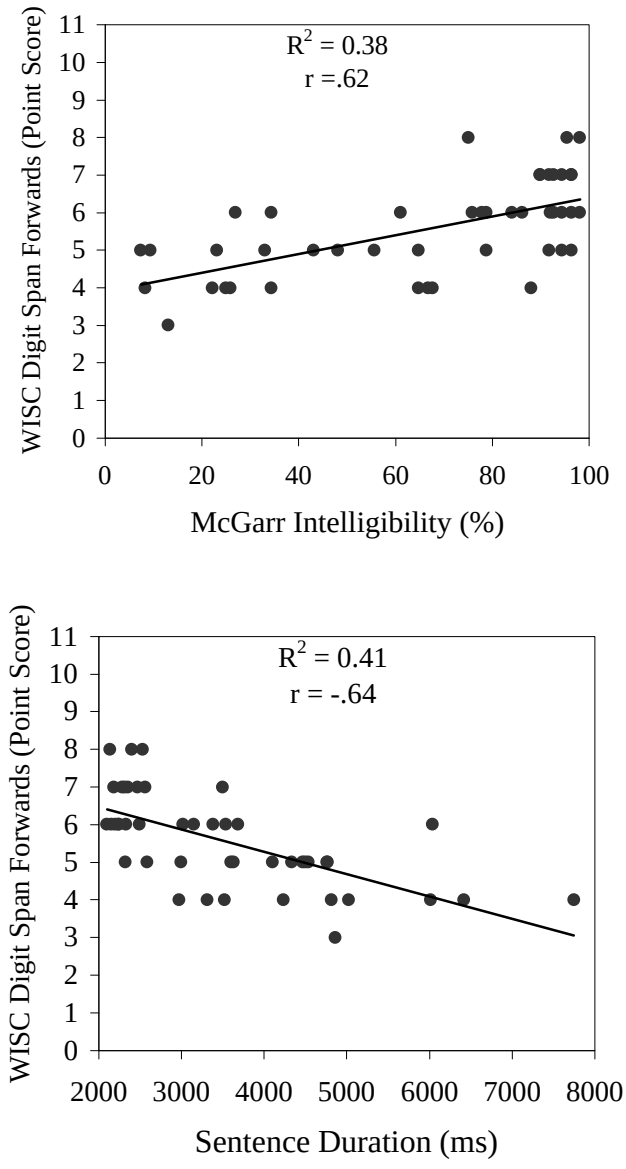


Figure 9. Top panel shows a scatterplot of WISC forward digit spans in points as a function of speech intelligibility; the bottom panel shows the WISC forward digit spans in points as a function of sentence duration in ms.

This finding is consistent with a large body of earlier research on verbal short-term memory which demonstrates a close relation between memory span and the fastest rate at which a person con pronounce a short list of words (Baddeley, Thomson & Buchanan, 1975; Chase, 1977; Schweickert & Boruff, 1986). Several recent studies have suggested that this global information processing rate may actually reflect the combined effects of two independent processes, one related to speed of articulation and the other related to

the retrieval of words from short-term memory (Schweickert et al., 1990; Cowan et al., 1998). Cowan et al. (1998) suggest that while both of these processes affect memory span, they are actually independent of each other. Thus, memory span may depend on several component processes that are occurring simultaneously. Without additional data, it is not possible to dissociate the contribution of these two effects in the present analyses but the results clearly demonstrate a strong relation between digit span and a specific processing mechanism related to the rate at which information is encoded, rehearsed and subsequently output in an immediate serial recall task. These findings provide additional converging support for the proposal that the variation in the underlying processes may account for the large differences in outcome performance on a wide range of audiological tests.

The correlations between WISC digit span and the four sets of outcome measures obtained from these children demonstrate very clearly that the working memory component of the human information processing system is involved in some way in mediating, modulating and controlling performance across a wide range of language related tasks like speech perception, speech production, spoken word recognition, language comprehension and reading. Thus, processes related to the encoding, rehearsal and short-term storage of spoken words appear to play an important role in the component underlying abilities and skills that are actually being measured by the four different language-related outcome measures.

The correlations observed between the WISC digit spans and the two measures obtained form the speech production task, the speech intelligibility scores and the sentence durations, provided some extremely valuable new information about the specific processing mechanism that may be responsible for the differences in working memory capacity and the correlations found with other language processing measures. Both findings suggest that the rehearsal process may be the locus of the individual difference observed in these children. Although these are only correlational data, and must be interpreted cautiously, the relationship between WISC digit span and rehearsal speed suggests a number of new research directions to pursue in order to test this hypothesis more directly. The next experiment was designed to investigate coding and rehearsal strategies using a new procedure to measure sequence memory.

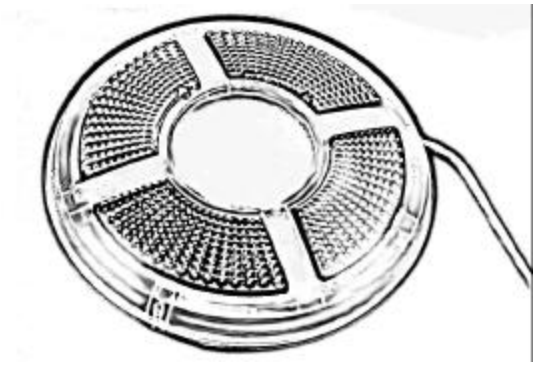


Figure 10. Schematic picture of the modified Simon ™ memory box game originally manufactured by Milton Bradley.

Coding and Rehearsal Processes in Working Memory

The findings obtained using the WISC auditory digit span as a measure of short-term memory capacity were very informative and suggested that some processing variable related to working memory

may underlie the large individual differences in outcome measures observed in children with cochlear implants. Gaining a detailed understanding how young children encode and manipulate the phonological representations of spoken words may provide further insights into the development of their spoken language abilities and skills and may help to explain the underlying basis of the variability in performance in terms of information processing variables.

Because some children with cochlear implants may have difficulty producing intelligible speech due to differences in speaking rate and fluency of articulation, it was necessary to find a procedure that did not require the child to produce an explicit verbal output response. In order to meet this need, we recently developed a new procedure to measure working memory span that was modeled after the popular electronic game called "Simon" developed by Milton Bradley. The apparatus has four large colored response buttons and is shown in Figure 10. Children are presented with a sequence of sounds in conjunction with a sequence of colored lights and are asked to immediately reproduce the sequence in the order in which it was originally presented by depressing the appropriate response buttons.

The Simon memory game apparatus and experimental procedures we developed have a number of useful attributes that were explored in the experiment described below. First, the difficulty of the task can be adjusted by simply increasing the length of the sequence to be reproduced. Second, using an adaptive testing procedure running under computer control, it is possible to quickly locate the longest sequence that a child can reproduce under a given test condition and use that value as a measure of the child's memory span. Finally, it is possible, as shown below, to manipulate both the visual and auditory stimuli separately or in combination and the contingencies between them at the time a sequence is presented. This particular feature permitted us to study how redundancies between visual and auditory cues are perceived, encoded and processed and how correlations between these two stimulus dimensions would affect memory span for reproducing sequences of sounds or sequences of lights or sequences of sounds and lights together. In the experiment described below, we obtained reproductive spans using the Simon memory game under three presentation conditions that manipulated the redundancies across dimensions. In the first condition, whenever one of the four colored lights was illuminated on the display, the color-name of the response button was also simultaneously output as an auditory signal. In the second condition, whenever a colored light was illuminated, a digit-name was output as an auditory signal. The digits were always consistently mapped to a response button and remained invariant for a individual child. Finally, in the third condition, visual patterns were generated using only the lights with no auditory stimuli. This last condition, light-only presentation, served as an important control condition to assess the extent to which a child would be able to take advantage of the cross-correlations between stimulus dimensions presented in two different modalities. In earlier research using these procedures with normal-hearing adults and children, Cleary (1997) and Carlson et al. (1998) observed a "redundancy gain" when the lights and sounds were both correlated together. Reproduction spans increased in length for the combined (A+V) condition compared to either auditory-only or visual-only presentation.

Methods

Subjects

Two groups of 45 children were recruited for this study. Another group of 45 prelingually deaf children with cochlear implants was obtained from the large-scale project underway at CID. All of these children were between eight and nine years of age and all of them had used their cochlear implant for at least five years. None of the children served in the previous experiment. Forty-five normal-hearing children were also recruited as a comparison group. They were matched for gender and chronological age with the

children from CID. The normal-hearing children were obtained from the Bloomington, Indiana community using names recorded from birth announcements that were published in the local newspaper. A hearing screening was carried out on each normal-hearing child to insure that there were no hearing problems at the time of testing. Left and right ears were screened separately using a Maico audiometer (Model MA 27) and TDH-39 headphones.

Procedures

The auditory signals used in this experiment were obtained from recordings made by a male talker. He recorded tokens of the following eight words: “red,” “blue,” “green,” “yellow,” “one,” “three,” “five,” and “eight.” The words were spoken in a clear voice at a moderate-to-slow speaking rate and were recorded digitally in real-time using 16-bit A-D converter running at 22 KHz. The amplitudes of the digital speech files were equated using software to achieve equal loudness. All of the auditory stimuli were output using a SoundBlaster AWE64 sound card and were presented over a loudspeaker at approximately 70 dB SPL.

Forward and backward WISC digit spans were obtained from each child using the procedures described earlier in the previous experiment. Digit spans were obtained using live voice presentation by the examiner with visual cues to lip-reading present. Presentation of the auditory and visual sequences in the Simon memory game and collection of the child’s responses was controlled by a computer program running on a PC. The response box consisted of a highly modified version of the Simon game that had been rebuilt and interfaced to the computer so that the lights and sounds could be varied and controlled independently by the experimenter under program control. The computer program automatically tracked the child’s performance in a given condition using an adaptive testing procedure developed by Levitt (1970) that is frequently used in psychophysical experiments.

This experiment was designed to measure reproduction spans using sequences of stimuli presented under three conditions, color-names and lights, digit-names and lights and lights-only. Both groups of subjects received all three conditions using a within subject design. The lights-only condition was always completed last in the series; the other two conditions were counterbalanced across subjects. In addition to the memory game task, both groups of children were administered the WISC forward and backward digit span tasks.

The children with cochlear implants were tested in St. Louis by clinicians and researchers who were highly experienced in working with hearing-impaired children. The normal-hearing children were tested in the Speech Research Laboratory at Indiana University in Bloomington by graduate students and undergraduate research assistants. All children were tested individually in a quite room. Subjects were introduced to the experimental task as a “memory game” and were shown how to press the buttons on the Simon response box. The subjects were told that they would be hearing sounds through the loudspeaker on the table in front of them and also seeing the buttons on the memory box light up. They were then instructed to pay attention to the computer and try to copy exactly what the computer does by pressing a sequence of buttons on the memory game box.

Results

WISC Digit Spans. Table VI shows the means, standard deviations and ranges for the forward, backward and total WISC digit spans for both groups of children. The spans displayed here were scored by total points using the procedures outlined in the WISC manual. Points are awarded for each list correctly repeated and no partial credit is given for incomplete lists. The left-hand panel of the table shows the digit

spans for the children with cochlear implants, the right-hand panel shows the spans for the normal-hearing children. The forward, backward and total summed digit spans were consistently shorter for the children with cochlear implants than the normal-hearing children replicating the previous findings reported by Pisoni and Geers (1998) with another group of cochlear implant users.

The results of the WISC digit span tests shown here demonstrate fundamental differences in working memory capacity between these two groups of children using highly familiar stimulus materials. Unfortunately, at this time without manipulating some other variable, it is not possible to identify precisely which specific aspect of working memory differs between the two groups. It is possible these differences in digit span are due to initial encoding operations, rehearsal processes, scanning or response output and retrieval of motor control programs used in speech articulation. Despite the ambiguity, however, the differences shown here are large and consistent and point to one possible locus of individual differences in processing stimulus input. The results of the Simon memory game provide additional information about the sources of these differences in working memory capacity.

Table VI

Summary of WISC Digit Spans for Implant Users and Normal-hearing Children

	WISC Digit Spans (N=43)			**WISC Digit Spans (N=44)**		
	Cochlear-Implant Users			**Normal-Hearing Children**		
	Summed Total	Direction of Recall		Summed Total	Direction of Recall	
	Points	Forwards	Backwards	Points	Forwards	Backwards
Maximum:	15	10	7	20	12	9
Minimum:	2	1	1	7	4	2
Mean:	**8.49**	**5.21**	**3.28**	**12.43**	**7.98**	**4.45**
Std. Dev.:	2.97	1.85	1.56	3.44	2.11	1.56

Simon Reproduction Spans. The averaged results for both groups of subjects on the Simon Memory Game are displayed in Figure 11 separately for each of the three presentation conditions. The normal-hearing children are shown on the left; the children with cochlear implants are shown on the right. The dependent measure plotted here is the longest list length that the child could reproduce correctly at least once during a given condition. Examination of these data reveals several differences. First, the normal-hearing children have longer reproduction spans for all three conditions than the children with cochlear implants. Second, the normal-hearing children display a "redundancy gain" in the colornames + lights condition compared to the lights-only condition. These children were able to benefit and increase their memory spans when the colornames and lights were congruent and paired together simultaneously. In contrast, however, the hearing-impaired children with cochlear implants did not show this same pattern. There is no difference between the three presentation conditions for these children and they do not appear to be able to use the additional redundant auditory information to improve their reproductive spans in the colornames + lights condition. In addition, we also observed an unexpected finding in the lights-only conditions. Even in this condition that did not involve the presentation of any auditory information, the children with cochlear implants had significantly shorter reproduction spans than the normal-hearing children. This finding suggests that the differences in working memory span between these two groups are not directly related to encoding of auditory inputs via the cochlear implant but reflect some aspect of the rehearsal process or output routines used to generate a sequence of motor responses.

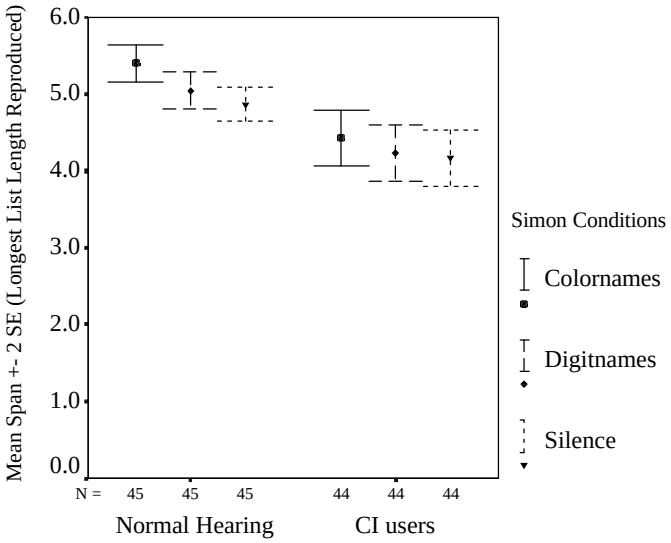


Figure 11. Mean Simon reproduction spans for three presentation formats (colornames, digitnames and lights-only) obtained from normal-hearing children (left) and deaf children with cochlear implants (right).

The absence of a redundancy gain in the colornames + lights condition and the shorter span length observed in the lights-only condition suggests that deaf children with cochlear implants may encode and process both auditory and visual information in ways that are fundamentally quite different from normal-hearing children who are able to make efficient use of cross-modal redundancies between correlated stimulus dimensions. Support for this hypothesis comes from an analysis of the WISC digit spans that were obtained from both groups. Not surprisingly, as we noted earlier, we found that the WISC forward and backward digit spans of the CI users were significantly shorter than the digit spans obtained from the normal-hearing children. Clearly, there are substantial differences in working memory span between these two groups even for highly familiar materials like digits that may reflect differences in encoding, working memory capacity or speed and/or efficiency of processing sensory information. These are all possible explanations of the differences found in digit spans between the two groups.

Of more interest, however, were the correlations between the WISC digit spans and the sequence reproduction spans obtained in the colornames + lights conditions. For the normal-hearing children the correlation between these two measures was positive and quite strong (r= +.58, p<.001), suggesting the operation and use of a common verbal rehearsal strategy in both memory tasks. However, the relationship between these two memory measures was quite different for the children with cochlear implants. There was no correlation at all between WISC digit span and colornames + lights (r= .09, NS), a finding that strongly suggests that the children with cochlear implants are carrying out the two memory tasks using different rehearsal strategies. For the colornames + lights memory game task, the CI children appeared to be using a visual-spatial rehearsal strategy based on encoding these multi-modal sequences as visual patterns. In contrast, in the WISC digit span task, they were forced to use a verbal rehearsal strategy because these items cannot be encoded or rehearsed using another alternative coding strategy.

An examination of the intercorrelations among the memory measures obtained from both tasks also showed a clear dissociation between the two groups of subjects in the pattern of correlations across these

two tasks. The colornames + lights and lights-only and the digitnames + lights and lights-only were both highly correlated for the cochlear implant group (r= +.71, p<.001 and r= +.64, p<.001, respectively) but these same conditions were not correlated at all for the normal-hearing group (r= +.20, NS and r= +.09, NS, respectively). Taken together, the pattern of results obtained from these two memory tasks suggests that the children with cochlear implants were not encoding the stimulus input in the same way as the normal-hearing children did in this task. When there is an optimal way of encoding a multi-modal sequence, the cochlear implant children appear to prefer a visual-spatial rehearsal strategy while the normal-hearing children automatically use a verbal rehearsal strategy. It is possible that differences in encoding are related to processing speed as well as automaticity. Children with cochlear implants, even children who have used their implant for at least 5 years, may not be able to rapidly encode and maintain complex multi-dimensional inputs in working memory for short period of time. Normal-hearing children have developed very efficient strategies for verbal coding and rehearsal based on their experiences perceiving and using spoken language. It is quite possible that hearing-impaired children with cochlear implants have much less efficient and slower verbal rehearsal strategies which not only effect their working memory capacity as measured by WISC digit spans but also influence how they perceive and encode multi-modal inputs that could be perceived and coded in memory in an alternative manner using visual-spatial cues available in the stimulus display.

Although some children with cochlear implants can perceive speech and understand spoken language at reasonably high levels of performance as measured by standardized outcome measures, they may nevertheless encode and process speech signals in ways that are non-optimal given the high redundancy of human language. Traditional behavioral tests that are based on response accuracy may not be able to detect and measure some of these subtle differences in processing speed, efficiency and the use of stimulus redundancy, which are characteristic markers of normal-hearing listeners.

General Discussion

In 1995 the NIH published a Consensus Statement on Cochlear Implants in Adults and Children to provide clinicians and other health care providers with a current summary of the benefits and limitations of cochlear implants (NIH, 1995). The 14-member consensus panel consisted of experts representing the fields of otolaryngology, audiology, speech-language pathology, pediatrics, psychology and education. The panel concluded that while cochlear implants improve communication abilities in most postlingually deafened adults with severe to profound hearing loss, the outcomes of implantation are much more variable in children, especially prelingually deafened children. Among other findings related to the efficacy and effectiveness of cochlear implants, the report emphasized the wide variation in outcome measures among implant users and recommended that additional research be carried out on individual differences in both adults and children. The report also suggested that new methods and tools should be used to study how cochlear implants activate the central auditory system.

An examination of the literature on the effectiveness of cochlear implants in prelingually deaf children reveals several closely related findings which strongly suggest that “central” auditory, cognitive and linguistic factors may play an important role in accounting for the enormous variation and individual differences observed in traditional outcome measures. Although the NIH Consensus Statement mentioned “central” auditory factors, the panel was not specific about precisely what these factors might be, or what role higher-level cognitive processes might play in outcome measures. We can speculate that these would include, at the very least, processes such as perception, learning, memory, and language. We believe that the recommendations of the NIH panel are fundamentally correct in drawing attention to central auditory factors as another unexplored source of variance and recommending new research on cognitive processes

and language development. Higher-level cognitive processes have not received much attention in the past and investigation of these factors may provide new insights into the underlying basis for the large individual differences as well as the specific effects of early experience on speech and language development in children with cochlear implants.

Five Key Findings

Five "key" empirical findings on cochlear implants in children were discussed earlier in the introduction to this report. In this final section, I first review these findings briefly because they provided the motivation for the new research we have carried out on process measures of performance. Then I attempt to tie several themes together and draw a few general conclusions that follow directly from our recent findings on working memory and coding strategies.

The existence of large individual differences in outcome is probably the most important issue in pediatric cochlear implantation at the present time. Until recently, we knew very little about the nature of these differences and the factors that were responsible for the variation in performance. In addition to the issue of individual differences, I described four other related findings that suggest a possible underlying basis for the wide variation among deaf children with cochlear implants. A careful examination of these findings also suggests that it might be fruitful to adopt a new and somewhat different research strategy in the future that focuses on "process" rather than the final "product" of perceptual analysis. The study of traditional audiological outcome measures and the effects of demographic variables have provided a great deal of valuable information about cochlear implants and changes in performance over time. However, these kinds of outcome measures are somewhat limited because they only assess the final product of what is generally regarded to be a complex set of interacting processes that draw on many different sources of knowledge.

As noted earlier, both age of implantation and length of deprivation have been shown to be very strong predictors of outcome performance in deaf children with cochlear implants. Children implanted at a younger age generally do better than children implanted at an older age and children who have been deaf for shorter periods of time generally do better than children who have been deaf for longer periods of time. How can we explain these related findings? Both results demonstrate the important contribution of "sensitive" or "critical" periods in development and both findings suggest close links between neural development, on the one hand, and behavior, on the other hand. This conclusion seems to be especially true for the skills and abilities that underlie the development and use of speech and language and the underlying biological component that drives the process of development.

It has been known for many years that language development in normal-hearing children has a strong biological basis and follows a genetically programmed maturational schedule (Lenneberg, 1967; Pinker, 1994). There is every reason to suspect that language development in deaf children who receive a cochlear implant also follows the same biologically based developmental schedule as well. As soon as these children begin to receive some auditory stimulation, novel interactions begin to emerge as the perceptual system and specialized phonetic module begins to detect, perceive and encode regularities in the input patterns. The actual time-course of language development in deaf children with cochlear implants may be delayed somewhat compared to normal-hearing children because of variable periods of sensory deprivation before implantation. However, there is also a wide range of variation in the onset and time-course of language development in the normal-hearing children too so it should not be surprising to see some variation in speech and language development in deaf children with cochlear implants.

Variation is an inherent part of the process of normal language development and small differences in normal-hearing children may simply be magnified and exaggerated in deaf children who have received cochlear implants. Once deaf children begin receiving auditory stimulation via their cochlear implant, even if this information is impoverished and degraded, it may be sufficient to get the neural mechanisms going, so to speak, and to start the "normal" process of language on a developmental trajectory. Without auditory stimulation to the nervous system during critical periods of development, it may never be possible for spoken language to develop fully or for the underlying sensory, perceptual and cognitive processes to reach their optimal states in a mature adult. Thus, the findings that age of implantation and length of deprivation affect outcome can be understood within a somewhat broader theoretical framework of critical periods for sensory-motor development and vocal learning which reflect the underlying neural specialization for speech and language (Pytte & Suthers, 1999; Suthers et al., 1999).

The second finding we touched on earlier concerned the effects of communication mode on outcome measures. Numerous studies have shown that communication mode affects outcome measures, especially outcome measures that rely heavily on the use of oral-language skills and phonological processing strategies in traditional assessment tasks used to measure outcome such as open-set word recognition, speech intelligibility, spoken language comprehension and even reading. The findings from several recent studies investigating the effects of communication mode strongly suggest that the specific language-learning environment that the child is exposed to and develops in may play a critical role in modulating, shaping and facilitating vocal learning and the process of language development. These conclusions seem to be especially relevant for spoken language development and the development oral skills and abilities that are used in receptive and expressive language processing tasks. These particular tasks make use of skills that rely heavily on phonological representations of spoken words and phonological processing strategies that transform highly variable sensory inputs into stable internal representations that encode the sound contrasts of the language and provide the basis for the motor programs used in speech production.

Once again, these are not particularly surprising findings for anyone who is familiar with the recent literature on speech and language development in normal-hearing infants and young children (see, Jusczyk, 1997; Hart & Risley, 1995). Young children learn language very quickly and are unusually sensitive and highly attuned to the regularities and frequencies of sound patterns in their language-learning environment (Jusczyk & Aslin, 1995; Saffran, Aslin & Newport, 1996). Numerous studies have shown that this "attunement" process closely tracks very subtle acoustic-phonetic differences in the input signals that children are exposed to during the first year or two after birth (Aslin & Pisoni, 1980; Aslin, Jusczyk & Pisoni, 1998). It is very likely that these early perceptual strategies play an important role in segmentation and word learning and form the basis for later syntactic and semantic development (Jusczyk, 1997). Past research on cochlear implants has generally failed to acknowledge the important contribution of learning and memory to performance although audiologists have shown that habituation and acclimatization effects do affect most outcome measures (Robinson & Summerfield, 1996; Tyler & Summerfield, 1996).

The third finding that we discussed in the introduction was the apparent lack of reliable preimplant predictors of success with a cochlear implant. From a theoretical perspective, we consider this to be a very important result because it suggests that basic underlying cognitive factors such as learning, memory and attention may be the kinds of outcome measures that should be used to assess performance in children with cochlear implants. One of the major assumptions of the information processing approach to cognition that guides our research program is that perception is not immediate but is a result of a series of processing stages that take place over time. Sensation, perception and memory are conceptualized within this framework as representing a continuum of processing activities that are organized in a hierarchical manner

(Haber, 1969). Very complex interactions exist in the language-learning environment that affect the way raw sensory information is perceived, encoded, stored and interpreted by the child. Investigation of these intermediate processes may provide valuable new insights into the wide variation observed in outcome performance. While the lack of preimplant predictors based on traditional outcome measures may be somewhat troubling for clinicians and researchers who would like to maximize the benefits of cochlear implants by modifying or adjusting intervention strategies soon after implantation, the findings reported in this paper suggest that other more basic measures of performance that are related to the information processing operations and skills such as working memory, coding and rehearsal strategies may be worth exploring in greater detail in addition to the traditional audiological outcome measures that have been used over the years.

It is very likely that earlier studies of preimplant predictors of outcome performance may not have succeeded in measuring the critical processing variables that reflect the encoding, perception and storage of sensory and perceptual information. All of the traditional outcome measures that have been used in the past are based on standardized assessment tests developed within the fields of clinical audiology and speech pathology. These tests use measures of performance that are "static" and reply exclusively on accuracy scores. More importantly, these tests assess the final "product" of performance not the intermediate "processes" and structures leading up to a final response. Thus, the lack of "true" process measures of performance in both adults and children with cochlear implants may be one of the primary reasons why the past efforts to find preimplant predictors have been so unsuccessful so far. Two recent studies have looked at this problem in greater depth and report encouraging findings that it may be possible to identify reliable preimplant predictors of outcome performance.

In a study of postingually deafened adults, Knutson et al. (1991) found that preimplant performance on a visual monitoring task predicted audiological outcome after 18 months of implant use. Strong and highly significant correlations were found between visual monitoring performance in a signal detection task and scores on four sound-only audiological measures, sentence perception, consonant and vowel perception and phoneme recognition in words. These results obtained with adult patients demonstrate that the cognitive processing operations and skills needed to rapidly extract information from sequentially arrayed visual stimuli may also be used in processing auditory signals and may underlie the successful use of a cochlear implant (see also Gantz et al., 1993). The findings obtained in Knutson et al.'s study support the hypothesis that higher-level cognitive factors related to perception, attention and working memory capacity play an important role in predicting outcome with an implant. More importantly, these results show that preimplant measures of information processing in the visual modality can be used to predict speech perception performance in the auditory modality.

More recently, Tait, Lutman and Robinson (submitted) reported moderate correlations between pre-verbal communication measures extracted from an analysis of videotapes and several outcome measures of speech perception obtained from prelingually deaf children three years after implantation. Video recording of 33 children were transcribed and scored for various turn-taking and autonomy behaviors before implantation. Outcome measures of sentence perception, discourse tracking and telephone use were obtained without the use of visual cues and correlations were computed with the behaviors obtained from the coded videotapes. Although positive correlations were found between each of the outcome measures and the pre-implant behaviors coded from the videotape analysis, none of the correlations with the turn-taking behaviors reached significance. However, the correlations with the autonomy behaviors were significant suggesting that some pre-verbal communicative behaviors that are present before implantation are associated with audiological outcome measures of speech perception and language processing obtained three years later.

Although somewhat limited in scope and generality at this point, the findings of Tait et al. are very intriguing and suggest that several important aspects of the development of spoken language are already present in infancy in these deaf children. These underlying pre-verbal communication skills may function as the "prerequisites" for speech and language and therefore may be quite general in nature reflecting multi-modal interactions between perception and action that are not tied to a specific sensory modality (see also Rizzolatti & Arbib, 1998 for recent findings on "mirror neurons" in monkeys that link actions of an observer and actor). The Tait et al. findings are, of course, correlational in nature and it will be necessary not only to attempt a replication of these findings but also try to specify more precisely the underlying neural and perceptual mechanisms that are responsible for these differences. It is possible that differences in imitation behaviors, gestures and perceptuo-motor links between perception and action are the fundamental processes that actually underlie the observations from the analysis of the videotapes.

Finally, and this is worth emphasizing strongly here, it is entirely possible that no preimplant measures will ever be found that will predict outcome performance in children with cochlear implants. The reason for this assertion is obvious if we consider the contribution of learning and memory processes to vocal development. If the underlying abilities for speech and language "emerge" after implantation and are the end product of a set of complex interactions between sensory, perceptual, cognitive and linguistic factors and sensory-motor learning processes that develop over time according to some built-in biological schedule, it may not be possible to find a unique signature or marker in only one measure of behavior that reflects all these different kinds of interactions. Because a substantial portion of the variance in outcome may be related to higher-level central auditory factors and vocal learning that reflect the way the initial sensory information is perceived, encoded and stored in long-term memory, it may be necessary to develop an entirely new set of outcome measures that can be used to assess these kinds of central auditory factors. Some of these new outcome measures might be behaviorally-based and "process-oriented" in nature like the measures of working memory span and the rehearsal strategies described earlier in this paper. Other outcome measures may use electrophysiological techniques to measure neural responses to sound more directly (see Kraus et al., 1996) or neural imaging (Naito et al., 1997; Wong, 1999) or measures of sensory-motor integration and vocal learning (Pytte & Suthers, 1999).

Analysis of the "Stars"

In the first section of this paper, we presented the results of a series of analyses that were carried out on two groups of prelingually deaf children who had received cochlear implants. An extreme groups design was used to identify differences in performance on a battery of speech and language measures that might provide new insights into the large individual differences observed on outcome measures with these children. One group consisted of children who were exceptionally good cochlear implant users-- the so-called "Stars." These were children who scored in the top 20% on the PBK test, a very difficult "open-set" test of spoken word recognition that has been used in the literature to identify exceptionally good implant users. A second group of children were selected as control subjects to draw comparisons. The children in this group scored in the bottom 20% on the PBK test and were unable to recognize any of the test words when they were presented in isolation using an open-set format. After the subjects were assigned to these two groups, scores on tests of speech perception, language comprehension, spoken word recognition, receptive vocabulary, receptive and expressive language development and speech intelligibility were obtained from an existing longitudinal database of 160 subjects. Descriptive analyses were carried out first to compare differences between the two groups on these different measures. Then a series of correlational analyses were computed for each group separately in order to study the relations among the dependent

variables and to uncover patterns that might reflect common underlying sources of variance that could be used to predict outcome.

The results of our descriptive analyses after one year of implant use revealed several interesting findings about the exceptionally good users of cochlear implants. First, we found that although the "Stars" showed consistently better performance on some measures such as speech perception, language comprehension, spoken word recognition and speech intelligibility than the "control" group, the two groups did not differ from each other on measures of receptive vocabulary knowledge, non-verbal intelligence, visual-motor integration or visual attention. We also found that some measures of performance continued to improve over a time period of six years whereas other measures remained fairly stable and did not show changes with experience after the first year. These overall findings demonstrate that the "Stars" differ in selective ways from the "Control" subjects and whatever differences are revealed by other descriptive measures, it is clear that the results are not due to some global difference in overall performance between the two groups. More importantly, we found that the "Stars" displayed exceptionally good performance on another test of spoken word recognition, the LNT, which demonstrates that the superior skills and abilities of these children are not due to the specific items on the PBK test or the methods used to administer the test.

The results of correlational analyses carried out on the test scores for the "Stars" one year after implantation showed a consistent pattern of very strong and highly significant intercorrelations among several of the dependent variables, particularly for the measures of word recognition, language development and speech intelligibility, suggesting a common underlying source of variance. These patterns of correlations were not observed for the control group. The common source of variance found in the correlational analyses of the "Stars" seemed to be related in some way to the processing of spoken words and to the encoding, storage and retrieval of the phonological representations of words. Of particular interest was the finding of strong correlations with speech intelligibility scores for these children, suggesting transfer of knowledge and a common shared representational system for speech perception and speech production (see also Shadmehr & Holcomb, 1997). These analyses suggested that the exceptionally good performance of the "Stars" may be due to their superior skills and abilities to process spoken language, specifically, to perceive, encode and retrieve phonological representations of spoken words from lexical memory and use these representations in a variety of different language processing tasks, especially tasks that make use of vocal learning.

While the results of these correlational analyses were suggestive and point to several new directions for future research on individual differences, the data available on these children were confined to traditional outcome measures in our database that were collected as part of the annual assessments. All of the scores on these tests represent the final product of perceptual and linguistic analysis. Process measures of performance were not part of the standard research protocol and were never collected from these children so it was impossible to investigate differences in speed, fluency or capacity at this time. Differences in information processing including topics such as perceptual learning, categorization, attention, and memory may underlie the individual differences observed between the two groups of children in our initial study. However, traditional audiological assessments of hearing and speech perception performance especially those used in assessing performance of deaf children with cochlear implants have not typically measured these types of processing activities.

Working Memory Spans

In the second section of this paper, we described the results of a recent study carried out by Pisoni and Geers (1998) who obtained measures of working memory using the digit span subtests from the WISC. Memory span data were collected from 43 eight- and nine-year old prelingually deaf children with cochlear implants as part of a larger project on speech and language development that is being conducted at CID. All of the children included in this study had used their cochlear implants for a period of at least five years. This study of digit spans was the first investigation to obtain a measure of processing capacity, specifically, measures of working memory span from a large number of children. Correlations were carried out between digit span and four sets of outcome measures that assessed speech perception, speech production, language and reading. Moderate to high correlations were obtained between forward auditory digit span and each of the four outcome measures. The pattern of correlations suggested the presence of a common source of variance that is related in some way to working memory, specifically, the encoding and rehearsal of phonological representations of spoken words. Thus, differences observed on various outcome measures of performance using standardized assessment tests may actually reflect more fundamental differences in the way sensory information is processed by the nervous system and used in specific language processing tasks that make use of this information.

The findings from the study by Pisoni and Geers on working memory span in deaf children with cochlear implants are consistent with a large and growing body of recent data on working memory and language development in normal-hearing children. Gathercole and her colleagues have reported strong correlations between measures of working memory span and early word learning, vocabulary knowledge and non-word repetition abilities (Gathercole et al., 1997). Gupta and MacWhinney (1997) have suggested that the working memory system is the common processing mechanism and serves as the "interface" between speech perception and speech production and the phonological knowledge of words stored in the mental lexicon. Thus, language processing and working memory are closely linked together in a variety of tasks that require access to phonological information about words in the lexicon.

The correlations between working memory and several different measures of language processing found in our study demonstrate the important contribution of "processing variables" – fundamental information processing operations that are used in the encoding, storage, retrieval and rehearsal of the phonological representations of spoken words. These new findings on auditory digit span in children with cochlear implants help to identify the "locus" of precisely where differences in performance are located within the larger information processing system and how they operate in specific tasks. The differences in performance observed in the Pisoni and Geers study may be due to the operation of a subcomponent of working memory known as the "phonological loop." The phonological loop is responsible for the rehearsal and maintenance of the phonological representations of spoken words in memory and plays a critical role in the learning of new words (Baddeley et al., 1998). The phonological loop is also assumed to play an important role in speech production as well by mediating access to retrieval of sensory-motor plans needed for speech motor control and articulation. Finally, the phonological loop is also used in reading, especially reading unfamiliar words or novel nonword patterns. All of these language processing tasks draw on the same common set of phonological representations of spoken words and all of them use the same processing resources in working memory, specifically, the phonological memory store and the articulatory subvocal rehearsal mechanism that serves as the primary interface between the initial sensory input and the representations of spoken words in the mental lexicon.

The correlations found between digit span and speech intelligibility and digit span and speaking rate suggest that rehearsal speed in working memory may be one of the factors that distinguishes good implant users from poorer ones. While these results are only correlational in nature, they do provide additional converging support for the proposal that differences in working memory are responsible for the

enormous variation in outcome scores on standardized assessments used with these children. At the very least, these findings point to a specific processing mechanism and suggest several new directions to pursue in future studies. Additional support for the importance of working memory and rehearsal speed was observed in another analysis. In addition to the correlations between digit span and the four outcome measures reported earlier, Pisoni and Geers (1998) also observed a moderate but significant correlation between digit span and communication mode (r=+.38 p<.05). This correlation suggests that early auditory experience in oral-only programs may have very specific effects on working memory capacity and the elementary information processing operations that are used in language processing tasks. Not only was there a positive correlation between communication mode and auditory digit span but a subsequent analysis showed that the children from oral-only programs had significantly longer digit spans than the children from TC programs.

Children from oral-only programs are not only exposed to more speech and language (see Hart & Risley, 1995) but they also engage in more meaningful processing activities that require them to construct more robust phonological representations of the sound patterns of spoken words in their language. The study of working memory in these children provides us with a new approach to two long-standing research issues in the field of cochlear implants, the enormous individual differences observed among children with cochlear implants and the role of early auditory experience in the language-learning environment. Spoken language processing and working memory appear to be closely related in both normal-hearing children and deaf children with cochlear implants. The present findings suggest that early experience and exposure to oral language affects the underlying perceptual and sensory-motor mechanisms used to process and code the sensory information. Thus, working memory appears to be influenced and shaped by early auditory exposure and experience with spoken language in the language-learning environment. Again, this suggests that specific experience with spoken language and exposure to spoken words has an effect on a specific processing mechanism—working memory and a subcomponent of this system related to speed of rehearsal.

Coding and Rehearsal Strategies

In the third section of this paper, we presented some recent findings obtained by Cleary, Pisoni and Geers (this volume) using a new experimental methodology, the Simon memory game, to measure reproductive memory spans for sequences of stimuli. This memory game procedure was originally developed to obtain measures of working memory span without requiring the child to produce an explicit verbal-motor response as output. When digit spans are obtained using the WISC, the child simply repeats back or imitates the sequence of digits and the examiner records the child's verbal response. Using the Simon memory game procedure, children were presented with a sequence of stimuli, either visual-only or combined auditory+visual that were selected from an ensemble of four possible signals and were simply required to reproduce the stimulus pattern by pressing a sequence of buttons on a response panel. The difficulty of the Simon memory task and the amount of concurrent "processing load" was manipulated by increasing the length of the sequence to be reproduced using an adaptive testing program and then recording the longest list length that a child was able to reproduce correctly under a given stimulus condition. This sequence length was taken as a measure of the child's "reproductive" memory span in a given experimental condition.

WISC digit spans and Simon reproductive memory spans using the new procedures were obtained for two groups of age-matched children, 45 deaf children with cochlear implants and 45 normal-hearing children. Both groups were tested under several different presentation conditions. For sequences of colored lights and color names, we found that the normal-hearing children showed a "redundancy gain, " an advantage for the auditory plus visual condition (lights and sounds) compared to the visual-only (lights)

when the color names matched the colors of the lights that were illuminated on the panel. The children with cochlear implants did more poorly overall than the normal-hearing group on all of the conditions we studied and they failed to show the same advantage under the combined simultaneous auditory+visual conditions. The difference in performance between the two groups on the visual-only sequences which was not originally anticipated suggested the possibility that the children with cochlear implants might be using a different coding strategy to carry out the task, a visual-spatial strategy that relied entirely on encoding and rehearsing visual patterns without phonological-verbal coding or mediation.

Correlations were then carried out separately for each group using the measures of working memory obtained from the digit span task and the measures of reproductive span from the Simon memory game. An examination of the patterns of correlations across these tasks revealed a very interesting dissociation between visual-spatial and verbal rehearsal strategies in the two groups. The normal-hearing children appeared to use verbal coding and verbal rehearsal strategies in both memory tasks. Strong correlations were observed between forward digit spans and the combined auditory+ visual presentation conditions in the Simon memory game. In contrast, the children with cochlear implants relied on visual-spatial cues to do the Simon memory game in both the visual-only and the combined auditory+visual conditions. The correlations within the Simon conditions were all positive and very strong for the implant children but were weak or non-existent for the normal-hearing children.

The low correlation observed between the Simon auditory+visual condition and WISC forward digit span suggested that the children with cochlear implants were not able to take advantage of the redundancy across stimulus dimensions in the combined auditory+visual condition of the Simon game and simply encoded these sequences using visual-spatial cues. The pattern of results across these two tasks suggests that fundamentally different modes of processing complex multi-dimensional stimuli are being used by these two groups of children to encode and reproduce these sequences. When the memory task can be performed by using either verbal coding or visual-spatial coding, the normal-hearing children rely on verbal coding and are able to take advantage of the cross-modal redundancies between stimulus dimensions while the deaf children with cochlear implants prefer to use the visual-spatial cues and seem to ignore the auditory signals that are presented simultaneously with the light patterns. Exactly why they have a preference to do this is not entirely clear right now but it may reflect differences in processing capacity, robustness of stimulus encoding or automaticity. We know from the data on speech feature discrimination obtained with the minimal pairs test, that children with cochlear implants do not encode fine phonetic details in speech the way normal-hearing children do. As a result, they may construct and use only partial under-specified representations of signals in their environment and they may not process redundant stimulus information especially when the redundancy requires integration of multi-modal inputs from separate sensory modalities.

Process Measures of Performance

It should be clear from the three sets of findings presented in this paper that new “process” measures of outcome performance will be needed to assess learning, memory, attention and categorization--the “central” cognitive processes that act on and use the initial sensory input provided by the cochlear implant. Traditional audiological outcome measures are simply not adequate to assess the underlying processes used in speech and language processing tasks. Instead, one can imagine the development of an entirely new battery of “process” measures of performance that could be used to assess the flow and content of information as it is processed and coded by the listener. These new measures would be designed to measure and quantify what the listener does with the limited sensory information he/she receives through the cochlear implant. Some of these new measures could be used to assess differences in processing speed,

efficiency and capacity. Thus, measures of working memory span, coding and rehearsal strategies, selective and divided attention and automaticity of processing may be much more informative and useful than the traditional battery of audiological outcome measures that have been used in the past which assess the final product of perceptual analysis. By refocusing research to the study of the elementary cognitive processes that are assumed to underlie the observed behavior in specific tasks, it may also be possible to develop a new set of preimplant measures that are more successful in predicting outcome than the current procedures available now. Given the tools that are currently available and the theoretical framework of human information processing, we think these are reasonable goals that can be achieved in a relatively short period of time.

Some New Research Directions

The findings from the three studies summarized in this paper strongly suggest that "central" auditory factors and higher-level cognitive processes contribute to the variability in performance found in traditional assessment measures. Using a new theoretical framework based on concepts and methodologies from the information processing approach in cognitive psychology, we have been able to gain some new insights into the specific sensory, perceptual, and linguistic factors that are responsible for the enormous individual differences in performance on a wide range of audiological outcome measures. Although our initial analysis of the descriptive data from the "Stars" pointed to a common source of variance associated with the processing of spoken words and the use of phonological knowledge, this account was based entirely on patterns of correlations among variables using existing data that were originally collected for audiological assessment purposes not hypothesis testing. When this research was originally carried out as part of a longitudinal study of outcome performance, process measures of memory, attention and learning or the time course of early perceptual analysis were not obtained from any of these children. As a consequence, we do not know what these deaf children actually do with the limited sensory information they receive through their cochlear implant and how their perceptual processing differs from normal-hearing and normal-developing children.

Our recent investigations of working memory capacity have provided the first direct evidence that process variables related to verbal coding and rehearsal contribute to the variability in audiological outcome measures. Two findings from these studies are particularly noteworthy. First, we found that working memory span was strongly correlated with several traditional outcome measures. This finding suggests that working memory and capacity limitations of working memory may be the locus of the individual differences observed in performance on other language processing tasks that draw on this common mechanism. The findings on speaking rate suggest that these capacity limitations are related in some way to rehearsal speed and processing efficiency. Second, we found that deaf children with cochlear implants do not automatically encode and process sequences of multi-dimensional stimuli using verbal rehearsal but instead use a visual-spatial coding strategy to perform the task. This finding demonstrates important differences in modes of processing complex stimuli and suggests that even very elementary cognitive processes like automaticity, attention and the allocation of processing resources may be carried out in fundamentally different ways by deaf children who have received cochlear implants after a period of sensory deprivation. The use of verbal rehearsal strategies in sequence memory tasks like those studied here may be an important early diagnostic marker that predicts success with a cochlear implant on a variety of outcome measures that make use of phonological coding and phonological processing skills.

Although these two initial studies have provided new knowledge about the operation of working memory and the use of different modes of processing, we still do not have any data on basic processes of learning, attention, automaticity and categorization in this population of children. In order to obtain a

complete picture of the information processing capabilities of these children and the possible differences in the way they process and encode sensory information because of the limitations of their cochlear implant, it will be necessary to carry out additional studies of perceptual learning, implicit and explicit memory and long-term retention of verbal and non-verbal materials. It is possible that the differences observed in our studies of working memory may actually reflect the contribution of long-term knowledge and exposure to spoken language. Differences in long-term memory and lexical knowledge may also affect the speed of rehearsal and the time-course of perceptual processing for familiar and unfamiliar words which, in turn, may affect measures of processing capacity in working memory tasks because of differences in the rate of scanning verbal materials in short-term memory. Other research should be carried out on automaticity and attention as well as categorization of novel stimulus materials in order to chart out the time-course of perceptual learning and study long-term retention of new knowledge gained under controlled laboratory conditions. Studies of episodic and semantic memory and incidental learning may provide some additional valuable insights into the extent to which instance-specific details of complex stimulus events are encoded and stored in long-term memory and later used in categorization and memory tasks.

The findings reported by Pisoni & Geers (1998) that children from oral-only environments have longer memory spans than children from total communication environments demonstrates that working memory capacity is not fixed or hard-wired at birth but is extremely flexible and reflects on-going learning processes and interactions between the sensory and perceptual environment and the real-time processing mechanisms used to encode, store and manipulate information. Recent findings on vocabulary acquisition and language development summarized by Hart & Risley (1995), suggest that frequency of exposure to language matters and that vocabulary knowledge is related to the number and frequency of words used by the parents in the child's language-learning environment. Other recent evidence suggests that memory spans for lists of spoken words are related to lexical knowledge and competition among phonetically similar items in long-term memory (Goh & Pisoni, 1998).

Studies on early word learning and research on the organization of words in the mental lexicon will also provide new information about how spoken words are encoded and stored in long-term memory and how children with cochlear implants gain access to this information and use it in receptive and expressive language processing tasks. It has been reported that normal-hearing children show very rapid word learning at a certain point in language development, a process often referred to in the literature as "fast mapping." These findings suggest that words are learned quickly and effortlessly in context with only a few exposures (see Carey, 1978; Carey & Bartlett, 1978; Dollaghan, 1985; 1987). The process of early word learning in normal-hearing children appears to be quite different from other forms of perceptual learning and development which often show a very slow and gradual process of acquisition (Woodward & Markman, 1997). At the present time, we have very little knowledge about how deaf children with cochlear implants learn novel words, how they organize spoken words in long-term memory or how they retrieve the phonological representations of spoken words from the mental lexicon in different types of language processing tasks, especially tasks that require them to reproduce and imitate actions and specific gestures. The close relations between working memory, measures of immediate serial recall and lexical development noted by Gupta and MacWhinney (1997) and discussed at some length in their recent theoretical papers, suggest that this particular topic may be a very important and fruitful area for future research on individual differences in these children. This topic seems to be especially attractive at this time given the importance of lexical processes that emerged from the correlations that were computed on the outcome measures obtained from the "Stars" in our first study and the role working memory plays in language development, specifically, phonological and lexical development (Gathercole, Hitch, Service & Martin, 1997).

Summary and Conclusions

Our long-term goal is to obtain new knowledge about the effectiveness of cochlear implants in adults and children. Along the way, we hope to gain a detailed understanding of the basis for the enormous individual differences in audiological outcomes that have been reported universally in the literature on prelingually deaf children with cochlear implants. To accomplish this task, we began our investigation of individual differences by studying the exceptionally good users of cochlear implants, the so-called "Stars," in order to narrow down the scope of the problem and formulate more specific testable hypotheses about the underlying processes that distinguish this group of children from the other deaf children who are not able to benefit as much from their cochlear implant. In our initial analyses, we found that the "Stars" differed in several theoretically interesting ways from a comparison group of children with cochlear implants. Correlational analyses revealed a common underlying source of variance associated with spoken word recognition in several different tasks that required lexical access and use of phonological processing skills. This pattern emerged in both receptive and expressive measures of language and was related to the language-learning environment, the communication mode that the child was exposed to after implantation. The correlational analyses of the "Stars" also suggested that working memory capacity might be one place to look for the locus of the individual differences and variation in performance on audiological outcome measures of performance.

To study working memory, we obtained auditory digit spans from a large group of deaf children with cochlear implants. We found that the working memory spans of deaf children with cochlear implants were highly correlated with a number of traditional audiological and language-based outcome measures, including measures of spoken language comprehension, speech intelligibility and reading. In another study using a new experimental methodology to measure working memory spans that did not require a verbal output response, we found that deaf children with cochlear implants do not automatically encode and integrate redundant stimulus dimensions for sequences that are presented cross-modally. Instead, they perceived and processed these patterns using visual-spatial coding and a rehearsal strategy that appears to be fundamentally different from the highly automatized verbal coding strategies that age-matched normal-hearing children employed in the same experimental task.

Taken together, the present findings support the hypothesis that central auditory processes related to perception, attention, memory, learning and language may underlie the large individual differences observed in traditional audiological outcome measures reported in the literature for this population of implant users. Knowledge and understanding of the processes and mechanisms responsible for differences in the effectiveness of cochlear implants and variation in audiological performance should be extremely valuable to both clinicians and researchers for several reasons. First, this new knowledge will provide a principled theoretical basis for the development of new intervention strategies for children who are not benefiting optimally from their cochlear implant at a time in their development of language when changes can still be made. Second, detailed knowledge of the underlying source of these individual differences may help in identifying new preimplant predictors of audiological outcome. These predictors can then be used to refine current criteria for candidacy based on direct behavioral measures of performance which are related in a straightforward way to a variety of outcome measures, particularly measures of speech perception, language processing and language development.

Finally, several new areas of research on individual differences were identified based on techniques and experimental procedures from cognitive psychology that are designed to measure the content and flow of information as it is processed by the central nervous system. These new research methods focus on the microstructure of cognition, processing speed, capacity limitations and modes of processing sensory

information, the intermediate processes and structures that underlie behavior in a given task rather than the final product of perceptual analysis which has been the primary defining characteristic of the traditional approach to audiological outcome measures used in the past.

References

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

Aslin, R.N., Jusczyk, P.W. & Pisoni, D.B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In W. Damon (Series Ed.), *Handbook of Child Psychology. Fifth Edition, Volume 2: Cognition, Perception & Language* (D. Kuhn & R. Siegler, Eds.). New York: Wiley. Pp. 147-198.

Aslin, R.N. & Pisoni, D.B. (1980). Some developmental processes in speech perception. In G. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child Phonology: Perception and Production*, New York: Academic Press, Pp. 67-96.

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

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

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

Carey, S. (1978). The child as word learner. In, M. Halle, J. Bresnan, & G.A. Miller (Eds.), *Linguistic theory and psychological reality*. Cambridge: MIT Press.

Carey, S. & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17-29.

Carlson, J., Cleary, M., & Pisoni, D.B. (1998). Performance of normal-hearing children on a new working memory span task. In *Research on Spoken Language Processing Report No. 22* (pp. 251-276). Bloomington, IN: Speech Research Laboratory.

Carpenter, P.A., Miyake, A., & Just, M.A. (1994). Working memory constraints in comprehension: Evidence from individual differences, aphasia, and aging. In M.A. Gernsbacher et al. (Ed), *Handbook of Psycholinguistics*. (pp. 1075-1122). San Diego, CA, USA: Academic Press, Inc.

Chase, W.G. (1977). Does memory scanning involve implicit speech? In S. Dornic (Ed.), *Attention and Performance VI*. Hillsdale, NJ: LEA.

Cleary, M. (1997). Measures of phonological memory span for sounds differing in discriminability: Some preliminary findings. In *Research on Spoken Language Processing Report No. 21* (pp. 93-140). Bloomington, IN: Speech Research Laboratory.

Cleary, M., Pisoni, D.B., & Geers, A.E. (this volume). Some measures of verbal and spatial working memory in eight- and nine-year-old hearing-impaired children with cochlear implants.

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

Dollaghan, C. (1985). Child meets word: "Fast mapping" in preschool children. *Journal of Speech and Hearing Research, 28,* 449-454.

Dollaghan, C.A. (1987). Fast mapping in normal and language-impaired children. *Journal of Speech and Hearing Disorders, 52,* 218-222.

Dunn, L. & Dunn, L. (1997). *The Peabody Picture Vocabulary Test—Third Edition*. Circle Pines, MN: American Guidance Service.

Ericsson, K.A., & Pennington, N. (1993). The structure of memory performance in experts: Implications for memory in everyday life. In Davies, G.M. & Logie, R.H. (Eds), *Memory in everyday life. Advances in psychology, 100*. (pp. 241-272). Amsterdam, Netherlands: North-Holland/Elsevier Science Publishers.

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

Gantz, B.J., Woodworth, G.G., Abbas, P.J., Knutson, J.F. & Tyler, R.S. (1993). Multivariate predictors of audiological success with multichannel cochlear implants. *Annals of Otology, Rhinology & Laryngology, 102,* 909-916.

Gathercole, S.E., Frankish, C.R., Pickering, S.J. & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25,* 84-95.

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.

Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59(2),* 267-333.

Haber, R.N. (1969). *Information-processing approaches to visual perception.* New York: Holt, Rinehart & Winston.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore, MD: Paul H. Brookes Publishing Co.

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

Jusczyk, P.W. (1997). Finding and remembering words: Some beginnings by English-learning infants. *Psychological Science, 6,* 170-173.

Jusczyk, P.W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press/Bradford Books.

Jusczyk, P.W. & Aslin, R.N. (1995). Infants' detection of sound patterns in fluent speech. *Cognitive Psychology, 29,* 1-23.

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.

Kluck, M., Pisoni, D.B., & Kirk, K.I. (1997). Performance of normal-hearing children on open-set speech perception tests. *Research on Spoken Language Processing Report No. 21,* Bloomington, IN: Speech Research Laboratory, (pp. 349-366).

Knutson, J.F., Hinrichs, J.V., Tyler, R.S., Gantz, B.J., Schartz, H.A. & Woodworth, G. (1991). Psychological predictors of audiological outcomes of multichannel cochlear implants: Preliminary findings. *Annals of Otology, Rhinology & Laryngology, 100,* 817-822.

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 vocalization*. Cambridge University Press. Pp. 29-48.

Kraus, N., McGee, T.J., Carrell, T.D., Zecker, S.G., Nicol, T.G., Koch, D.B. (1996). Auditory neurophysiologic responses and discrimination deficits in children with learning problems. *Science, 273,* 971-973.

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

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1-36.

Marler, P. & Peters, S. (1`988). Sensitive periods for song acquisition from tape recordings and live tutors in the swamp sparrow, *Melospiza georgiana*. *Ethology, 77,* 76-84.

McGarr, N. (1981). The effect of context on the intelligibility of hearing and deaf children's speech. *Language and Speech, 24,* 255-264.

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 & Hearing,20,* 363-371.

Naito, Y., Okazawa, H., Hirano, S., Takahashi, H., Kawano, M., Ishizu, K.L., Yonekura, Y., Konishi, J., & Honjo, I. (1997). Sound induced activation of auditory cortices in cochlear implant users with post- and prelingual deafness demonstrated by positron emission tomography. *Acta Oto-Laryngol, 117,* 490-496.

NIH Consensus Conference. (1995). Cochlear implants in adults and children. *Journal of the American Medical Association, 274*, 1955-1961.

O'Donoghue, G.M., Nikolopoulos, T.P., Archbold, S.M. & Tait, M. (1999). Cochlear implants in young children: The relationship between speech perception and speech intelligibility. *Ear & Hearing, 20*, 419-425.

Osberger, M.J., Miyamoto, R.T., Zimmerman-Phillips, S., Kemink, J.L., Stroer, B.S., Firszt, J.B., & Novak, M.A. (1991). Independent evaluation of the speech perception abilities of children with the Nucleus-22 channel cochlear implant system. *Ear and Hearing, 12*, S66-S80.

Osberger, M.J., Robbins, A.M., Todd, S.L. & Riley, A.I. (1994). Speech intelligibility of children with cochlear implants. *The Volta Review, 96*, 169-180.

Pinker, S. (1994). *The language instinct*. New York: William Morrow and Co.

Pisoni, D.B. & Geers, A.E. (1998). Working memory in deaf children with cochlear implants: Correlations between digit span and measures of spoken language. *Research on Spoken Language Processing Progress Report No. 22*. Bloomington, IN: Speech Research Laboratory, (pp. 336-343).

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*. Bloomington, IN: Speech Research Laboratory, (pp. 51-91).

Pytte, C. & Suthers, R.A. (1999). A bird's own song contributes to conspecific song perception. *NeuroReport 10*, 1773-1778.

Reynell, J.K. & Huntley, M. (1985). *Reynell Developmental Language Scales - Revised*, Edition 2. Windsor, England: NFER-Nelson Publishing Company, Ltd.

Rizzolatti, G. & Arbib, M.A. (1998). Language within our grasp. *Trends in Neuroscience, 21*, 188-194.

Robbins, A.M., Renshaw, J.J., Miyamoto, R.T., Osberger, M.J., & Pope, M.J. (1988). *Minimal Pairs Test.* Indianapolis, IN: Indiana University School of Medicine.

Robinson, K. & Summerfield, A.Q. (1996). Adult auditory learning and training. *Ear & Hearing, 17*, 51-65.

Saffran, J.R., Aslin, R.N. & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Schweickert, R., & Boruff, B. (1986). Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 419-425.

Schweickert, R., Guentert, L., & Hersberger, L. (1990). Phonological similarity, pronunciation rate, and memory span. *Psychological Science, 1*, 74-77.

Shadmehr, R. & Holcomb, H.H. (1997). Neural correlates of motor memory consolidation. *Science, 277,* 821-825.

Suthers, R.A., Goller, R. & Pytte, C. (1999). The neuromuscular control of birdsong. *Phil. Trans. R. Soc. Lond. B, 354,* 927-939.

Svirsky, M.A., Robbins, A.M., Kirk, K.I., Pisoni, D.B. and Miyamoto, R.T. (In press). Language development in profoundly deaf children with cochlear implants. *Psychological Science*.

Tait, M., Lutman, M.E. & Robinson, K. (submitted). Pre-implant measures of pre-verbal communicative behavior as predictors of outcomes in children. *Ear & Hearing.*

Tyler, R.S. & Summerfield, A.Q. (1996). Cochlear implantation: Relationships with research on auditory deprivation and acclimatization. *Ear & Hearing, 17,* 38-50.

Waltzman, S.B. & Cohen, N.L. (2000). *Cochlear implants.* New York: Thieme.

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

Wong, D., Miyamoto, R.T., Pisoni, D.B., Sehgal, M. & Hutchins, G.D. (1999). PET imaging of cochlear-implant and normal-hearing subjects listening to speech and nonspeech. *Hearing Research, 132,* 34-42.

Woodward, A.L. & Markman, E.M. (1997). Early word learning. In W. Damon, D. Kuhn & R. Siegler (Eds), *Handbook of Child Psychology, Vol 2: Cognition, Perception and Language* (pp. 371-420). New York: Wiley.

Zwolan, T.A., Zimmerman-Phillips, S., Ashbaugh, C.J., Hieber, S.J., Kileny, P. R., & Telian, S.A. (1997). Cochlear implantation of children with minimal open-set speech recognition skills. *Ear & Hearing, 18,* 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Some Measures of Verbal and Spatial Working Memory in Eight- and Nine-Year-Old Hearing-Impaired Children with Cochlear Implants[1]

Miranda Cleary, David B. Pisoni,[2] and Ann E. Geers[3]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH Research Grant DC00111 and Training Grant DC00012 to Indiana University. We extend grateful appreciation to B. Staley and J. L. Carlson for their help in testing of the normal-hearing children, to the CID staff for the testing of all hearing-impaired cochlear implant users, and to K. I. Kirk and L. Gerken for helpful comments and advice on an earlier version of this study.

[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

[3] Center for Applied Research on Childhood Deafness, Central Institute for the Deaf, St. Louis, MO 63110

Some Measures of Verbal and Spatial Working Memory in Eight- and Nine-Year-Old Hearing-Impaired Children with Cochlear Implants

Abstract. Two groups of 8 and 9-year-old children, 45 normal-hearing (NH) and 45 hearing-impaired users of cochlear implants (CI) completed a working memory task that presented either visual-spatial cues or visual-spatial cues paired with auditory signals. In this task colored buttons were illuminated in a sequence either with or without the synchronized auditory presentation of verbal labels for the buttons (color-names or digit-names). The child was required to reproduce each sequence by pressing the appropriate buttons. The longest list length reproduced under each set of conditions was recorded. The NH group demonstrated an advantage for the auditory plus visual-spatial task over the visual-spatial-only task when the labels matched the colors of the lights. The children in the CI group did not show this same advantage. The ability of many of the CI users to correctly identify the auditory stimuli when presented in isolation did not lead to their use of the informationally redundant auditory cues. Overall, the CI group did significantly worse than the NH children even on the visual-spatial-only sequences, suggesting important differences in encoding strategies. A modified task that eliminated the visual-spatial cues from the target sequence of auditory color-names was run with NH children and a second group of pediatric CI users. In this case, errors in stimulus identification were associated with reduced span scores. The highest scoring CI users obtained spans on this task equivalent to the scores of the bottom third of the NH group, indicating that when visual-spatial encoding was not an option, some CI users were capable of engaging in verbal encoding in a manner on par with some NH children.

Introduction

A long-standing debate surrounds the question of the extent to which information from the different sensory modalities is channeled independently through the mechanisms of short-term memory (Fastenau, Conant, & Lauer, 1998; Hale, Bronik, & Fry, 1997; Swanson, 1996). Over the years the theoretical construct of short-term/working memory[4] has been "fractionated" into modality-specific pathways in the brain's prefrontal cortex sharing a common "executive" control center (Baddeley, 1986; Smith & Jonides, 1999). Much effort has been expended in trying to understand how sensory experience contributes to these divisions as they appear to exist in the mature individual (Lewkowicz & Lickliter, 1994; Stein & Meredith, 1993). Examining the long-term impact of various forms of sensory deprivation may help shed further light on the nature of these "divisions" in normally developed organisms.

Past research on the development of short-term memory has tended to focus on whether increases in memory capacity, processing speed, or resistance to interference are primarily responsible for age-linked changes—with the modality specificity of these effects only now undergoing careful examination (e.g., Hale, Bronik, & Fry, 1997; Kail, 1991; Swanson, 1996). Effects of early sensory experience and learning have been examined by assessing how performance improves with greater familiarity with the to-be-remembered items and acquired strategies for circumventing capacity limitations (Dempster, 1981; Naus &

[4] *Short-term* memory tasks using verbal materials usually require no more than a simple parroting back of presented items, or immediate labeling of a test sequence as either identical to or different from a prior sequence. *Working* memory is argued to be used when maintenance of novel information is required while other incoming information is also processed, or some manipulation or transformation of items in memory is necessary.

Ornstein, 1983). Many encoding strategies, by their very nature, appear to favor a particular input modality. The learned strategy of *verbal rehearsal* for example, although best suited for the task of remembering the phonological characteristics of spoken or read words, tends to be utilized by normal-hearing adults whenever verbal labels can be applied—even to deleterious effect in the case of encoding and remembering ambiguous shapes, for example (Brandimonte, Hitch, & Bishop, 1992; Carmichael, Hogan, & Walter, 1932). This heavy reliance on verbal encoding strategies has naturally led to much interest in whether access to a typical oral/aural based language environment is a necessary component to the normal development of aspects of working memory such as Baddeley's "central executive" which is assumed to be modality-independent (Baddeley, 1992).

Literature published in the 1950-1970's suggested that hearing-impaired children in general were less adept overall than normal-hearing children in their short-term memory for some types of visual (but verbally code-able) sequences because these children lacked effective verbal strategies for rehearsal (Conrad, 1972; Furth, 1966; see Marschark & Mayer, 1998; Mayberry, 1992 for some recent reviews). In the 1980's and 90's however, the focus shifted to showing that manually signed equivalents to verbal strategies could be adopted to analogous benefit, although the comparable efficiency of these manual strategies, and the frequency of their use by signing individuals came under some debate (Hanson, 1990; Hanson & Lichtenstein, 1990; Shand, 1982).

The present study investigated the processing and short-term storage of auditory and visual-spatial sequences in prelingually-deafened pediatric cochlear implant (CI) users--profoundly deaf children that have been without auditory input for a period in their very early development, but who have then been supplied with access to sound via an electrically-coded signal presented directly to their auditory nervous system. Through an intervention program that follows the implant surgery, most pediatric CI users eventually come to gain at least an awareness of sound through their implant, and in some cases go on to develop good speech perception and production skills (Svirsky, Robbins, Kirk, Pisoni, & Miyamoto, in press). The skills attained through cochlear implant use vary quite widely among children, and the variables and processes that contribute to this outcome are not well understood (Pisoni, Svirsky, Kirk, & Miyamoto, 1997).

Large individual differences in spoken word recognition skills and language development continue to be observed in pediatric users of cochlear implants (Miyamoto et al., 1994; Tyler et al., 1997). A sizable proportion of this variability can be accounted for in terms of physiological and hardware-related factors. Miyamoto et al. (1994) suggest for example, that about 40% of the variance in open-set speech perception measures can be accounted for in terms of processor type, duration of deafness, communication mode, age of onset of deafness, duration of CI use and age at implantation. Other studies, using similar sets of variables, have obtained R-squared values ranging between 37-64% (Dowell, Blamey, & Clark, 1995; Snik, Vermeulen, Geelen, Brokx, & van den Broek, 1997). In young children, the age at which the hearing-impairment occurred, the duration of the period of auditory deprivation, and amount of experience using the implant have all been shown to play an important role in predicting outcome (Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997; Miyamoto et al., 1994; Nikolopoulos, O'Donoghue, & Archbold, 1999). From recent reexamination of audiological candidacy requirements, it has also been determined that the presence of even minimal amounts of residual hearing under aided conditions before implantation can also contribute positively to success with an implant (Seghal, Kirk, Pisoni, & Miyamoto, 1997; Zwolan et al., 1997). However, there is still a large proportion of unexplained variance in this multivariate prediction of speech perception performance (Pisoni et al., 1997).

Our current research program on individual differences and variation in CI users is motivated by the hypothesis that some significant part of this current error variance can be accounted for in terms of already established knowledge about the distribution of specific cognitive capacities in the general population of young children. In particular, one of our goals has been to tease apart possible contributions from modality-specific versus modality-independent mechanisms of short-term and/or working memory, in light of the pediatric cochlear implant user's newly acquired access to auditory information.

Although it is common practice to administer tests of "general intelligence" prior to implantation (Miyamoto, Robbins, & Osberger, 1993; Tiber, 1985), the early finding that overall IQ was a poor predictor of speech perception performance seems to have discouraged more detailed investigation of whether certain subscales within the IQ batteries might show more predictive power than others (though see Quittner & Steck, 1991; Tiber, 1985; and for general discussion of intelligence measures obtained from pediatric CI users, Knutson, Boyd, Goldman, & Sullivan, 1997). New research, however, has begun to explore the cognitive development of children with cochlear implants in greater detail and to investigate the specific information processing skills used in spoken language processing.

The focus on short-term memory in the present study was largely motivated by recent findings reported in Pisoni and Geers (1998) involving auditory digit span. Auditory digit span is a simple and widely used task that requires that the subject listen to a list of digits and then repeat back the list items in the correct order. Typically, the length of the list is increased over the course of a several trials until the subject can no longer do the task correctly. Pisoni and Geers showed that in a large group of pediatric CI users (N=43), simple forward digit span measures (administered live voice with lip-reading permitted, and requiring a spoken response during recall) were significantly correlated with open-set spoken word recognition scores even when the most obvious confounding variables were statistically controlled for (simple bivariate r = +.64 ; with variance from a test of speech feature discrimination removed r = +.37). Pisoni and Geers interpreted this finding as demonstrating the influence of "processing variables"—that is, "skills and abilities that have to do with the encoding, storage, retrieval and rehearsal of phonological representations of spoken words" (Pisoni & Geers, 1998, p. 342). Their findings on auditory digit spans were replicated more recently in a new sample of CI users (one of the two groups reported on in the current paper). Reanalysis of the pooled data set (N=88) showed that statistically significant correlations of at least +.30 still remained, even with a wholesale (albeit non-ideal) statistical "partialing-out" of the variability linked to slight age differences, communication mode, duration of deafness, duration of device use, age of onset of deafness, number of active electrodes, as well as a measure of auditory speech feature discrimination (Pisoni, 1999). Examination of the same data using the closely related procedure of factor analysis has led to a similar result (Geers, 1999). These findings on auditory digit span suggest that approximately 10% of the variance in open set speech perception measures may be accounted for by individual differences in the cognitive skills tapped by the forward digit span memory task, or perhaps some other co-varying but as yet unidentified variable. In making the case for the relevance of memory span measures, Pisoni (1999) has therefore proposed that an important cognitive processing variable related to how young children encode and manipulate the phonological representations of spoken words is contributing to the development of oral/aural language skills in pediatric cochlear implant users.

Surprisingly, relatively little other research has focused directly on short-term auditory memory in users of cochlear implants. Lyxell et al. (1996) have, however, examined whether for adult cochlear implant users it might be "possible to predict the level of speech understanding [post-implant] by means of a preoperative cognitive assessment." Since Lyxell had previously obtained results suggesting that individual differences in processing speed and working memory capacity could account for a portion of the variability found in the speech-reading/lip-reading scores of normal-hearing adults, he hypothesized that

this same relationship might also be observed in users of cochlear implants. Although Lyxell et al. reported negative results for the above study—that is, there was no evidence of working memory capacity serving as a strong predictor of speech reading ability—they have, as yet, only examined the case of post-lingually deafened adult cochlear implant users, and not pre-lingually deafened pediatric users of CIs.

In another, more clinically-oriented study of post-lingually deafened CI users, Gantz, Woodworth, Abbas, Knutson, and Tyler (1993) identified six preoperative measures to account for approximately two-thirds of the between-subject variability in sound-only open-set word recognition scores nine months post-implant. The Wechsler Memory Test, a battery of learning, memory, and working memory measures, though initially included in the battery of predictor variables, failed to exhibit any sizable correlation with the word recognition scores. Gantz et al. did however report that "the ability to extract information from sequentially arrayed signals and rapidly process that information as measured by the signal detection score from the Visual Monitoring Task, appears to be relevant to word understanding with an implant" (Gantz et al., 1993, p. 915). The visual sequence task used by Gantz et al. required participants to monitor a series of visually presented digits shown one at a time, for a specified subsequence—e.g., an odd-even-odd sequence of digits—thus requiring short-term storage of at least two items in recent memory during presentation (Knutson et al., 1991). In another paper by this research group on the skills of post-lingually deafened adults, the scores on this same visual monitoring task were again found to correlate between +.30 to +.40 with auditory-only measures of *phoneme* discrimination (Knutson et al., 1991).

At first glance it may appear somewhat surprising that a "visual" skill that seems minimally dependent on the verbal ability associated with spoken language would show any relation to subsequent gains made in speech understanding by hearing impaired listeners. One point that bears discussion however is whether hearing subjects typically approach the above visual monitoring task using non-verbal strategies--(i.e., do they tend to rehearse or keep a verbal tally using the spoken names of the items immediately prior in the monitored sequence—or can the visual patterns on the screen avoid "verbal mediation" en route to being remembered?). Here it is relevant to consider evidence that the skills tapped in such typical "visual monitoring" tasks develop somewhat atypically in at least some pediatric CI users as compared to normal-hearing children. Quittner, Smith, Osberger, Mitchell, & Katz (1994) used a task similar to Gantz et al.'s with normal-hearing and hearing-impaired children (cochlear implant and hearing aid users) ages 6 to 13 years. The participants were required to monitor a series of single-digit numerals on a screen and respond only when a certain specified sequence of two successive numbers was seen. No auditory stimuli were presented in this task. Both groups of hearing-impaired children did significantly worse than the hearing controls in the 6-8 years age range, and while this difference was not statistically significant in the older age range of 9-13 years (probably due to high within group variability), average performance was still worse in the hearing-impaired groups. Although no word recognition or language development outcome measures were reported for the Quittner et al. sample, these results suggest that some pediatric cochlear implant users (even those with several years experience with an implant) may encounter more difficulty than is typical for their age-matched hearing peers even when they are presented with a memory/attention task that *could conceivably* be performed on the basis of vision alone.

EXPERIMENT 1

The task employed in this report is neither the traditional verbal auditory digit span measure used in Pisoni & Geers (1998) (although digit span measure was also gathered for comparison purposes), nor the visual monitoring task used by Quittner et al. (1994) (though it shares some characteristics with this task). Instead, the "memory game" procedure used in this study involves the presentation of a sequence of sounds in conjunction with a sequence of colored lights located behind a series of four large translucent

buttons mounted on a response box modeled after the popular Milton Bradley electronic game "Simon™". (See Figure 1.) The task requires the child to immediately reproduce the target sequence by pressing on the appropriate buttons in the proper order thereby causing the synchronized sounds to be heard as the buttons are pressed. The difficulty of the task is adjusted by increasing the list length of the sequence to be reproduced when the child is doing well, and shortening it after an error is made. A computer program monitors the child's performance using an adaptive testing algorithm and records the longest list length a child is able to attain under a given set of conditions as a measure of "memory span". Further details can be found in Carlson, Cleary and Pisoni (1998) and in our Methods section below.

The motivation for using this particular response format was to obtain a measure of memory span using a non-linguistic manual response, rather than a spoken or signed verbal response as is used in traditional digit span tasks. A non-linguistic manual response was adopted to avoid confounds involving individual differences in rate and fluency of articulation and/or manual signing (i.e., productive language skills). The choice of target stimuli was determined by our interest in differences in memory span as a function of whether or not redundant auditory cues in addition to visual-spatial cues were presented as part of the target sequence. That is, we tested the effect of presenting auditory stimuli during a task that could also be performed on the basis of vision alone. We also manipulated whether or not the redundant sound stimuli were semantically congruent with the light sequence. In one condition, whenever a sound was presented, the auditory stimulus was the color-name of the currently illuminated button. In a second auditory-plus-visual-spatial condition, a spoken digit-name was arbitrarily assigned to each of the four response buttons and consistently presented whenever the matched button was illuminated (see Figure 1).

From a previous study by Cleary (1997), we knew that it was possible to obtain a measure of working memory span from normal-hearing adults using our proposed manual response format. In a later study we then attempted to use a similar procedure to compare preschool-age children's memory spans in three different conditions--when a light sequence was presented in conjunction with a sequence of nonsense syllables, in conjunction with a sequence of auditory digit-names, and alone, without auditory stimuli (Carlson, Cleary, & Pisoni, 1998). Although the procedure used was very similar to that of the current study, we obtained no significant differences between the stimulus conditions. However the children in the Carlson et al. study were 45 normal-hearing three- to five-year-olds, not all of whom were able to do the required task with any reasonable amount of facility. Additionally, we had not yet at that point introduced the new color-name condition, which we are now using to examine the effect of informational redundancy across two modalities of input.

Prior to conducting the current study, we had also established that children with cochlear implants could be reasonably asked to complete the task since we had piloted the procedure with a group of 19 pediatric CI users, ranging in age from 5; 7-11; 11 (Cleary, Pisoni & Kirk, 1998). The results from this previous study were, however, difficult to interpret due to the wide variation in age, experience with the implant and duration of hearing loss present within the sample. The pediatric CI users in the current study constituted a much larger and far more homogenous sample. Since all children in the current study had used a cochlear implant for at least four and a half years, we also judged it more likely that the mechanisms of phonological working memory might have developed at least partially in some members of this sample, than if relatively inexperienced users had been selected.

We believed that our new choice of stimulus materials, access to a relatively homogeneous CI group, and the use of an age group that was consistently able to do the task would permit us to examine the effects of interest. We expected to find improved performance as a function of multi-modal stimulus presentation to be less apparent in the hearing-impaired group as a whole, although we expected the CI

users with better open-set word recognition skills to behave more like the normal-hearing children in utilizing this redundant information. We were particularly interested in comparing the performance of

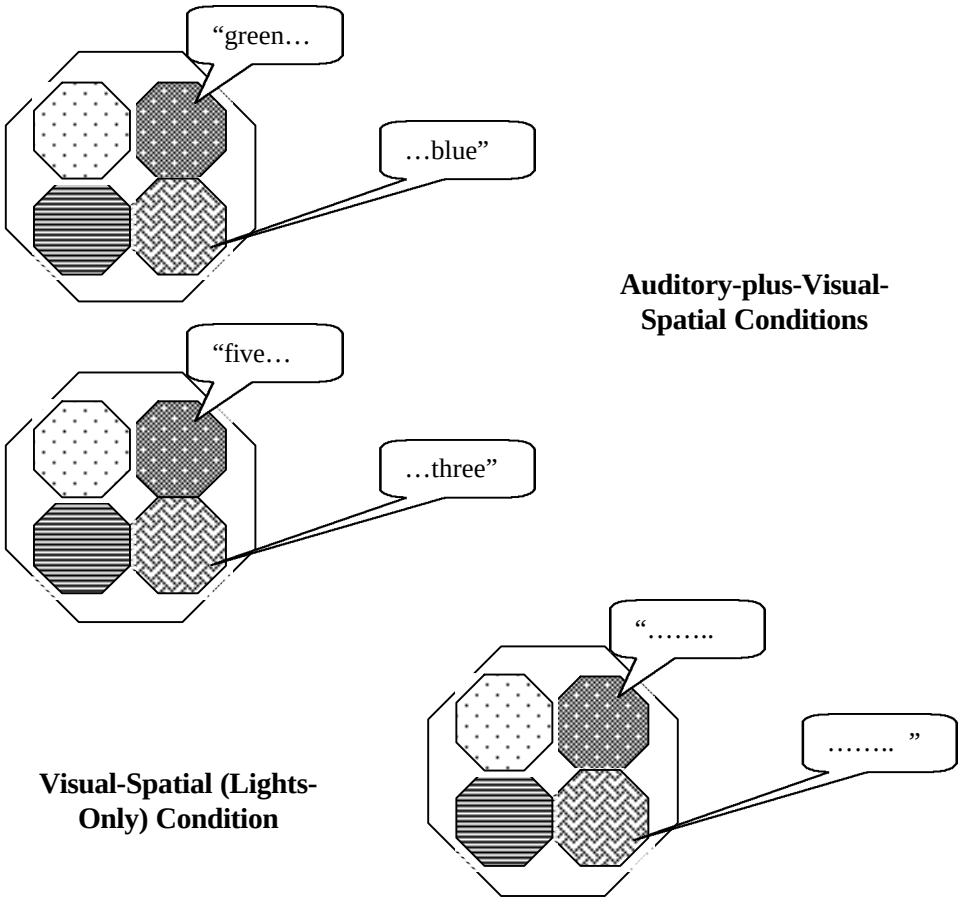


Figure 1. Diagram of memory game apparatus and experimental conditions given a list length of two items. Each shaded hexagon represents a large colored button back-lit by a light. All auditory stimuli were presented via a loud speaker located just behind the memory game response box. The verbal labels simply illustrate the consistent mapping between a particular auditory stimulus and a given button location.

the subset of cochlear implant users who were able to correctly discriminate/identify all the auditory stimuli when presented in isolation, with the normal-hearing group's performance on the memory game task. Although the previous research on visual monitoring in pediatric CI users was suggestive, the CI group was not necessarily expected to perform differently from the normal-hearing children on the *visual-spatial-only* version of the memory game task, unless the encoding of the button locations was perhaps attempted by one or both groups using a verbal encoding strategy.

A comparison of performance between traditional verbal digit span measures and the three different conditions of our memory game task in both groups of children was also planned. Since the traditional verbal digit span task makes no demands on visual-spatial aspects of working memory, smaller correlations between verbal digit spans and auditory-plus-visual-spatial memory game scores were

expected for those children who could be shown to be doing the memory game task by vision alone, and larger correlations were anticipated for those children who demonstrated use of the redundant auditory stimuli.

Method

Participants

Forty-five pediatric cochlear implant users were recruited as part of a large ongoing study by the Central Institute for the Deaf (CID). Participants were pre-selected by CID to demonstrate relatively low variability in chronological age (between eight and nine years of age) and device experience (at least four years of use) (see Geers, 1999 for details). The forty-four children who were retained in the final analysis evenly spanned the age range between 8 years, 1 month and 9 years, 10 months. Nineteen were female and 25 were male. The average age at onset of deafness ranged from 0 - 36 months. The mean age at onset of hearing-loss was four months of age with the majority of participants reported as congenitally deaf. Experience with the CI device varied between 4 years, 3 months, to 6 years, 9 months. The mean duration of CI use was approximately 5 years, 5 months. The duration of deafness prior to implant ranged from three months to five years, the mean being just slightly over three years. Each child's experience with oral vs. manual communication methods was quantified by determining the type of communication environment experienced by the child in the year just prior implantation and then each year over the four years of CI use prior to the current testing. A score was assigned to each year, with a "1" corresponding to the use of "total communication" with a sign emphasis (that is, extensive use of manual signs in addition to spoken language using an English, not ASL, grammar), and a "6" indicating an auditory-verbal environment with a strong emphasis on auditory communication without the aid of lipreading (see Geers et al., 1998 for details). Communication methods intermediate between these two extremes were assigned intermediate scores ranging from 2 to 5. These scores were then summed over the five points in time to produce a communication mode score ranging from 5 to 30. The mean summed communication mode score for the forty-four participants was approximately 19 on this scale. The cochlear implant devices used by the children contained anywhere between 6 to 20 active electrodes, with the group mean being 17+ active electrodes.

Forty-five normal-hearing (NH) children were also recruited for this study as a comparison group. The normal-hearing children were matched for gender and chronological age with the CID group. The mean age for both groups was 106 months (8 years, 10 months). A hearing screening was conducted for each child at 250, 500, 1000, 2000 and 4000 Hz at a level of 20 dB HL using a Maico Hearing Instruments pure-tone audiometer (MA27) and TDH-39P headphones. (A response at 25 dB HL was accepted for 250 Hz due to ambient room noise). Left and right ears were screened separately. Of fifty-three NH participants recruited, data from eight children were not included in the analysis due to one of the following reasons: the child failed the hearing screening, equipment malfunction, experimenter error, or a suitable age and or gender match was not among the hearing impaired CI group (some recruitment for this study began before the CID testing was complete).

Materials and Procedure

The auditory stimuli were created by recording a single male speaker of American English in a sound-attenuated, single-walled anechoic recording chamber (Industrial Acoustics Company Audiometric Testing Room, Model 402) using a head-mounted close-talking microphone (Shure, Model SM98). The talker was recorded saying the words "red," "blue," "green," "yellow," "one," "three," "five," and "eight"

in a clear voice at a moderate to slow rate of speech. Each word was spoken in isolation. The recordings were digitally sampled on-line at 22.05 kHz with 16-bit amplitude quantization using a Tucker-Davis Technologies (TDT) System II with an A-to-D converter (DD1), and a low-pass filter of 10.4 kHz (anti-aliasing filter, FT5), controlled by an updated version of Dedina's 1987 "Speech Acquisition Program" (Dedina, 1987; Hernandez, 1995). The amplitudes of the individually edited speech files were then adjusted using a digital leveling procedure such that the average RMS amplitudes for each file were approximately equated. All auditory materials were presented via a single loudspeaker (Advent AV280, 10 Watts amplifier output power, THD < 1%, frequency response 70 Hz-20 kHz) at approximately 70 dB SPL as determined via a hand-held sound level meter (Triplett Model 370) held at approximately the level of the child's head.

Presentation of the stimuli was controlled by a computer program specially created for this purpose, running on a PC computer. The response box used to collect the child's button presses consisted of a large round disk-like plastic case approximately ten inches in diameter housing four wide plastic buttons on its surface. The four buttons were each approximately one quarter of the surface area of the response box in size and could be easily depressed by a child. Each button was made of a different color plastic and could be illuminated by a light located beneath its surface. The colors of the buttons matched the color-names that were recorded as stimuli. The button response box was interfaced to the PC computer so that the control program could illuminate the lights when the sound stimuli were played, and dim the lights once the stimuli ceased outputting. The computer also recorded all button presses and automatically tracked the subject's performance using an adaptive testing procedure further described below.

The CI users were all tested by professional clinicians and researchers experienced in working with hearing-impaired children as part of a larger study being conducted by the Central Institute for the Deaf (see Geers, 1999). The normal-hearing participants were tested at the Speech Research Laboratory at Indiana University by graduate and undergraduate research assistants. All children were tested individually in a quiet room. Each condition of the memory game task took approximately four minutes to complete.

Identification Testing. Before the memory game was administered, each child was asked to identify the recorded tokens of the digits and color names by pointing to one of four numerals printed in large lettering on a piece of paper or one of four large colored squares. These tokens were presented one at a time through the same loudspeaker as used for the memory game. The digit-names and color-names were presented separately in two sets. If a child correctly identified all four items in a set on the first attempt, no further identification testing was administered. If one or more errors were made, the identification task was repeated up to three times, or until zero mistakes were made on a given set of stimuli, whichever occurred first. Twenty-one CI children identified all four digit-names correctly on their first try. Twenty-two CI children did the same for the four color-names. Identification of the stimulus "eight" appeared to pose a problem for many of the pediatric CI users. Eighteen of the forty-four CI children misidentified this item one or more times during this pretest. Errors on the set of color-names were randomly distributed across the set of four stimuli. None of the normal-hearing children misidentified any of the stimuli from either set during the same identification task. Regardless of identification errors made, the memory game task was administered next.

Memory Game Task. Participants were shown how the buttons on the response box could be pressed and were told that they would be hearing sounds through the loudspeaker and seeing the buttons light up. They were then instructed to "pay attention and copy exactly what the computer does by pressing on the buttons." A hypothetical example was given to insure that the child understood what was being of asked of them: "So for example, if you hear 'blue' ('one') and then 'green' ('three') (pointing consecutively

to two buttons), I want you press the blue button/this button and *then* the green/that one. Do you understand?"

Sequences used for the 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 very brief inter-stimulus interval of 200 ms was used between sequence items. However, since the individual stimuli had been recorded at a relatively slow speech rate, the rate of presentation was about 1.5 items per second. Each child started with a list length of one item. If two lists in a row at a given length were correctly reproduced, the next list presented was increased by one item in length. If on any trial the list was incorrectly reproduced, the next trial used a list one item shorter in length. This "adaptive tracking procedure" is similar to methods typically used in psycho-acoustic testing (Levitt, 1970). If the child made no response within four seconds after the last target item was presented, the sequence was presented again up to two additional times. The child was not told what rules were governing the presentation of the stimulus items. The strategy of waiting for re-presentation of the sequence (e.g., a second, or third time) was not permitted by the tester. The computer recorded when and if a child was permitted an extra repetition of the list. Very few trials involved such repetitions. Twenty unique trials were presented total, for a maximum possible list length of ten items. The experimenter provided no explicit feedback regarding the accuracy of the child's responses. Verbal encouragement was given only to keep the child on task during the procedure.

The independent variable of stimulus type (color-names vs. digit-names vs. silence/lights-only) was administered within subject. All children were presented with twenty lists to reproduce in each of these three conditions. The silent/lights-only condition was always completed last in the series of three conditions, but the two auditory-plus-visual-spatial conditions (color-names and digit-names) were counterbalanced across subjects. This procedure was followed based on practice effects evident in pilot testing, and a desire to not encourage purely visual-spatial encoding in a third of the children by introducing them first to the task in which no auditory stimuli were presented. Since the silent/lights-only condition was assumed to be the most difficult of the three conditions, placing it last in the series, when coupled with practice effects raising scores later in the testing session, can be reasoned to conservatively work against finding significant differences between the conditions.

WISC Digit Span. Children in both the CI and normal-hearing groups also completed the WISC Digit Span Supplementary Verbal Sub-test of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III) (Wechsler, 1991). This task requires the child to repeat back a list of digits as spoken live-voice by an experimenter at a rate of approximately one digit per second (WISC-III Manual, Wechsler, 1991). In the "digits-forward" section of the task, the child is required to simply repeat back the list as heard. In the "digits-backward" section of the task, the child is told to "say the list backward." An example of reversing a list length of two items is provided and a practice trial is given for the backward task. In both parts of the WISC task, the lists begin with two items, and are increased in length upon successful repetition until a child gets two lists incorrect at a given length, at which point testing stops. Points are awarded for each list correctly repeated with no partial credit. The WISC-III testing manual provides two lists for use at each list length, with the possible list length ranging from two to nine items in the forward condition, and from two to eight items in the backward condition. Items are not repeated consecutively within any list and each list is unique.

Although the CID research group gathered a large number of other language-related measures from the CI children as part of a larger study (Geers, 1999), only the short-term/working memory tasks (the memory game and WISC digit span) are reported here. Other analyses will appear in a future report.

Vocabulary Screening Measure. Finally, the Peabody Picture Vocabulary Test, a measure of receptive vocabulary (PPVT-III Form A, Dunn & Dunn, 1997) was administered to all normal-hearing children to obtain a general measure of verbal language development. In this task, a word is spoken by the experimenter and the child is asked to "point to the picture that means the same thing" from a selection of four different line drawings. The child begins the task hearing words known to be familiar to children in his/her age group, and the vocabulary is then adjusted in difficulty for the individual child until "floor" and "ceiling" word sets are established. Every NH child tested obtained a standardized score above one standard deviation below the mean.

Results and Discussion

One CI participant (98313) failed to complete the memory game task, reducing the number of participants in that group to 44. The matched hearing child (8251) was also dropped from the analysis. Our primary dependent measure for performance on the memory game task in each condition was the longest list length that the child was able to correctly reproduce at least once during that condition. However, we also computed the longest list length each child was able to correctly repeat on at least half the trials at that length for each condition, as well as the longest list length he/she correctly reproduced on *every* trial administered at a given length. After our initial analysis however, we became concerned that our planned scoring methods might be overly "coarse" and could be collapsing meaningful differences between individual children. Therefore, we also computed a "weighted" score for each child. This weighted score was calculated by finding the proportion of lists correct at each list length (e.g., the child correctly reproduced four of six lists presented at a list length of four items = 0.667), and this proportion was summed across all list lengths. (The use of an adaptive testing algorithm entailed that different subjects experienced a variable number of lists of any one given length.) Across all 88 children, this weighted scoring method correlates $r = +.95$ with the "at least once" measure, $r = +.93$ with "half the time" measure and $r = +.80$ with the "all lists at that length" measure. The weighted score is also more continuously distributed and results in a more normal-looking, less jagged distribution of scores--a necessary characteristic if correlational analyses are to be attempted. Figure 2 provides a comparison of results from the two groups using the four different scoring methods for the data. We will not consider the "half" or "always" scoring methods any further, but similar conclusions can be drawn from these alternative ways of analyzing the data.

An alpha level of .05 was adopted for all statistical tests reported below. A 2 x 3 mixed factorial repeated measures analysis of variance for hearing status (NH vs. CI) and stimulus condition (digit-names, color-names, lights-only) on the "at least once" memory game scores demonstrated a within-subjects main effect of stimulus type, $F(2, 172) = 7.02$, $p = 0.001$, and a between-subjects main effect of hearing status, $F(1, 86) = 23.41$, $p < .001$. Normal-hearing children obtained significantly higher scores than the pediatric CI users in every condition of the memory game task. No significant interaction was found between stimulus type and hearing status ($F(2, 172) < 1$, $p = 0.47$) indicating that the advantage of the NH children was not significantly greater in the auditory-plus-visual/spatial conditions than in the visual-only conditions. Post-hoc Bonferroni pair-wise comparisons between the three stimulus type conditions over both hearing status groups indicated a significant difference between the means for the color-name and silent/lights-only conditions (adjusted $p = .01$), and a marginally significant difference between the color-name and digit-name conditions (adjusted $p = .054$). No statistically significant difference was obtained between the digit-name and silent/lights-only conditions (adjusted $p = .69$). This same analysis using the weighted scores resulted in an analogous set of results: main effects of stimulus type ($F(2, 172) = 11.77$, $p < .001$) and hearing status ($F(1, 86) = 32.9$, $p < .001$), and no significant interaction ($p = .24$).

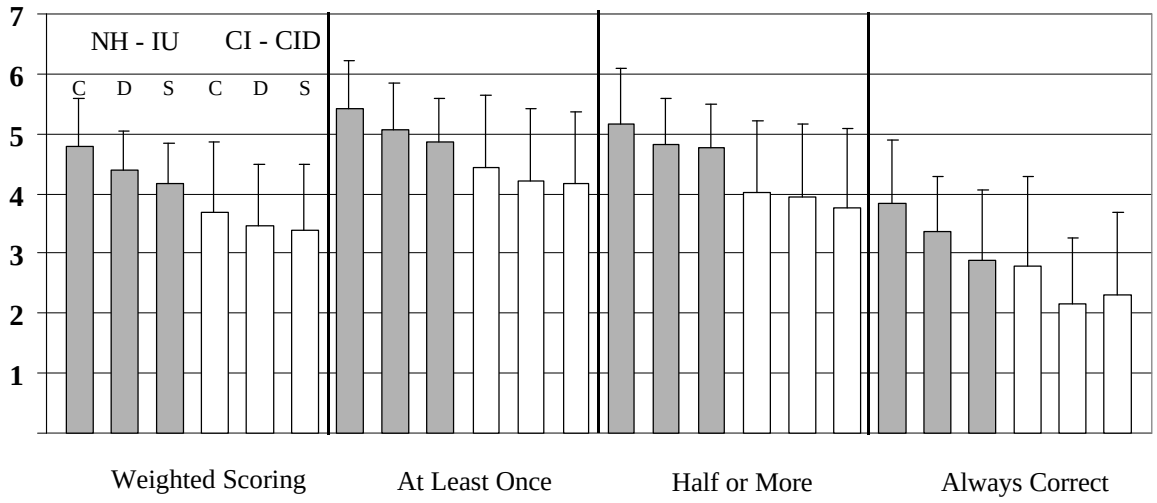


Figure 2. Comparison of performance on the memory game task across different stimulus types in normal-hearing and pediatric cochlear implant users. Hatched bars indicate the normal-hearing participants. Light bars indicate the CI users. "C" = color-names, "D" = digit-names, "S" = silent/lights-only. Heavy weight vertical lines separate the four different scoring methods described in the text. Error bars indicate one standard deviation.

A one-way repeated measures ANOVA and three pair-wise comparisons between the stimulus conditions were then conducted individually within each group of children (NH, CI) to determine if the main effect of stimulus type held up within each group considered independently. Using the "at least once" dependent measure, no significant effect of stimulus type was found for the CI group (p = .21). Using the "weighted score" measure, this effect still did not reach significance (p = .08). Within the normal-hearing group, however, the effect of stimulus type on the "at least once" measure was significant, $F(2, 86) = 6.34$, p = .003. Post-hoc Bonferroni pair-wise comparisons indicated that within this NH group, only the difference between color-name and silent/lights-only conditions was statistically reliable--(adjusted p = .002). The comparison between color-names and the digit-name conditions yielded an adjusted p = .11, between digit-names and silent/lights-only, adjusted p = .61. When the "weighted scoring" method was used, the difference between the color-name and digit-name conditions reached statistical significance (p = .026), otherwise duplicating the above results.

As expected, these results show that the normal-hearing children did obtain longer "spans" on average, when auditory stimuli were presented in conjunction with the visual-spatial light sequence. However, this advantage was only statistically reliable when the verbal cues were "semantically redundant" with salient characteristics of the spatial-locations to be remembered. That is to say, merely providing a consistently matched auditory stimulus (i.e. the spoken digit-names) did not provide reliable benefit. Examination of Figure 2 suggests that while the differences in performance among the stimulus conditions followed the same overall pattern in the CI group, the within group variability among the pediatric CI users was too large for the stimulus type manipulation to yield reliable differences. Children in the CI group did not appear to use the informationally redundant auditory cues as effectively as the normal hearing children.

We had, however, naturally expected that the CI children who were obtaining lesser benefit from their implant would be less able to utilize the supplementary auditory information during sequence presentation and would therefore do similarly on all three memory game conditions. To test whether the stimulus type manipulation had any differential impact depending on the children's level of auditory speech perception skill, the CI group was subsequently divided in two subgroups according to whether or not errors were made on the stimulus identification pretest. As shown in Figure 3, *no* significant differences in memory game performance were found between the two groups. The small differences obtained between means were not even consistently in one direction, although somewhat greater variability was evident within the CI group that made one or more identification errors on the pretest. We were surprised by this result, since we had had strong expectations that the CI users who had better word identification skills would behave more like the normal-hearing children. We then conducted two other splits of the CI group that we thought should also yield the expected pattern. The first split was by number of color-name identification errors (zero vs. one or more errors). No difference was found between the groups. In the second split we sorted the CI users into two groups using an independent measure of open-set word recognition ability. The CID research group had obtained these open-set word recognition scores from the children for a separate purpose within about a week of the memory game testing. Children that scored above 40% correct on the Lexical Neighborhood Test "Easy" word lists (Kirk, Pisoni, & Osberger, 1995) were placed in one group, and those that scored below 40%, in the other. Once again, contrary to expectation, the two CI groups created in this manner performed equivalently on the memory game task regardless of stimulus type.

This pattern of results suggested to us that most of the CI children who were *able* to consistently identify the memory game speech tokens were *not using* auditory information to do the memory game task. That is to say, despite the fact that a subset of the pediatric CI users had fairly good auditory-only word recognition skills, these children were performing our memory game task in a manner more similar to the CI users with poor word recognition skills, than to our age-matched normal-hearing controls. We suspected that this result might have to do with the visual/spatial component that is present in all conditions of the memory game task. To further test this notion, we retained the split of the CI users into two groups according to each child's identification of the four digit-names: zero errors (n=21), one or more errors (n=22). We then compared the performance of the two groups on the verbal WISC digit span measure, a task that includes little to no visual-spatial component. This memory span measure is, however, scored in a different manner than our memory game measure: the child is tested on two lists at each list length starting at three items, with the list length increasing after every two trials until both lists at a given length are incorrectly repeated, at which point, testing stops. A point is awarded for every list correctly reproduced. Thus, this score cannot be directly equated with a "list length" measure of "span capacity." The two right-most bars in Figure 4 show the obtained difference in scores between the two CI groups. A t-test for independent groups showed this difference to be statistically reliable ($t(41) = 2.57$, $p = 0.014$).

A difference between the two CI groups is thus clearly evident when the memory task used is primarily auditory and requires verbal coding (i.e., the WISC digit span task), but not when performance on the memory game task (with its visual-spatial component) is examined. This result is consistent with the fact that although significant correlations existed in this CI group between communication mode and WISC verbal digit spans (Pisoni & Geers, 1998),[5] these correlations were not replicated with any of our memory game measures and communication mode. All correlations computed between the memory game task conditions and communication mode were in fact, negative, though negligible in size. (Recall that high communication mode scores reflect a more oral communication emphasis, and lower

[5] The correlations between communication mode and forward WISC digit span were as follows: simple bivariate r = +.42; with speech feature discrimination and age partialed out, r = +.30; with speech feature discrimination, age and open set-word recognition scores partialed out, r =+.12.

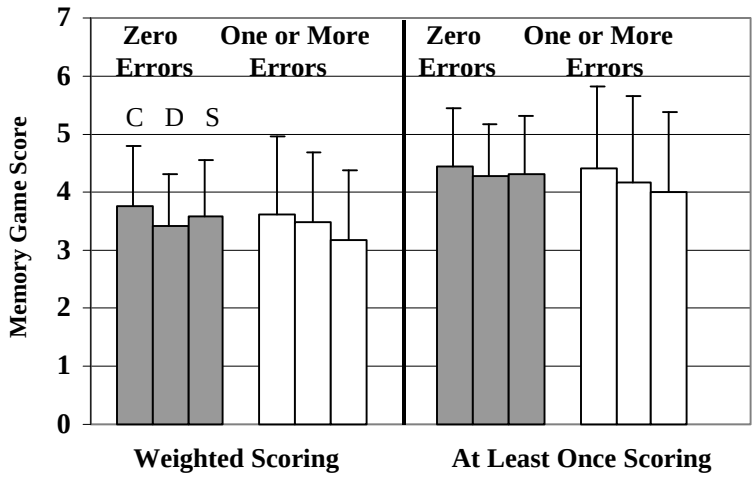


Figure 3. Performance on the memory game task in the group of pediatric cochlear implant users split by number of errors made in the digit-name identification pre-test. Striped bars indicate mean spans for the twenty-two children that made no errors. White bars indicate mean spans for the twenty-two children that made one or more errors. "C" = color-names, "D" = digit-names, "S" = silent/lights-only. The dark vertical line separates two different scoring methods as described in the text. Error bars indicate one standard deviation.

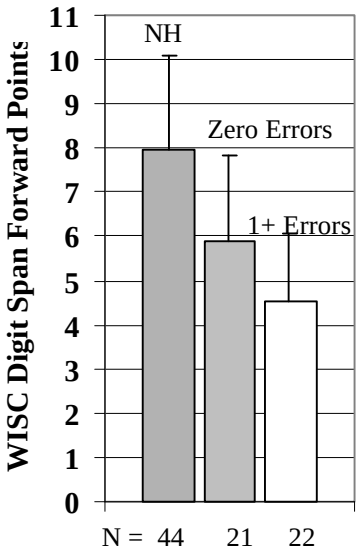


Figure 4. A comparison of the WISC forward verbal digit spans across groups. NH = normal-hearing. The CI group is subdivided into two sub-groups depending on whether zero vs. one or more errors were made identifying the four digit-names used in the memory game task. One CI child (98114) did not complete the WISC verbal digit span task although memory game spans were obtained from this child. This child's data is not included in Figure 2. Error bars indicate one standard deviation from the mean.

communication mode scores, a more manual-sign oriented emphasis.) This result indicates that greater experience with oral/aural methods of communication did not predict improved performance on the memory game conditions involving auditory stimuli.

Table I lists the intercorrelations between the children's scores in the three different within-subject conditions of the memory game task. The observed high intercorrelations between the different conditions of the memory game would be expected in the CI group, if it were the case that the children in this group were approaching each of the three conditions in the same manner using a visual-spatial coding strategy regardless of whether the additional auditory component were present. On the other hand, the normal-hearing children demonstrate low correlations between the memory game conditions, suggesting the use of different strategies in each condition. This proposal is somewhat weakened by the lack of substantial correlation between the two conditions using auditory stimuli. The data indicate, however, that although the normal-hearing children found hearing the color-names to be helpful, the digit-names produced less of a redundancy gain because they were arbitrarily, not semantically, paired to each colored button location.

Table I also shows that the WISC verbal digit span measure in general showed little correlation with the memory game measure. This lack of correlation cannot be attributed to lack of range in the scores, because the "points" method of scoring the WISC forward span task, and the weighted method used to score the memory game task are both normally distributed enough to permit correlation calculations. The only indication of a sizable positive correlation is between WISC forward digit span and the memory game color-name scores in the normal-hearing group – probably the only set of conditions under which verbal strategies contributed measurably to performance on the memory game task. The failure to find this same pattern of correlations among the CI participants, taken together with this group's failure to demonstrate a redundancy gain in the color-name condition, suggests that the pediatric CI users tended to use primarily visual-spatial cues to perform the memory game task.

Table I also includes a column that lists the correlations of the memory tasks just discussed with WISC *backward* digit span (WISC-BPTS). This task requires the child to reverse the order of the numbers when recalling the presented list and is scored in the same manner as the forward digit span measure. The common theoretical interpretation of the backward span task is that since it requires an effortful strategic manipulation of the presented items, it is a better measure of "working memory" than the immediate forward repetition digit span task (Lezak, 1995). Table I shows that although there are consistent small to moderate correlations between the memory game task conditions and WISC backward digit span, there is almost zero correlation between the memory game task conditions and WISC forward digit span in five of the six correlations computed between these measures. This pattern of correlations suggests that the memory game task has more in common with the backward WISC digit span task than with the forward digit span task. This certainly appears to be true in the case of the CI users, and less strongly so, in the normal-hearing group as well. We speculate that this may have to do with component of the reversal task that could be described as "spatial manipulation" of the items in order to effect the reversed list. Perhaps this shared visual-spatial aspect is responsible for the larger correlations between the memory game and the backward rather than forward verbal digit span task in these children.

From the results thus far presented, then, several findings emerge that may have general implications beyond the circumstances of our particular study. Firstly, the normal-hearing children demonstrate a redundancy gain--longer memory spans for auditory-plus-visual-spatial sequences than for visual-spatial-only sequences--when the auditory stimuli bear semantic relevance to characteristics of the matched button locations. When the auditory stimuli are paired in an unfamiliar, seemingly arbitrary way

Table I.

Intercorrelations between WISC forward digit spans (WISC-FPTS), the three memory game conditions ("weighted" scoring), and WISC backward digit spans (WISC-BPTS).

	Cochlear Implant Group, N=43					Normal-Hearing Group, N=44				
	WISC-FPTS	**Colors**	**Digits**	**Silent**	**WISC-BPTS**	**WISC-FPTS**	**Colors**	**Digits**	**Silent**	**WISC-BPTS**
WISC-FPTS										
Colors	.09					.58**				
Digits	-.02	.66**				.08	.16			
Silent	-.02	.71**	.64 **			.01	.20	.09		
WISC-BPTS	.52**	.46**	.23	.48**		.32*	.36*	.17	.19	

Note: *$p < .05$. **$p < .01$. Although age in months has been partialed out of the calculations, age showed only a small positive correlation in the NH group and a near-zero correlation in the CI group, making the simple correlations almost identical to those in Table I.

to the spatial locations, even normal-hearing children show little benefit from having multi-modal sources of input. From these results we are also led conclude that within the CI group our memory game task is yielding measures primarily of visual-spatial memory span. If this is the case, however, we still need to account for the fact that the hearing-impaired children as a group did less well than the normal-hearing children, not just on the conditions of the memory game task that utilized auditory stimuli, but also on the visual-spatial "lights-only" condition. This is an interesting and somewhat unexpected finding. We discuss this point further below and offer some tentative explanations why this result was obtained.

Gender Differences in Performance. Next we briefly report one analysis that was conducted post-hoc stemming from an inquiry we received regarding possible gender differences in the development of verbal vs. visual-spatial short-term memory skills. In previous research, males have been reported to show an advantage over females on measures of spatial memory span (e.g., Grossi, Orsini, Monetti, & De Michele, 1979). It has also been claimed that female children tend to have more advanced verbal skills compared to their age-matched male peers (Kramer, Delis, Kaplan, O'Donnell, & Prifitera, 1997), with the origin of this difference being hotly debated. Several studies however, report *no* advantage of females over males on the traditional verbal digit span task (e.g., Grossi et al., 1979).

No significant gender differences were found in either group on the memory game tasks, although in the CI group the males did somewhat better on all conditions of the memory game (see Table II). There was, however a significant difference between the male and female groups overall on the WISC verbal forward digit span measure, with the females (n=37) scoring significantly higher than males (n=50), t(67.596--adjusted df, equal variances not assumed) = 3.06, $p < .01$. This gender difference shows up independently in both the NH and CI groups. Note that this result differs from the usual lack of difference reported in some of the literature previously cited.

Table II.

Memory Game mean scores as a function of Gender and Hearing Status. The normal-hearing group's Vocabulary scores (PPVT-III) are also shown.

		Female		Male	
Normal-hearing		Mean	SD	Mean	SD
Vocabulary	Vocab. Raw Score	130.74	12.48	131.96	18.18
	Vocab. Standard Scr.	111.63	9.45	112.40	14.21
Memory Game Weighted Scoring	Color-names	4.88	1.02	4.72	.63
	Digit-name	4.45	.75	4.36	.57
	Silent/Lights-Only	4.15	.65	4.16	.75
Memory Game At Least Once Scoring	Color-names	5.47	.96	5.36	.70
	Digit-name	5.05	.85	5.08	.76
	Silent/Lights-Only	4.84	.69	4.88	.78
CI Users					
Memory Game Weighted Scoring	Color-names	3.42	1.21	3.89	1.15
	Digit-name	3.34	1.07	3.54	1.02
	Silent/Lights-Only	3.08	1.19	3.62	.98
Memory Game At Least Once Scoring	Color-names	4.21	1.27	4.60	1.15
	Digit-name	4.11	1.24	4.32	1.18
	Silent/Lights-Only	3.79	1.18	4.44	1.16

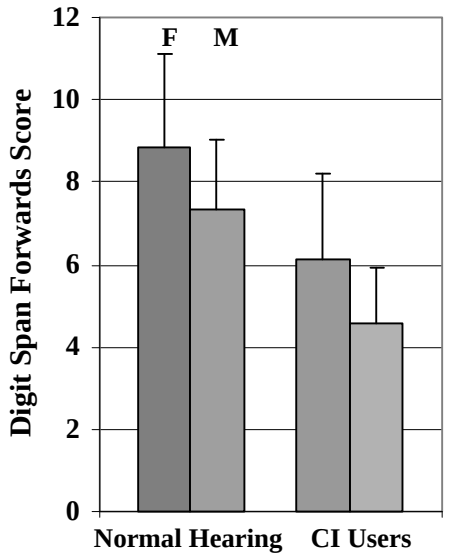


Figure 5. Differences in Forward WISC verbal digit span (scored by points) as a function of gender ("F" = female, "M" = male) and hearing status.

Figure 5 displays the main effects of gender and hearing status, and lack of any interaction. An overall age difference between the male and female groups can be ruled out, because the means and range of ages for all four sub-groups (gender x hearing status) are virtually identical. (The groups are however, unequal in size.) As can be seen in Table II, the standardized vocabulary scores obtained for our NH group showed no difference as a function of gender. This may have to do with how the PPVT-III is constructed—namely to avoid the inclusion of gender-biased test-items. Although we think it is worth mentioning these findings here, the gender-related results require further study before any firm conclusions can be drawn since they were conducted entirely post-hoc and could well be due to chance. The one conclusion that we feel *can* be safely drawn from these results, however, is that once again, we have evidence that results using our memory game task fail to resemble results obtained using an auditory-only short-term/working memory task such as WISC forward digit span, since only the latter showed clear gender-based differences.

Conclusions

In summary, we were able to show in Experiment 1 that normal-hearing school-age children were sensitive to the informational redundancy present in the color-name condition of the memory game task. The hearing-impaired pediatric CI users that participated in our study did not reliably demonstrate this same sensitivity. Contrary to expectation, the CI users that demonstrated good open-set word recognition skills failed to behave similarly to the normal-hearing comparison group. The data also indicated that the hearing-impaired children did less well than hearing children even on a condition of the memory game task that provided only visual-spatial target sequences to be remembered.

From the results obtained from the CI group, we realized that the large visual-spatial component of our memory game task might endanger its utility as a means of obtaining measures of auditory short-term/working memory in CI users. We therefore decided to explore ways of making the memory game task more "auditory" in nature without having to dispense with the manual response method. To address this issue, a subset of our normal-hearing subjects were asked to complete an additional condition of the memory game task, as described below in Experiment 2.

EXPERIMENT 2

In order to try to discourage purely visual-spatial processing of the target sequences, a subset of the normal-hearing participants tested in Experiment 1 was asked to play an additional "round" of the memory game in which sequences of color-names were played through the loudspeaker. This time however, the auditory stimuli were presented *without* the lights flashing synchronously on the memory game response box. Thus, on a given trial, a child might hear the list "red, green, blue, red," output via the loudspeaker, but no lights would be presented on the response box as the target sequence was played. The child would then have to manually reproduce the sequence by pressing the appropriately colored buttons on the response box.

We planned to compare memory spans obtained in this condition to spans obtained when lights *were* presented at the same time as the color-names. Lower spans were expected for the "auditory-only" condition. This finding, if obtained, would provide additional demonstration of the informational redundancy present in the original color-name condition used in Experiment 1.

Method

Participants

Twenty-seven of the original forty-five normal-hearing subjects participated in this condition. This task was added after the data from the CI users in Experiment 1 had been collected, and the collection of data from the matched group of normal-hearing children had already begun. Thus, this task was not planned as part of the experiment proper, but rather as a procedure to be piloted for future use (as will be reported in Experiment 3).

Materials and Procedure

On each trial, the target sequence consisted of an auditory list of color-names presented without illumination of the colored lights. The mapping between auditory stimulus and spatial button location was thus made less visually salient during the target list presentation. The child's task was still to reproduce the sequence by pressing the color-matched buttons in the proper order. Both sound and light were initiated by the subject's button press responses in order to convey that each press had registered and to provide the subject with the same minimal amount of feedback as had been present in the original task.

The administration of this task was not counterbalanced across subjects, but was always run after the three memory game conditions described had been completed, following a short break. Admittedly, due to the lack of counterbalancing, it is possible that any differences obtained might be artifactual.

All hardware and stimulus materials were otherwise identical to those used with the normal-hearing children in Experiment 1. The child was told before starting that "you will not see the lights light up this time, so listen carefully, and copy what you hear by pressing on the buttons just like before."

Results and Discussion

Figure 6 (left panel) shows the mean span for the new "audio-only" version of the task ("A") plotted next to the mean spans from Experiment 1 for color-names in the auditory-plus-visual-spatial condition ("A+V-S") and the visual-spatial-only condition ("V-S"). This comparison suggests that although the sound-only condition means are reduced, the normal-hearing children are clearly able to complete the task without the visual stimuli present in the target sequence. A paired-samples t-test on the weighted scores in the "A" vs. "A+V-S" conditions approached significance ($t(26) = 1.82$, $p = 0.08$). Figure 6 also shows that the normal-hearing children's mean performance for the "audio-only" condition was actually quite similar to this group's mean span in the visual-spatial-only condition. Since the conditions were not counterbalanced in administration, the conclusions we can draw from this finding are limited. We suggest however, that this result provides some further evidence that the difference in performance shown by the normal-hearing children in the single-modality vs. multi-modal cue conditions used in Experiment 1 was, in fact, due to the presence of redundant information in the later condition. Use of only a single modality of presentation appears to reduce the NH children's performance on this task by about the same amount, regardless of which modality is omitted.

From this last set of results, we were eager to try the "audio-only" version of the memory game task with a sample of pediatric CI users. In order to reproduce the sequence of color-names, the CI users would be forced to encode the auditory stimuli since no visual-spatial information would be available.

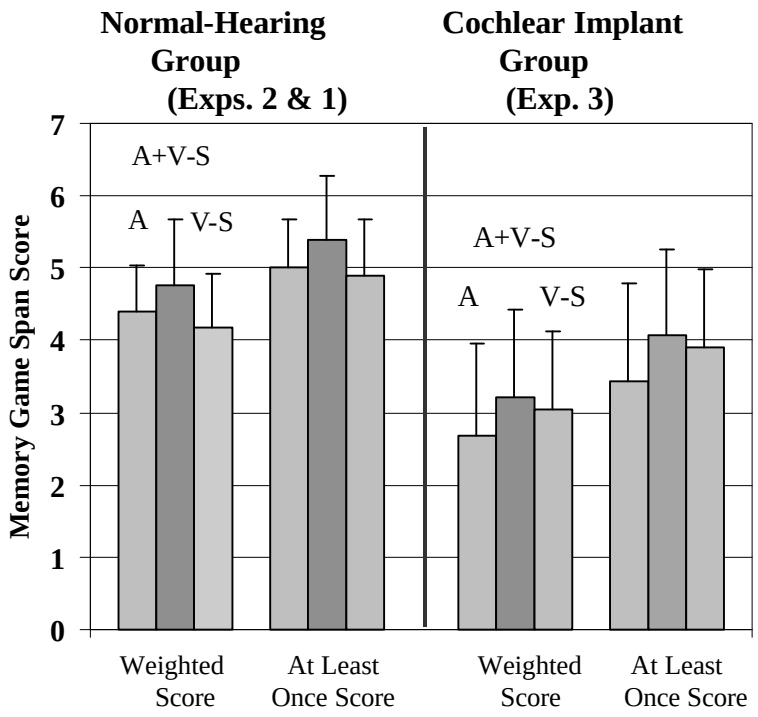


Figure 6. Comparison of group means for the audio-only version of the memory game using color-name stimuli ("A"), with the original version of the memory game using both sounds and lights ("A+V-S"), and the silent/lights-only visual-spatial condition ("V-S"). The left panel contains results relevant to the discussion of Experiment 2 and contains data repeated from Figure 2 (Experiment 1). The right panel presents results relevant to Experiment 3, to follow. Error bars indicate one standard deviation from the mean.

EXPERIMENT 3

Experiment 1 showed that the memory game task in its original form provided measures of primarily visual-spatial memory span from the pediatric CI users rather than tapping mechanisms for auditory short-term/working memory. Experiment 3 gave us the opportunity to confirm this finding in a new sample of 28 eight- and nine-year old users of cochlear implants, but also, more interestingly, to test their performance using the audio-only version of the memory game task as introduced in Experiment 2.

Methods

Participants

Twenty-eight hearing-impaired pediatric users of cochlear implants participated in this study. None of the children in the current experiment took part in Experiment 1. All children were participants in a larger study currently being conducted at the Central Institute for the Deaf. The backgrounds of these children were similar to the pediatric CI users reported on in Experiment 1. Thirteen female and fifteen male children completed the task. The children ranged in age between 8 years 2 months, to 9 years 11 months. The mean summed communication mode score for the twenty-eight participants was approximately

19 on the scale described in Experiment 1, thus once again indicating a group tendency slightly towards more oral communication methods.

Materials and Procedure

At time of writing, only the memory game data from these children are available. Other measures were also collected from this group but will be reported on in a future paper.

Identification Testing. The stimulus identification pre-test was run as described in Experiment 1 except that each child was only asked to identify each stimulus once. Twelve of the twenty-eight children made at least one error on the digit-name identification task. Five children made one or more errors on the color-name identification task. Fourteen of the twenty-eight children that participated identified all eight stimuli correctly.

Memory Game Task. The stimulus materials and hardware used for this study were the same as in Experiments 1 and 2, except that a different PC computer was used to run the program controlling the stimulus presentation. Due to a problem with the setup of this second computer, on some trials, the computer registered a double press when the child pressed a given button for too long. There was a risk therefore of the child being inappropriately penalized for such trials. In tabulating the data collected using this setup, we counted the number of times each child appeared to have inappropriately registered a double press. On average, for each child, this occurred on about two of twenty trials per condition. Some children never encountered this problem. Due to the liberal way in which scores were tabulated, the effect of this technical problem on the "at least once" method of scoring was quite small. The audio-only condition was not affected by this problem and scores on this condition were not similarly influenced.[6] As with the normal-hearing children, the audio-only condition was run last, after all other conditions of the memory game task had been administered.

Results and Discussion

For this new sample of CI users, once again, no significant differences were obtained between the original memory game task conditions (color-names, digit-names, vs. silent/lights-only) although the group means were in the predicted direction (color-names > digit-names > silent/lights only). The group of fourteen children that made zero errors on the identification task did not benefit any more from the redundant auditory information than the children who made one or more identification errors. The correlations among the three original memory game measures were again found to be quite high, as might be predicted if the same strategies were being used in all three tasks. From this result we can be fairly certain that most of the CI users, even those with the ability to identify the auditory stimuli, were not using this redundant auditory information to aid them in doing the original memory game task. Instead, what we obtained from the original task was a measure of visual-spatial short-term/working memory span.

More interesting was the group's performance on the auditory-only color-name condition. Many of the CI users proved able to do this new task with some facility. The right-hand panel of Figure 6, above, provides a comparison of the CI group's performance in the auditory-only color-name condition as compared to the auditory-plus-visual-spatial condition using color-names, and the visual-spatial lights-only condition. (Note that the aforementioned technical problem could only have increased the difference

[6] Since the technical problem could only have resulted in lower-than-normal span scores for the original set of stimulus conditions, the difference obtained between the original audio-plus-visual-spatial conditions vs. the new audio-only condition would probably have been somewhat larger if this problem not occurred.

between the bar on the far left of each cluster ("A") as compared to the two bars to the right of each cluster ("A+V-S" and "V-S").) The data shown in Figure 6 indicate that the hearing-impaired children did less well on the auditory-only condition of the memory game. Their memory game spans in this condition were quite reduced as compared to the conditions in which they were provided with visual-spatial cues: planned comparisons (paired-samples t-tests) between the audio-only condition and the audio-plus-visual-spatial condition using color-names yielded a $t(27) = 2.17$, $p = .04$, and between the audio-only condition and the lights-only condition, $t(27) = 1.61$, $p = .12$. Moreover, unlike Experiment 1, those children who had made errors during the stimuli identification pre-test (n=14) did significantly less well on the audio-only version of the memory game task than children who had made no errors (n=14) (t(21.1-equal variances not assumed) = 2.11, $p = .047$). These results provide evidence that it is feasible to use this manual response task with a CI population with the assurance that it will measure phonological working memory span to some degree rather than purely visual-spatial memory span. While it is true that the audio-only condition as described here does not necessarily preclude the child from looking at the appropriate buttons on the response box as the auditory stimuli are played, the elimination of the lights clearly made the task more difficult than when both modalities of input were provided, or when just the visual-spatial input was provided.

An additional analysis was conducted to try to shed further light on whether the revised task truly utilized the desired modality. Since Pisoni and Geers (1998) had reported significant correlations between WISC forward digit span and communication mode, as discussed previously, we computed simple correlations between communication mode and each of the memory game tasks. Again, as in Experiment 1, we found negligible to mildly negative correlations between each of the three original memory game tasks and communication mode. (Recall, once again that high communication scores are associated with more oral/aural communication environments.) The new audio-only variant of the memory game task, on the other hand, showed a small but positive correlation with communication mode ($r = +.30$, $p = .11$). Rather surprisingly, this correlation is actually stronger when only those children who made no errors on the identification task (n=14) were considered alone ($r = +.50$, $p = .06$). From this result we conclude that the auditory-only version of the memory game task can be used to assess auditory short-term/working memory span in a manner similar to the WISC digit span task. The memory game task in this form may in fact, be preferable in some cases, as it avoids the articulatory component inherent in the WISC digit span response format.

Finally, as a point of interest, we present in Figure 7, a profile of the individual subject scores for the auditory-only memory game task for both the normal-hearing subjects in Experiment 2 (N=27) as well as the cochlear implant users in Experiment 3 (N=28). These groups are only approximately matched for age and gender, but there is an interesting conclusion that we feel can legitimately be drawn from this figure. Note that despite their profound deafness as infants and subsequent use of a cochlear implant, the top one-third of the CI children in the distribution of white bars are demonstrating auditory memory spans *equivalent* to the lower one-third of children in the normal-hearing distribution (dark bars). Plotting the WISC forward digit spans of the two groups of children tested in Experiment 1 yields a very similar graph with about the same amount of overlap in scores between the CI and normal-hearing groups. This is a considerable achievement on the part of the pediatric CI users, in light of the circumstances.

General Discussion

One of the long-term goals of our current research program is to develop a practical methodology for assessing and tracking the development of verbal working memory in pediatric cochlear implant users over time. Our interest in this area partially stems from recent debate in the area of cognitive development

having to do with the role of working memory in language development and lexical acquisition (Baddeley, Gathercole, & Papagno, 1998; Gupta & Dell, 1999). The results presented

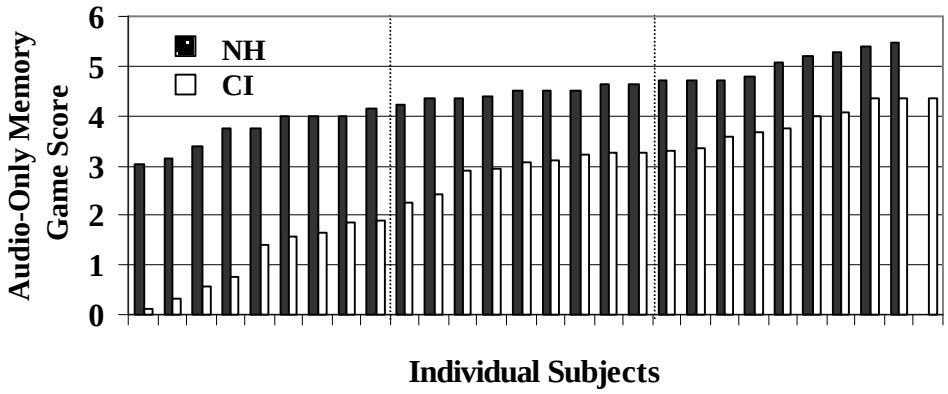


Figure 7. Individual subject data for all children who completed the audio-only version of the memory game task. Normal-hearing children are shown by the dark bars, pediatric CI users using clear bars. The scores for each of the two groups of children have been sorted in rank order of performance. The memory game scores were tabulated using the weighted method as described in the text.

in this paper address some methodological issues regarding the assessment of short-term/working memory in a special population of children for whom the auditory modality has been partially compromised. We conclude that given the opportunity to utilize an intact sensory modality, even experienced school-age users of a cochlear implant will not "automatically" make use of auditory information and verbal rehearsal that might aid them in performing a short-term memory task. That is to say, unlike normal-hearing children, the pediatric cochlear implant users we tested showed little evidence of "integrating" semantically related auditory and visual/spatial stimuli in order to complete the original version of the memory game task. This suggests that fundamentally different sensory coding and/or rehearsal processes may exist in these two populations. Although the cochlear implant is now providing the CI children with access to sound and spoken language, the atypical early experiences of these children are still evident in how they perceive and encode sensory information.

The argument that the cochlear implant users failed to make use of the auditory information simply because it was inaccessible to their sensory systems must account for the results reported for the auditory-only version of the memory game task. Given only the auditory information, most of the cochlear implant using children were able to do reasonably well on the task, as evidenced by a direct comparison to the distribution of scores obtained from normal-hearing children of the same age. This finding supports a tentative account that the CI users were less likely than NH children to draw benefit from the multi-modal form of stimulus presentation due to idiosyncratic habits of information processing and sensory integration across modalities. These differences in processing informationally redundant signals may have developed as a result of the early auditory deprivation experienced by the pediatric CI users.

The CI children reported on in this study performed significantly less well as a group overall than the normal-hearing children even when the memory game task utilized only visual-spatial stimuli. This was somewhat unexpected as the literature regarding the development of short-term memory in hearing-impaired populations suggests that the performance of deaf children on span tasks lacking a verbal component should not be impaired relative to that of normal-hearing children (Furth, 1966; Mayberry,

1992). When differences are, in fact found, investigators tend to find that the visual/spatial stimuli used in the particular task lent themselves to linguistic labeling. When linguistic labeling is possible, hearing-impaired children appear to be at a disadvantage relative to their normal-hearing counterparts (see discussion in Mayberry, 1992). Various studies have also suggested that hearing-impaired children who have grown up around spoken language sometimes attempt rehearsal/encoding strategies used by normal-hearing persons, involving self-generated verbal labels for even visual-spatial stimuli.[7] Since the verbal skills of hearing-impaired children with oral/aural backgrounds are often not as fully developed as normal-hearing children their age (i.e., with regards to processing speed, efficiency, robustness, etc.), their performance on a memory span task that could be approached in terms of even self-generated verbal labels might be reduced relative to that of NH children. It is possible that this occurred in the Quittner et al. (1994) study involving visual monitoring of a stream of orthographically presented digits discussed in our introduction. This situation could also have arisen in the current study—the CI users may have attempted to encode the sequence of lights in the "visual-spatial-only" condition using verbal names for the colors being illuminated even though the auditory stimuli were not explicitly presented.

We acknowledge this as a possible explanation for the reduced spans of the CI group in the visual-spatial-only condition, but if this relatively sophisticated verbal approach was truly being attempted, we would have expected the effect of informational redundancy to have had greater impact in the CI group. The fact that no significant differences were found as a function of stimulus-cue type using the original version of the memory game task strongly suggested to us that the CI group carried out all three conditions as purely visual/spatial tasks. Many of these CI children *were* capable of completing auditory versions of both the WISC digit span task and the modified memory game, and yet these same children did not perform any differently on the original memory game than the CI children that made errors in simply identifying the number names. This result was surprising and again suggests that the artificially imposed time synchrony between the light presentation and sound presentation was simply ignored by a majority of the CI children. This did not occur with the normal-hearing children in these same tasks.

In summary, the present results suggest that even those CI children who are able to accurately identify speech signals as they are heard, may not have phonological working memory mechanisms or processing strategies that are developed to a point equivalent to chronologically age-matched normal hearing counterparts. This outcome would not exactly be surprising, as many important milestones in the development of speech perception and memory are reached during the first two years of life. Despite their prelingually-deafened status, most of the CI users reported on here received their implant at a point in time when the FDA did not permit implantation of children under two years of age. Additionally, since the implantation procedure requires that candidates show a demonstrated failure to benefit from conventional hearing aids, we can be fairly certain that most of these eight- and nine-year-old children were without any sensory input from the auditory modality for one quarter to one third of their lives. It should not be surprising, then, that the encoding strategies and working memory mechanisms of pediatric CI users *are*, in fact fundamentally different from those of normal-hearing children. Ongoing research in our lab is attempting to answer the question of *how* these coding/rehearsal mechanisms differ, and what kind of developmental changes can be observed or effected in these children. Increasingly, clinicians are beginning to see pediatric CI users that have reached ceiling levels of performance on traditional standardized measures of speech perception and spoken word recognition that are usually used with this population--and yet these children are still clearly having problems with reading and other more advanced language skills that are based on listening, phonological encoding, and other metalinguistic abilities. Further investigation of how pediatric cochlear implant users engage in cognitive processing of information originating from this

[7] In contrast to fluent users of a sign language as their first language, who many researchers report as having normal to above average memory for visual-spatial sequences (see reviews in Mayberry, 1992).

reintroduced sensory input modality may help us provide new evaluation and treatment techniques (Pisoni, 1997). Eventually we would like to settle the question of whether individual differences in modality-*in*dependent aspects of working memory within the pediatric CI population might have a meaningful causal relation to the level of verbal language skill attained by individual children. The present research begins to address this issue since it provides some of the first data on short-term/working memory in pediatric cochlear implant users involving tasks in which the potential contribution of each available modality was varied.

References

Baddeley, A. D. (1986). *Working Memory*. London: Oxford University Press.

Baddeley, A. D. (1992). Working memory. *Science, 255,* 556-559.

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

Brandimonte, M. A., Hitch, G. J., & Bishop, D. V. (1992). Verbal recoding of visual stimuli impairs mental image transformations. *Memory and Cognition, 20,* 449-455.

Carlson, J. L, Cleary, M., & Pisoni, D. B. (1998). Performance of normal-hearing children on a new working memory span task. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 251-275). Bloomington, IN: Speech Research Laboratory, Indiana University.

Carmichael, L, Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived forms. *Journal of Experimental Psychology, 15,* 73-86.

Cleary, M. (1997). Measures of phonological memory span for sounds differing in discriminability: Some preliminary findings. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 93-138). Bloomington, IN: Speech Research Laboratory, Indiana University.

Cleary, M. Pisoni, D.B., & Kirk, K. I. (1998). Performance of a sample of hearing-impaired children on an auditory-spatial working memory task and its relation to open-set word recognition skills. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 231-250). Bloomington, IN: Speech Research Laboratory, Indiana University.

Conrad, R. (1972). Short-term memory in the deaf: A test for speech coding. *British Journal of Psychology, 63,* 173-180.

Dedina, M. J. (1987). SAP: A speech acquisition program for the SRL-VAX. In *Research on Speech Perception Progress Report No. 13* (pp. 331-337). Bloomington, IN: Speech Research Laboratory, Indiana University.

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

Dowell, R. C., Blamey, P. J., & Clark, G. M. (1995). Potential and limitations of cochlear implants in children. *Annals of Otology, Rhinol Laryngology Suppl. 166,* 324-327.

Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test, Third Edition.* Circle Pines, Minnesota: American Guidance Service.

Fastenau, P. S., Conant, L. L., & Lauer, R. E. (1998). Working memory in young children: Evidence for modality-specificity and implications for cerebral reorganization in early childhood. *Neuropsychologia, 36,* 643-652.

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M., 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.

Furth, H. G. (1966). *Thinking Without Language: Psychological Implications of Deafness*. The Free Press: New York.

Gantz, B. J., Woodworth, G.G., Abbas, P. J., Knutson, J. F., Tyler, R. S., (1993) Multivariate predictors of audiological success with multichannel cochlear implants. *Ann Otol. Rhinol. Laryngol., 102,* 909-916.

Geers, A. (1999). Factors contributing to speech perception in children before age 5. Presented at the *138th Meeting of the Acoustical Society of America.* Columbus, OH, November 1-5, 1999.

Geers, A., Nicholas, J., Tye-Murray, N., Uchanski, R., Brenner, C., Davidson, L, Torretta, G. M. (1998). Effects of communication mode on skills of long-term cochlear implant users. Presented at the *Seventh Symposium on Cochlear Implants in Children*. Iowa City, IA, June 4-7, 1998.

Grossi, D., Orsini, A., Monetti, C., & De Michele, G. (1979). Sex differences in children's spatial and verbal memory span. *Cortex, 16,* 667-670.

Gupta, P. & Dell, G. S. (1999). The emergence of language from serial order and procedural memory. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 447-481). Mahwah, NJ: Lawrence Erlbaum.

Hale, S., Bronik, M.D., Fry, A.F. (1997). Verbal and spatial working memory in school-age children: Developmental differences in susceptibility to interference. *Developmental Psychology. 33,* 364-371.

Hanson, V. L. (1990). Recall of order information by deaf signers: Phonetic coding in temporal order recall. *Memory and Cognition, 18,* 604-610.

Hanson, V. L. & Lichtenstein, E. H. (1990). Short-term memory coding by deaf signers: The primary language coding hypothesis reconsidered. *Cognitive Psychology. 22,* 211-224.

Hernandez, L. (1995). Current computer facilities in the Speech Research Laboratory. In *Research on Spoken Language Processing Progress Report No. 20* (pp.389-393). Bloomington, IN: Speech Research Laboratory, Indiana University.

Kail, R. (1991). Developmental changes in speed of processing during childhood and adolescence. *Psychological Bulletin, 109,* 490-501.

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.

Knutson, J. F., Boyd, R. C., Goldman, M., & Sullivan, P. M. (1997). Psychological characteristics of child cochlear implant candidates and children with hearing impairments. *Ear & Hearing. 18,* 355-63.

Knutson, J. F., Hinrichs, J. V., Tyler, R. S., Gantz, B.J., Shartz, H.A., & Woodworth, G. (1991). Psychological predictors of audiological outcomes of multichannel cochlear implants. *Annals of Otology, Rhinology and Laryngology, 100,* 817-822.

Kramer, J. H., Delis, D. C., Kaplan, E., O'Donnell, L., & Prifitera, A. (1997). Developmental sex differences in verbal learning. *Neuropsychology, 11,* 577-584.

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Lewkowicz, D. J. & Lickliter, R. (1994). *The Development of Intersensory Perception: Comparative Perspectives.* Hillsdale, NJ: Lawrence Erlbaum.

Lezak, M. D. (1995). *Neuropsychological Assessment, Third Edition.* New York: Oxford University Press.

Lyxell, B., Andersson, J., Arlinger, S., Bredberg, G., Harder, H., & Ronnberg, J. (1996). Verbal information-processing capabilities and cochlear implants: Implications for preoperative predictors of speech understanding. *Journal of Deaf Studies and Deaf Education, 1,* 190-203.

Marschark, M. & Mayer, T. S. (1998). Mental representation and memory in deaf adults and children. In *Psychological Perspectives on Deafness: Volume 2.* J. Marschark & M. D. Clark, (Eds.) Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Mayberry, R. I. (1992). The cognitive development of deaf children: recent insights. In *Handbook of Neuropsychology, Vol 7.* S. J. Segalowitz & I. Rapin (Eds.). Elsevier Science Publishers.

Miyamoto, R., Robbins, A. M., & Osberger, M. J. (1993). Cochlear Implants. In Cummings et al. (Eds.), *Otolaryngology, Head and Neck Surgery, 2nd Edition, Vol. 4.* St.Louis, MO: Mosby Publishers.

Miyamoto, R. T., Osberger, J. J., Todd, S. L., Robbins, A. M., Stroer, B. S., Zimmerman-Phillips, S., Carney, A. E. (1994). Variables affecting implant performance in children. *Laryngoscope, 104,* 1120-1124.

Naus, M. J. & Ornstein, P. A. (1983). Development of memory strategies: Analysis, questions, and issues. In M. T. Chi (Ed.), *Trends in memory development research* (pp.1-30). New York: Karger.

Nikolopoulos, T. P., O'Donoghue, G. M., & Archbold, S. (1999). Age at implantation: its importance in pediatric cochlear implantation. *Laryngoscope, 109,* 595-9.

Pisoni, D. B. (1997). Cognitive factors and cochlear implants: An overview of the role of perception, attention, learning, and memory in speech perception. *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory, (pp. 335-347).

Pisoni, D. B. (1999). Some measures of working memory span in deaf children with cochlear implants. Talk presented at Central Institute for the Deaf, Summer 1999.

Pisoni, D. B. & Geers, A. E. (1998). Working memory in deaf children with cochlear implants: Correlations between digit span and measures of spoken language. *Research on Spoken Language Processing Progress Report No. 22*. Bloomington, IN: Speech Research Laboratory, (pp. 336-343).

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*. Bloomington, IN: Speech Research Laboratory, (pp. 51-91).

Quittner, A. L., Smith, L. B., Osberger, M. J., Mitchell, T. V., & Katz, D. N. (1994). The impact of audition on the development of visual attention. *Psychological Science, 5,* 347-353.

Quittner, A. L. & Steck, J. T. (1991). Predictors of cochlear implant use in children. *The American Journal of Otology, 12(Supplement),* 89-94.

Sehgal, S. T., Kirk, K. I., Pisoni, D. B., Miyamoto, R. T. (1997). Effect of residual hearing on children's speech perception abilities with a cochlear implant. Poster presented at *the Vth International Cochlear Implant Conference*, New York City, New York, May 1-3, 1997.

Shand, M. A. (1982). Sign-based short-term coding of American Sign Language signs and printed English words by congenitally deaf signers. *Cognitive Psychology, 14,* 1-12.

Smith, E. E. & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283,* 1657-1661.

Snik, A.F., Vermeulen, A.M., Geelen, C. P., Brokx, J. P., & van den Broek, P. (1997). Speech perception performance of children with a cochlear implant compared to that of children with conventional hearing aids. II. Results of prelingually deaf children. *Acta-Otolaryngol-Stockh., 117,* 755-9.

Stein, B. E. & Meredith, M. A. (1993). *The Merging of the Senses.* Cambridge, MA: MIT Press.

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

Swanson, H. L. (1996). Individual and age-related differences in children's working memory. *Memory and Cognition, 24,* 70-82.

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

Tyler, R. S., Fryauf-Bertschy, H., Kelsay, D. M. R., Gantz, B. J., Woodworth, G. P., & Parkinson, A. (1997). Speech perception by prelingually deaf children using cochlear implants. *Otolaryngology Head and Neck Surgery, 117*, 180-187.

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

Zwolan, T. A., Zimmerman-Phillips, S., Ashbaugh, C. J., Hieber, S. J., Kileny, P. R., & Telian, S. A. (1997). Cochlear implantation of children with minimal open-set speech recognition skills. *Ear & Hearing, 18*, 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Use of Partial Stimulus Information in Spoken Word Recognition Without Auditory Stimulation[1]

Lorin Lachs

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by a grant from the NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington. For extremely valuable discussion and guidance, I thank David Pisoni. Special thanks also go to Luis Hernàndez and Patrick Kelley for their invaluable assistance during the completion of this study.

Use of Partial Stimulus Information in Spoken Word Recognition Without Auditory Stimulation

Abstract. The identification of isolated words in speechreading environments is extremely prone to error. Much of the reason for these errors is due to the extremely impoverished nature of spoken stimuli when the only perceptual information available is visual; some estimates place the number of visually discriminable segments at just over 25% of the number discriminable in auditory-only environments. However, previous research has shown that the confusions made by lipreaders are in fact patterned with respect to the perceptual confusability of phonetic segments in visual-only environments. In addition, it is well known that other sources of information such as phonotactic, lexical, and semantic constraints can play a role in speechreading performance. The current study investigated whether the errors made by speechreaders identifying isolated English words were random or in fact patterned in way that belies the use of partial information and lexical constraints during the process of visual-only spoken word recognition.

Multimodal speech perception has become a topic of considerable interest in the speech community (Massaro, 1998). The pioneering findings of Sumby and Pollack (1954) and the illusory "McGurk" effect (McGurk & MacDonald, 1976) demonstrate the importance of optical information in the process of speech perception. In both of these studies, visual information about speech was shown to have a considerable effect on the identification of spoken utterances. Consequently, some researchers have argued that current theories of speech perception based solely on the auditory properties of the speech signal must necessarily be deficient (Summerfield, 1987). At best, those theories are lacking in critical detail; at worst, they are fundamentally flawed. Determining how and why information from multiple sensory modalities is useful will not only provide explanations of perceptual illusions like the McGurk effect, but will also inform theories of speech perception in general.

In order to answer these questions, some integration and synthesis of what is known about speech perception in unimodal environments is necessary. Surprisingly, the properties of speech in auditory-only ("AO") and visual-only ("VO") environments show an extraordinary complementarity (Massaro, 1998). For example, data on AO perceptual confusions among consonants shows that information for place of articulation is easily lost in noisy environments, while the very same information is highly salient in VO environments **(**Dodd & Burnham, 1988; Miller & Nicely, 1955). In fact, Summerfield (1987) goes so far as to assert that, "to a first approximation.. the visible distinctiveness of consonants ... is inversely related to their auditory distinctiveness" (p. 15).

Unfortunately, much more is known about speech perception in AO environments than in VO environments. The ability to perceive VO speech (often called "lip-reading" or "speechreading") has mainly been explored in the clinical audiology literature (Demorest & Bernstein, 1991; Jeffers & Barley, 1971). Consequently, the application of this knowledge to general theories of speech perception has only recently begun. Several key concepts, however, are central to all speechreading investigations. One of the most important proposals is the concept of the "viseme". A viseme is a class of phonetic segments whose

constituents are treated as equivalent or identical based on their perceptual confusability in VO settings (Fisher, 1968; Jackson, 1988). For example, the phonemes [p], [b], and [m] are virtually indistinguishable without sound; all are articulated bilabially, but are distinguished by the manner of their articulation ([p] is unvoiced, [b] is voiced, and [m] is voiced and nasalized). As such, the segments [p], [b], and [m] frequently make up one viseme class.

It should be emphasized, however, that the constituency of a viseme is not necessarily determined by the linguistically motivated featural descriptions implied in the preceding example. On the contrary, the precise structure of a viseme depends on the perceptual confusability of the phonetic segments themselves and the *ad hoc* criterion for confusability used to cluster them (Auer & Bernstein, 1997). For example, using a strict criterion for confusability could yield 28 distinct viseme groups, each with a constituency of up to 3 phonemes. On the other hand, an extremely lax criterion might yield two viseme groups: one for consonants, and one for vowels. One cutoff was proposed by Walden, Prosek, Montgomery, Scherr and Jones (1977) who used a criterion at which at least 75% of confusions for a given segment were segments in the same viseme.

Further complicating the issue of defining the viseme is the fact that the confusability of the individual segments is affected considerably by environmental and instance-specific factors. For example, lighting, angle of viewing, and distance from a talker obviously affect the quality of the optical information provided in a VO environment and the ability of the speechreader to use that information (Auer & Bernstein, 1997; Jackson, 1988). Similarly, the idiosyncratic speechreading abilities of the observer receiving the message play a role in visual-only intelligibility. Walden, et al. (1977) showed that training techniques significantly affected the number of confusions made by speechreaders in segmental identification. After only a few hours of training, the number of visemes (as defined above) went up and phonetic segments were more likely to be confused with the segments from within their own viseme class, rather than segments from outside their viseme class. This pattern suggests that the ability to distinguish between speech segments in VO conditions can be improved through learning.

Finally, the distinctive and idiosyncratic way in which a particular talker articulates speech also plays a large role in affecting his/her intelligibility when observed in VO environments (Lesner, 1988; Lesner & Kricos, 1981). In their classic investigations, Kricos and Lesner (1982, 1985) reported talker-specific differences in visual intelligibility of consonants in nonsense syllables. Both the number and nature of the visemes produced showed marked variation across talkers. The visible characteristics of vowel production, too, point to variation across talkers (Montgomery & Jackson, 1983), as do patterns of vowel confusions across talkers (Montgomery, Walden, & Prosek, 1987). In addition, Jackson (1988) reviewed several investigations of VO speech perception and concluded that the *number* of visemes revealed by the various experiments was different. The specific speech segments that made up the constituencies of the visemes in each study were also different. Jackson points out, however, that the variation was not completely random; the /p, b, m/, /f, v/, and /θ, ð/ visemes are consistently confused with each other, regardless of the talker. Similarly, the /ʃ, ʒ, t͡ʃ, d͡ʒ/ and /w,r/ visemes are commonly grouped together. However, Jackson (1988) claims that "there is no one viseme system that accurately describes the visual characteristics of all phonemes for all talkers" (p. 103-104).

Another interesting aspect of talker variation in VO intelligibility is the finding that such variability is not simply derivative of a particular talker's AO intelligibility (Gagné, Masterson, Munhall, Bilida, & Querengesser, 1994), although for some talkers such a relationship can be shown. Gagné, et al. (1994) showed that the ranking of the intelligibility of the various talkers in their study was not consistent across

VO and AO conditions. As they point out, the patterns of articulatory movement that benefit speech intelligibility in one sensory modality may not be beneficial to the other.

All of these factors play an important role in the underlying confusability of phonetic segments in the VO environment. Auer and Bernstein (1997) suggest that generating several sets of visemes using different criterion levels can adequately model the effects of these factors. Most researchers, however, explicitly test the segmental confusability of a particular talker and analyze the resultant identification matrix using hierarchical clustering (e.g., Walden, et al., 1977) or multidimensional scaling techniques (e.g., Bernstein, Demorest, & Eberhardt, 1994). As of yet, there is no precise model for ***predicting*** the effects of the environmental, talker-specific, and observer-related factors that affect intelligibility, independently or in conjunction.

In addition to these instance-specific attributes of the stimulus, other effects on speechreading performance arise due to the interaction of different levels of processing. Boothroyd (1988) lists topical, semantic, syntactic, lexical and phonological constraints as potential sources of information in the disambiguation of a speechread stimulus. As with auditory speech perception (Lively, Pisoni, & Goldinger, 1994), the information provided by each of these "levels" facilitates performance, acting in combination for maximum benefit (Boothroyd, 1988). For example, Gagné, Tugby and Michaud (1991) found that embedding test items in semantically related sentences increased speechreading accuracy relative to test items embedded in unrelated sentences. Similarly, Lansing and Helgeson (1995) showed that priming a target word with a semantically associated prime word facilitated VO identification of the target. The utility of the various higher order sources of information is conditional on the availability and utility of the other higher order sources of information available (Lyxell & Rönnberg, 1989).

Several recent investigations by Bernstein and her colleagues have revealed a complex and important role for the mental lexicon in VO spoken word recognition. Auer and Bernstein (1997) performed computational analyses on several "visually-transcribed" lexicons. For a given set of visemes, the "visual-transcription" transformation of the mental lexicon effectively collapses across phonemic distinctions that lie within a viseme class. For example, given a viseme set that contains a viseme whose constituents are [b],[p], and [m], the words "bob", "mom" and "pop" would all be collapsed into the same word-level equivalence class. Much like earlier investigations into the properties of the lexicon given "coarsely-coded" or broad phonetic categories (Huttenlocher & Zue, 1984), Auer and Bernstein demonstrated that the loss of perceptual uniqueness across transformations of the lexicon using increasingly strict viseme sets is in fact patterned, and can potentially be useful during VO identification. For example, a frequency-weighted proportion of the number of unique words revealed that over half of the words in the lexicon remain distinct after transcription with a set of 12 visemes, even though this broadly-defined set of visemes reduced the number of segments by roughly 75%. In addition, for those words that do not remain unique at these lax criteria, the number of words with which they are equivalent ("expected class size") remains relatively low, with the average being around 5.1 for the 12 viseme set. Of course, these numbers are contingent on the total number of words presumed to be in the mental lexicon, but Auer and Bernstein's data show that these patterns remain qualitatively the same given different lexicon sizes. One consequence of the properties revealed by their analyses is that high frequency words tend to remain distinct from other high frequency words across these transformations. Thus, a speechreader could theoretically maximize his/her performance by simply choosing the most frequent word implicated by the visemes available in the stimulus display.

Given these findings and the importance of the mental lexicon in AO spoken word recognition (Lively, et al., 1994; Luce & Pisoni, 1998), it is likely that continued investigations of spoken word recognition under VO conditions will provide useful insights into speechreading ability and the nature of the neural representation of spoken words in long term lexical memory.

The present investigation examined the responses of a large number of participants who were asked to speechread isolated word tokens from the Hoosier Audiovisual Multimodal Database (Lachs & Hernàndez, 1998; Sheffert, Lachs & Hernàndez, 1997). The Hoosier Audiovisual Multimodal Database ("HAVMD") is a 3000 token collection of digitized, audiovisual movie clips of 10 talkers uttering 300 isolated words. These 300 words consist of 150 lexically "easy" words, and 150 lexically "hard" words (Luce & Pisoni, 1998). Easy words are defined as high frequency words that reside in sparse similarity neighborhoods whose average frequency is low; hard words are defined as low frequency words that reside in dense similarity neighborhoods whose average frequency is high (Goh & Pisoni, 1998). A word is a neighbor of a given target word if it differs from the target word by one phoneme, either substituted, inserted, or deleted.

Sheffert et al. (1997) reported that the overall audiovisual speech intelligibility of the tokens was 98.57%, with a range of 74.09% to 100%. This score was based on correct identification of the stimulus word. In addition, a per-item analysis showed that most of the tokens were identified near ceiling. 99.97% of the tokens in the 3000 token database were identified across talkers with above 90% accuracy. Finally, Sheffert et al. (1997) report a main effect of lexical category. Lexically "easy" words were identified more often than lexically "hard" words. No intelligibility differences were found across talkers. AO and VO identification of the same tokens yielded similar effects of lexical status. Inter-talker differences were also observed (Lachs & Hernàndez, 1998). However, VO intelligibility was extremely low, with an overall mean intelligibility of 14.13% correct. 90% of the tokens were identified with less than 40% accuracy.

The present analysis examined the responses to VO identification of the HAVMD tokens in greater detail, in the hope that a more in-depth analysis would provide new insights into the nature of lexical and talker effects on VO speech recognition. As shown by the studies mentioned above, requiring correct *phonemic* identification for accuracy measurements does not accurately reflect the performance of speechreaders. Although some phonemic contrasts remain distinctive in VO environments, many of them do not. Scoring responses in accordance with visemic criteria for perceptual equivalence, then, may reveal the availability of partial information about word identity in VO tasks and show how this information is used by speechreaders. It was expected that the errors made by speechreaders in the present identification experiment were not haphazard, but structured on the perceptual confusability of individual segments and the relationships between spoken words in the mental lexicon.

Method

Subjects

Two hundred Indiana University undergraduates participated as observers either in exchange for course credit in an introductory psychology course or for payment five dollars. All subjects had normal hearing, normal or corrected-to-normal vision, were native English speakers, and reported no history of speech or hearing disorders at the time of testing. The subjects were all drawn from the same population of college students in Bloomington, IN.

Materials

Two Apple Macintosh Power PCs and three Macintosh clone (PowerComputing 604|150) computers, each equipped with a 17" Sony Trinitron Monitor (0.26 dot pitch) and its own video processing board were used to present the stimuli to subjects. The video processing boards were each capable of handling clips digitized at 30 frames per second (fps) with a size of 640 x 480 pixels and 24-bit resolution. The stimuli consisted of all the movies from all the talkers in the HAVMD database.

Procedure

A computer program was written to control stimulus presentation and collect responses. The custom-designed software for presentation of the digitized movies was altered so that the audio track for each movie was not presented during presentation.

Each participant was presented with a randomly ordered list of movies for identification; each list of movies consisted of all the tokens spoken by one of the 10 talkers. Stimuli were presented over the Sony Trinitron 17" monitors.

Before the presentation of the first stimulus, participants in both conditions were handed a set of typed instructions that explained the task and procedures. Listeners were informed that they would see a series of video clips in which a person would be saying an isolated, single English word. Participants were informed that each stimulus would be presented only once. After each stimulus, participants were required to identify the word by typing their response on the keyboard. They were instructed that the next stimulus would not be presented until they pressed the RETURN key. Participants were also reminded to take time to make sure that the response they typed was the response that they intended to make before entering it. Each response was then recorded and collected in a text file that contained the name of the movie, its order in the presentation, and the subject's response.

Data Analysis

All responses to all stimuli were compiled in a large textfile for further analysis. In all, there were twenty responses to each of the three thousand stimulus tokens. The responses were initially checked by hand for any obvious typos. An obvious typo was defined as a word containing one letter in a string that deviated from the target letter at that position, and whose key on a standard keyboard is adjacent to the target letter's key. Alternatively, a typo was also defined by the insertion of a letter in a response string which was within one key of any of the surrounding letters in the response string. The textfile containing the target words and the responses to those words was fed into a DECtalk DTC03 Text-to-Speech System, configured such that it could output an ASCII-based phonemic transcription of each target and response (Bernstein, Demorest, & Eberhardt, 1994). This transcription was performed so that there was an algorithmic process for dealing with the phonemic transcription of (a) nonwords and (b) words subject to inter-talker pronunciation differences (e.g., "been", "pen", etc.). In addition, the automated procedures for the transcription process made the task of transcribing sixty thousand individual responses more tractable than if carried out by hand.

Nonwords were handled in two different ways by DECtalk. Some nonwords were evidently "pronounceable" (e.g., "gak", "sokat"), and for these nonwords, DECtalk output its best guess as to the phonemic transcription of the nonword. Other nonwords were "unpronounceable" (e.g., "sdjkb", "tnnnfhg"), and DECtalk simply output a phonemic transcription as though it were reading the individual segments in the order they occurred. The latter type of nonword responses was eliminated from any further analysis.

The output of the DECtalk transcription process was then submitted to a custom-designed scoring program that computed several measures for each target-response pair. All measures were computed over several sets of pre-defined viseme groups.

Confusability Continuum Heuristic. Because no data on the perceptual confusability of individual segments for the talkers in the HAVMD was available, these viseme groups were taken from the

sets used by Auer and Bernstein (1995). One can assume, however, that for all talkers, segments like [f] and [v] are perceptually 'closer' to one another than [f] and [k]. Likewise, it can be assumed that there are common underlying relational patterns in the perceptual 'distances' between all speech segments. These patterns specify a perceptual confusion space in which distance is inversely related to similarity (Nosofsky, 1986). It is assumed that some dimensions of the confusion space will necessarily collapse before others do in all VO environments. In our example, segments [f] and [v] would collapse earlier than [f] and [k] for all talkers. Individual variation, using this conception, will occur because for some talkers [f] and [v] are perceptually indistinguishable, while for others they are not. The pattern of perceptual confusability for a particular talker, then, could be described as some function of the criterion for perceptual equivalence in the underlyingly common confusion space.

We can represent this criterion as a number on some continuum. High values of the criterion represent subspaces where segments are highly distinct from one another (i.e., not many dimensions are collapsed). Low values of the criterion represent spaces where segments collapse onto one another in perceptual equivalence. Thus, scoring at the various viseme levels could represent sampling the continuum of confusability at various intervals. I will refer to this representation as the Confusability Continuum Heuristic (CCH).

CCH Sampling Points. Because the *actual* phonemic confusion spaces associated with each talker in the HAVMD are not known, six viseme sets were used in scoring the target-response pairs. These viseme sets tested a range of criteria for perceptual confusability and represented different points along the Confusability Continuum. The most stringent viseme set, of course, was the full set of phonemes, where no segment was considered indistinguishable from any other. There were 46 phonemes in all. In order from strict to lax, the next viseme sets contained 28, 19, 12, and 10 visemes each. The most lax viseme set contained only two visemes: one for consonants and one for vowels. These viseme sets were taken from Auer and Bernstein (1995) and were generated using confusion data based on a different set of audiovisual materials.

For each “real” viseme set, another “random” set was created using the same number of visemes. For these “random” visemes, phonemes were assigned randomly to each viseme. These random sets were created in order to control for improvements in speechreading accuracy due simply to the relaxation of the perceptual distinctiveness criterion. Appendix A contains the detailed structure of both the real and random viseme sets used during scoring. Notice that the randomly generated “28 viseme” set lists only 20 sets. This is because the random assignment procedure generated 8 sets with no constituents. In total, 12 viseme sets (6 “real” and 6 “random”) were used for all subsequent scoring procedures. Viseme translation was accomplished via a set of transcription rules similar to those used in Auer and Bernstein (1995).

	Dependent measure class		
Responses to the target: tæb	FL score	ORD score	SGM score
tæb	1.0	1.0	1.0
tæbi	0	0.75	0.75
tæn	0	0.66	0.66
bæt	0	0.33	1.0

Table 1. Scores for hypothetical responses to the target word "tab" under the various measures. The target and responses are shown in their IPA transcriptions.

Measures of Accuracy

First-order Measures. Three broad classes of accuracy measures were computed using each of the viseme sets. These measures were chosen because they represent several important ways of assessing accuracy and the use of partial stimulus information in spoken word recognition. These classes are not independent, as discussed below, but they do measure different aspects of speechreading accuracy. Twelve viseme-translated versions of both the target and the response were measured using each of the measures below. Thus, for each target-response pair, thirty-six measures of accuracy were obtained, each measuring a different aspect of performance at the various criteria for perceptual confusability.

Table 1 illustrates the function of each of the dependent measure classes given four responses to the hypothetical target word "tab". The target and responses are shown using their IPA transcriptions, in order to avoid confusion. The scores shown represent scoring using the strictest, phonemic criterion. The "full" ("FL") class of measures was binary and determined whether, given a specific criterion for confusability, a response was *exactly* the same as the target or not. Thus, the column marked "FL score" in Table 1 shows that all responses to the visually presented target word "tab" were incorrect, unless the response itself was also the word "tab".

Of course, the criterion for an exact match changed somewhat depending on which viseme criterion was being used. So, for example, the response "bat" given for the target "mat" would receive a score of 0 using the strictest criterion, but receive a score of 1 at all those criterion levels where [b] and [m] were constituents of the same viseme. This is true for all the other classes of measures as well.

The "ordinal" ("ORD") class measured the proportion of segments in the target word that were contained in the response *in the order that they occurred in the target*. In order to control for responses of different lengths, all ORD scores were normalized by the length, in segments, of the response. In the column marked "ORD score" in Table 1, the function of the ORD score is illustrated. The response "tab" receives a 1.0, since it contains all the segments found in the target, in the correct order. The response "tabby" meets these requirements as well; however, because ORD scores are divided by the length of the response, the response "tabby" receives an ORD score of 0.75. A response of "tan" receives an ORD score of 0.66, since that is the proportion of segments in the response that are contained in the target in the correct order. In contrast, the response "bat" contains all the segments found in the target; however, because they are in the wrong order, this response receives an ORD score of 0.33. Of course, responses with no segments in common with the target receive ORD scores of 0. From this example, it can easily be seen that the ORD class is a more fine-grained submeasure of the FL class.

Target-response pair	ORD	Ov28	oP28
bæt - mæt	0.66	1.0	1.0

bæt - mæʃ	0.33	0.66	0.5

Table 2. Calculation of the transitional ORD score for the transition between phonemes and the 28 viseme level. The column marked ORD represents the ORD score calculated using the phoneme criterion. The column marked Ov28 represents the ORD score calculated using the 28 viseme criterion. The column marked Op28 represents the transitional ORD score for the phoneme to 28 viseme transition.

In an analogous way, the "segmental" (SGM) class of measures was a submeasure of the ORD class. SGM scores measured the proportion of segments in the target that were contained in the response, irrespective of the order in which they occurred. SGM scores, like ORD scores, were normalized for the length of the response. The "SGM" column of Table 1 illustrates the function of the SGM score using the same target-response pairs as above. The responses "tab", "tabby", and "tan" all receive SGM scores equivalent to their ORD scores, for the same reasons described above. The response "bat", however, is treated differently by the SGM scoring procedure and receives a score of 1.0, since it contains all, and only, the segments in the target. Unrelated targets received a SGM score of 0.

It can be seen from the relationships among the various measures that they represent another kind of criterion-relaxation along a different accuracy dimension. The FL class represents the strictest criterion, the ORD represents a somewhat more relaxed criterion, and SGM scores represents the most lax criterion. It was hoped that computing accuracy at these different levels would reveal the fine-grained structure of response patterning.

Transitional or Second-order Measures. In addition to the first-order measures of accuracy, transitional measures were also computed to detect inter-talker differences in the degree to which specific visemic criterion relaxations helped or hindered scores. The transitional measures reflected the gain in accuracy due to the relaxation of perceptual confusability criteria. For example, the transitional measure "oP28" measured the gain in ORD accuracy due to scoring with the strictest viseme set, phonemes, to the next viseme set, with 28 visemes. These gains in accuracy were normalized relative to the possible gain in accuracy across the transition in question. In general, the formula for computing the transitional measures was as follows:

$$X_i - X_{i+1} \Big/ 1.0 - X_i$$

where X is a particular class of first-order measure like ORD or SGM. The index "i" represents the ordinal index of the stricter viseme set involved in the transition, and "i+1" represents the next most lax viseme set.

Viseme level	F(9, 60143)	η^2
Phonemes	51.306*	0.008
28 visemes	73.288*	0.011
19 visemes	91.780*	0.014
12 visemes	106.325*	0.016
10 visemes	120.523*	0.018
2 visemes	66.832*	0.010

Table 3. F statistics and associated η^2s for the main effect of the talker variable for the FL class at the various viseme levels. *All statistics are significant with p <=0.0009.

Table 2 illustrates the function of the transitional scores using the ORD scores for several hypothetical target-response pairs. Consider the pair "bat - mat" first. At the phoneme level, the ORD score for this pair is 0.66; at the next most lax viseme set, the score becomes 1.0, since [b] and [m] are considered equivalent at this level. Thus, the actual improvement across this transition was 0.33. The possible improvement across this transition is also 0.33. So, the transitionary ORD score "oP28" for this target-response pair would be 1.0.

Now, consider the target-response pair "bat - mash". At the phoneme level, the ORD score for this pair is 0.33. At the 28 viseme level, the ORD score is 0.66, since [ʃ] and [t] remain distinct using this criterion. So, the actual improvement across this transition was 0.33; however, the possible improvement was 0.66. Thus, the transitionary ORD score "oP28" for this target-response pair would be 0.5.

Results and Discussion

First-order Accuracy Measures

Full ("FL") Class. Table B.1 in Appendix B shows the average FULL accuracy associated for Easy and Hard words for each of the talkers at each of the viseme criterion levels. Inspection of the table reveals that, for all talkers, scores increase as the visemic criterion is relaxed. However, there are differences in the absolute accuracy for each talker, as well as overall differences in accuracy for Easy and Hard words. The data scored using each of the visemic criterion levels were treated as separate dependent measures in a 10 (Talker) by 2 (Easy/Hard) MANOVA. η^2 values were computed for each of the main effects and interactions in the MANOVA, to determine the relative sizes of any significant effects. An effect was considered significant if p <= 0.05.

Viseme level	F(1, 60143)	η^2
Phonemes	1885.853*	0.030
28 visemes	1279.356*	0.021
19 visemes	1095.707*	0.018
12 visemes	361.425*	0.006
10 visemes	85.907*	0.001
2 visemes	25.375*	<= 0.0009

Table 4. F statistics and associated η^2s for the main effect of the lexical status (Easy/Hard) variable for the FL class at the various viseme levels. *All statistics are significant with p <=0.0009.

For all the measures performed using the "real" viseme sets, the main effect of talker was significant. Table 3 shows the F values and associated η^2s for the main effect of talker at each of the viseme levels. The η^2s increase with increased relaxation of the viseme criterion, except for the 2 viseme level, indicating that differences in the talker became more important with this relaxation. Post-hoc Tukey's HSD tests were conducted to assess these main effects. Of all the talkers at the phonemic level, M1 is the least visually intelligible, while M3 is the most visually intelligible. For all the visemic criterion levels, M1 remains the least intelligible of all the talkers, but the rankings do fluctuate according to the particular viseme level being measured. This may be due to the fact that certain viseme criterion levels more accurately model the underlying perceptual confusability of each talker's utterances. This possibility is investigated more fully below in the discussion of the second-order dependent measures.

The η^2 values for each of these F statistics, however, are extremely low, varying between less than 0.008 (phonemes) to 0.018 (10 visemes). This means that while the differences between scores across talkers are detectable, they are extremely small and do not explain a large portion of the variation in these measures. Given our discussion above concerning the plethora of factors affecting speechreading intelligibility, however, this fact is not surprising.

Table 4 shows the F statistics and associated η^2s for the main effect of Lexical Status. It is clear from the F statistics shown that there is a marked difference in the lipreading accuracy of easy vs. hard words. However, as the visemic criterion is relaxed, this difference becomes less robust, as indicated by the η^2s. In contrast to the trend exhibited for the main effect of talker, these values start out explaining much more of the variance, but take a drastic drop between the scores for 19 and 12 visemes. Figure 1 shows the difference in accuracy between easy and hard words for each of the ten talkers at each visemic criterion level. For all talkers, the difference in accuracy between Easy and Hard words decreases, with the biggest drop occurring between 19 and 12 visemes. Strangely, the difference between Easy and Hard words reversed for M1 at the most lax criterion: Hard words identified with greater accuracy than Easy words. Aside from this unexpected finding, the data conforms to what is already known about spoken word recognition and the effects of neighborhood competition. Participants were more accurate in identifying

easy words than hard words. The reason for the reduction in the difference between performance on these two kinds of words is explained more fully below in the section on lexical factors.

Not surprisingly, the lip-read intelligibility of each of the talkers increases as the visemic criterion is relaxed. The difference between the real and random FL scores of the Easy words for each talker at each of the viseme levels is shown in the top panel of Figure 2. The same data for Hard words is shown in the bottom panel. As the visemic criterion is relaxed, the difference in "real" and "random" scores becomes bigger. This indicates that increases in accuracy across the various viseme groupings were not due to an uninteresting general relaxation of the criterion for accuracy. Instead, scores get better because they were measured using an increasingly accurate model of the underlying perceptual confusability of the segments uttered by these talkers.

Summary of First Order FL Measures. The data reviewed above demonstrate that the errors made in speechreading identification are patterned with respect to the confusability of segments in VO environments. Relaxing the criterion for correct identification by reducing the number of visemes resulted in higher scores. Random relaxation of the criterion through the use of randomly generated viseme sets, however, did not improve scores as much. In fact, the difference between accuracy calculated using real versus random visemes increased with increased relaxation of the criterion, indicating that improvements were due to the use of a more accurate model of the perceptual confusability of the talker, and not merely due to an improvement of the odds of picking a segment that would be scored correct. Furthermore, differences among talkers became more exaggerated as the

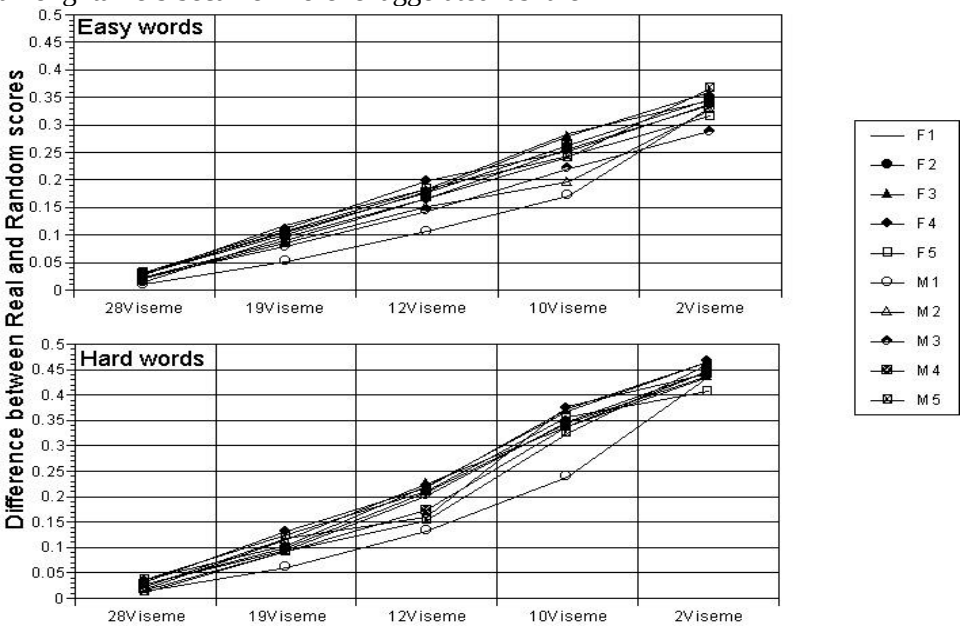


Figure 1. The average difference between FL scores for Easy and Hard words at the various viseme criterion levels. Each line represents the data from a different talker.

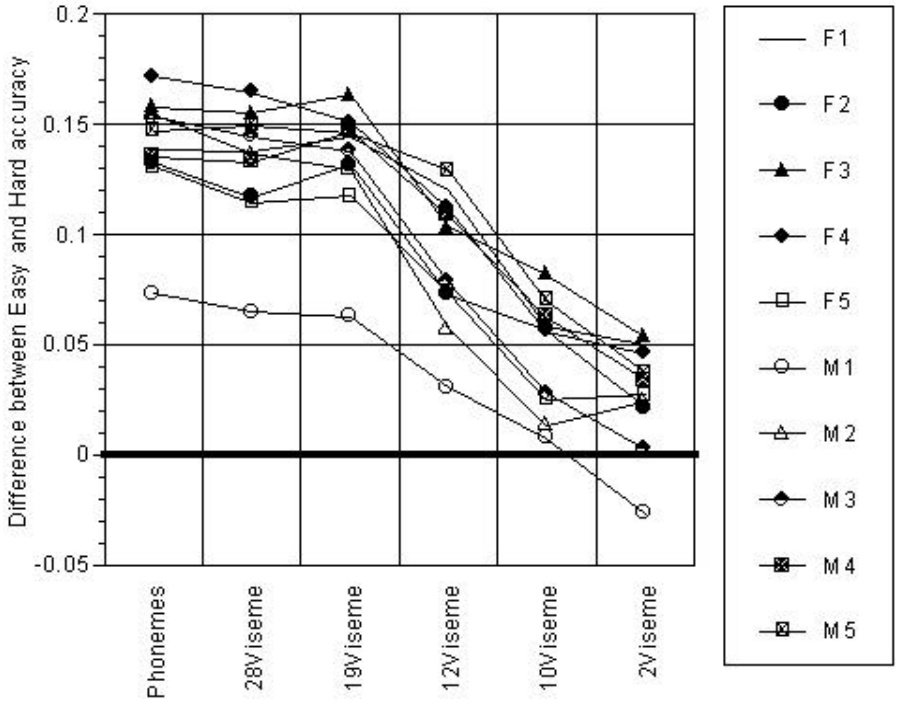


Figure 2. The average difference between FL scores computed using real and random viseme sets at the various viseme criteria. The top panel shows the average differences for Easy words; the bottom panel shows the average differences for Hard words. Each line represents the data from a different talker.

criterion was relaxed, indicating that the availability of partial information for use by speechreaders was dependent on the talker being viewed. Finally, differences in the accuracy of Easy and Hard words decreased as the visemic criterion was relaxed, indicating that lexical factors may become less important as the requirement to make fine-grained discriminations is relaxed (for more discussion on lexical factors, see below).

Ordinal ("ORD") and Segmental ("SGM") Classes. Another requirement for accuracy that can be relaxed is the proportion of the word recognized accurately. The FL class of measures only scores a response as correct when the entire word matches the entire target. However, it would be interesting to know whether VO words are speechread in an all-or-none fashion, or whether parts of the word are more clear than others. This question can be addressed by scoring responses segmentally with the ORD and SGM classes of measures.

Because the ORD and SGM scores are fairly similar in that they both measure a kind of segmental accuracy, the results using these measures are discussed together. However, it should first be pointed out that the mean difference between the ORD score and SGM score for a target-response pair was only 0.08. This difference was consistent across the various viseme criterion levels, confirming the earlier hypothesis that the SGM class represents a kind of submeasure of the ORD class.

Viseme level	ORD F (9, 60143)	ORD η^2	SGM F (9, 60143)	SGM η^2

Phonemes	93.958*	0.014	116.412*	0.017
28 visemes	107.243*	0.016	140.239*	0.021
19 visemes	124.258*	0.018	165.701*	0.024
12 visemes	143.860*	0.021	170.666*	0.025
10 visemes	130.012*	0.019	157.880*	0.023
2 visemes	94.661*	0.014	93.552*	0.014

Table 5. F statistics and associated η^2s for the main effect of the talker variable at the various viseme levels. Results from the MANOVAs performed on ORD (ordinal) and SGM (segmental) scores are shown. Asterisks denote statistics significant with $p <= 0.0009$.

Table B.2 in Appendix B shows the average ORD accuracy for Easy and Hard words for each of the talkers at each of the viseme criterion levels. Table B.3 shows the same data using SGM scores. As with the FL scores, scores increased as the visemic criterion was relaxed for all talkers. Again, there are differences in accuracy based on the Easy/Hard variable as well. One interesting result was that, even for the least intelligible talker (M1) at the phoneme level, *on average* roughly one segment in each target word (ORD = 0.333) was correctly perceived. When the requirement for a response segment to be in the correct order (by switching to SGM scores) is removed, the average accuracy goes up. This indicates that for many words, more than one segment was correctly perceived, but those segments were separated by erroneous response segments, or even transposed in the response.

The data from the ORD and SGM classes were submitted to separate 10 (Talker) by 2(Easy/Hard) MANOVAs, with the measures at each viseme criterion level being submitted as separate dependent variables. η^2 values were also computed using this data.

Again, inter-talker differences were evident in this analysis as well. Table 5 shows the F statistics and associated η^2s for the main effect of talker in the ORD and SGM analyses at the various viseme levels. For ORD scores calculated using the “real” viseme sets, the main effect of talker was significant. The same is true for the SGM scores. Post-hoc Tukey's HSD tests indicated that, at the phoneme level for both ORD and SGM scores, M1 is the least visually intelligible of all the talkers. F5 is the most intelligible talker using ORD scores at the phoneme level, and M2 is the most intelligible talker using SGM scores at the phoneme level. As with the FL scores, the talkers' ranks vary across the different viseme criterion levels; however, M1 remains the least intelligible at all viseme levels under both measures. The effect size of the talker variable (as indicated by the η^2 values) increases for both ORD and SGM scores, but drops off after the 19 viseme point. In contrast to the pattern exhibited by the FL scores, this shows that, when scoring segmentally, there is a point on the viseme criterion continuum at which differences in the talker become less relevant. Again, there seems to be something unusual about the transition between scoring with 19 visemes and scoring with 12 visemes. Further discussion of this will be included below.

Viseme set	ORD F (1, 60143)	ORD η^2	SGM F (1, 60143)	SGM η^2
Phonemes	1012.403*	0.017	758.833*	0.012
28 visemes	494.353*	0.008	495.952*	0.008
19 visemes	474.903*	0.008	443.339*	0.007
12 visemes	63.818*	0.001	63.426*	0.001
10 visemes	5.971**	< = 0.0009	0.014	0.000
2 visemes	0.000	0.000	54.101*	0.001

Table 6. F statistics and associated η^2 values for the main effect of Lexical Status (Easy/Hard) in the MANOVAs conducted on ORD and SGM class scores at the various viseme levels. Single asterisks denote Fs that are significant with p <= 0.0009; double asterisks denote Fs that are significant with p = 0.015.

Differences in intelligibility due to lexical factors were also revealed in the segment-based measures. Table 6 shows the F statistics and associated η2 values for the main effect of Lexical Status (Easy/Hard) in the ORD and SGM MANOVAs at the various viseme levels. As with the FL scores, there was a decrease in the effect size with the increased relaxation of the viseme criterion. Also, the biggest drop occurs, for both classes, between the 19 and 12 viseme criterion points. In contrast to the FL class, however, there is no difference between Easy and Hard word ORD scores at the 2 viseme level. In addition, there is no difference at the 10 viseme level for the SGM scores.

Clearly, the pattern of these differences was not the same as that observed for the FL measures. Figures 3 and 4 show the difference in the average easy and hard scores for each of the talkers under each viseme set using the ORD and SGM measures of accuracy. Using the more strict viseme sets (i.e., those with more visemes), a greater proportion, on average, of each Easy word was identified correctly when compared with the average proportion of each Hard word identified correctly. The opposite is true for some talkers at the more lax viseme sets (specifically, the 10 and 2 viseme sets). That is, more of each Hard word was identified correctly than each Easy word. More on this seemingly paradoxical result will be discussed below in the section on lexical factors.

However, when averaged across talkers, no difference was observed between ORD scores for Easy and Hard words, as denoted by 95% confidence intervals around each mean. In fact, no difference was observed between Easy and Hard words (according to the confidence interval) for F1, F5, M2, and M5. For F2, F3, F4 and M4, Easy scores are higher than Hard scores. For M1 and M3, the opposite is true, with Hard scores higher than Easy scores. Thus, for only two talkers is the anomalous pattern observed.

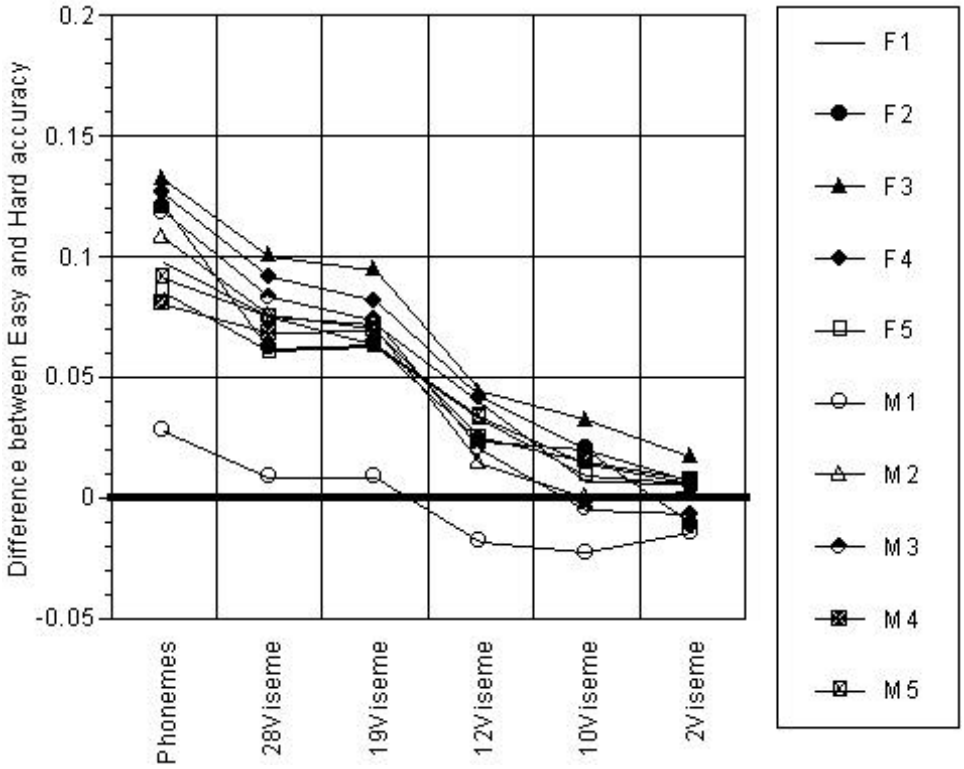


Figure 3. The average difference between ORD scores for Easy and Hard words at the various viseme criterion levels. Each line represents the data from a different talker.

The picture is not as clear for the SGM (segmental) scores. Using the 95% confidence interval, a significant difference in accuracy was observed between Easy and Hard words, when averaged across talkers. In this case, scores for Hard words were higher than those for Easy words. In fact, the Hard words for F1, F2, M1, M3, and M5 are all identified with better SGM accuracy than Easy words. All the other talkers show no difference in accuracy between the two lexical classes. These results seem problematic because they contradict virtually all known findings about the word frequency effect. Hard words are by definition lower frequency than Easy words and should be identified with lower accuracy, regardless of the viseme set being used to evaluate performance. A fuller discussion of this phenomenon is included below in the lexical factors section.

Scores computed using the "real" viseme sets tended again to be consistently higher than their "random" equivalents. Figures 5 and 6 illustrate this point. In both figures, the difference between "real" and "random" scores is shown as a function of the viseme set size, with the data for Easy words in the top panel and the data for Hard words in the bottom panel. Figure 5 shows the ORD data, while Figure 6 shows the SGM data. The pattern is not as straightforward as it was for the FL scores, however. Although the ORD difference in accuracy between real and random tends to increase across the viseme sets, it tapers off at the 2 viseme set. For the SGM differences, there is a drop in the difference between real and random at the 2 viseme set, too.

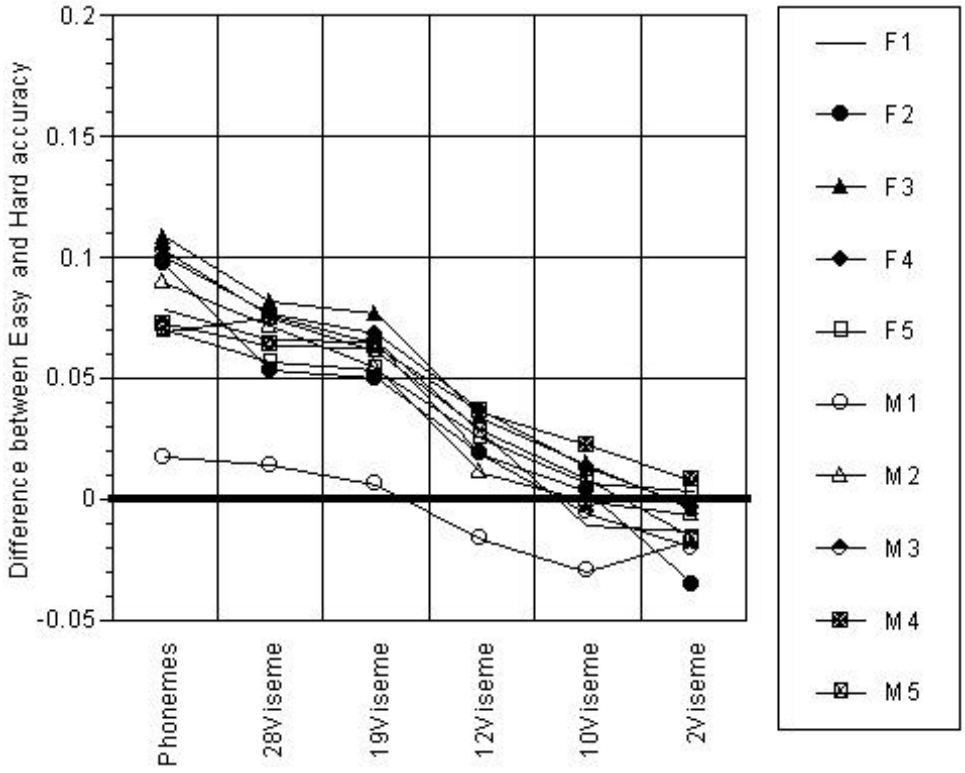


Figure 4. The average difference between SGM scores for Easy and Hard words at the various viseme criterion levels. Each line represents the data from a different talker.

The deviation of these difference scores from the trend observed using the FL class is troublesome. One likely explanation, however, can be found in the distribution of phonemes across the two-viseme sets. For the "real" sets, phonemes are distributed unequally across the two visemes. The "consonant" viseme had 28 constituents and the "vowel" viseme had 18 constituents. However, a set of randomly distributed visemes tends to have equally distributed phonemes. Each viseme within a set tends to have the same number of constituents. As the number of visemes decreases, the tendency for each viseme to contain an equal number of constituents increases. Thus, for the 2 viseme set, this tendency is greatest. Indeed, in the random set used here, there was a more evenly distributed viseme membership than for the real 2 viseme set. The consequence of a more even distribution is that every segment in a response has close to a 50% chance of being correct. Because of this, scores must necessarily be high for these random sets. Thus, there is a kind of ceiling effect shown by the difference scores using the 2 viseme set. It is remarkable that the scores using the "real" 2 viseme set remain more accurate than the "random" one at all. This fact confirms that the improvements in accuracy due to relaxation of the visemic criterion are not due simply to chance. Instead, improvements are due to the fact that the viseme set specified by the relaxed criterion provide a more accurate model of the actual perceptual confusability of the talker.

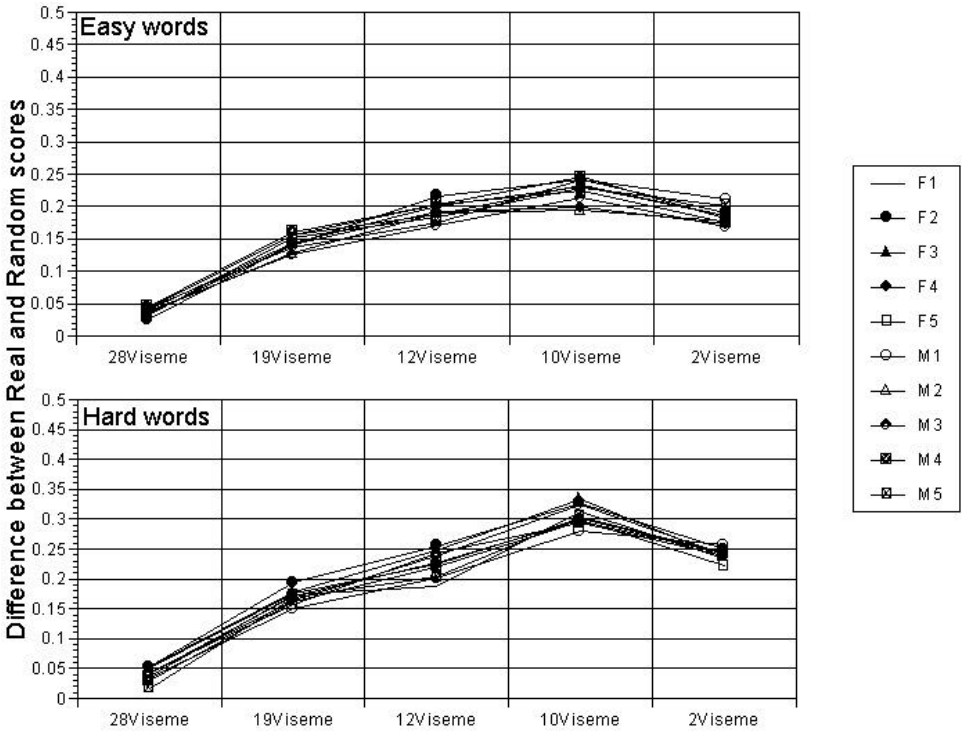


Figure 5. The average difference between ORD scores computed using real and random viseme sets at the various viseme criteria. The top panel shows the average differences for Easy words; the bottom panel shows the average differences for Hard words. Each line represents the data from a different talker.

In addition, it is worthwhile to note that the differences between real and random accuracy at the different viseme levels are different for Easy and Hard words, for both the ORD and SGM classes of scores. Furthermore, the difference between real and random accuracy is increasingly larger in Hard words than it is in Easy words. This provides evidence that the decreasing difference between Easy and Hard word accuracy mentioned above is not simply due to chance. If it were, then the difference between real and random accuracy would increase at the same rate for both Easy and Hard words. A fuller discussion of the role of lexical factors in speechreading performance is included below.

Summary of First-order Measures. Lexical factors were shown to play a role in speechreading accuracy. Lexically easy words were identified more accurately than lexically hard words. However, the difference in performance between these two classes of words decreased with increased relaxation of the visemic criterion for all talkers. This pattern is not surprising because transformations of the lexical similarity space into visemes necessarily collapse across perceptual dimensions presumed to be indistinguishable from each other when viewed in VO conditions. As a consequence, the neighborhood characteristics of a particular word change. At the two-viseme level, all CVC words are roughly segmentally equivalent, so differences in neighborhood density should become irrelevant. What is surprising is that the advantages in accuracy that are normally afforded by Easy lexical status seem to disappear when scores are calculated using this most lax criterion. In the section on lexical factors included below, one explanation for this effect is proposed.

In addition to lexical factors, it is also clear that the initial analyses have revealed a great deal of variation in the properties of the talkers and the stimulus items themselves. Most importantly, scores computed using "real" viseme sets were consistently more accurate than the scores based on "random" visemes. This means that increases in accuracy due to relaxation of the visemic criterion were not due to changes in chance-level performance. Instead, some aspect of the underlyingly common confusion space of lipread phonemes was being picked up and captured in the scores computed using the generic, empirically motivated visemes taken from Auer and Bernstein (1997).

Inter-talker differences were also found in differing levels of accuracy across the varying measures and perceptual confusion (viseme) criteria. Several patterns in the variation, however, emerged from more detailed analyses. M1 is clearly an outlier in terms of his visual intelligibility. Under all measures and at all visemic criterion levels, he was the least intelligible of the ten talkers, displaying the lowest recognition scores. In addition, the fact that the rank ordered intelligibility of each talker varied across viseme levels implies that some viseme levels more accurately reflect the underlying structure of each particular talker's phonemic confusion space than others. If each relaxation of the visemic criterion affected the scores for each talker in the same way, then the ranks would have remained the same, regardless of the criterion. A more detailed analysis of these inter-talker differences was carried out using the transitional measures of accuracy. These transitional scores measure the amount of improvement in accuracy as a result of a relaxed visemic criterion.

Second-Order Measures of Accuracy

Transitional measures were also calculated on this data set using relaxations of the criterion in "real" and "random" viseme sets. Changes in accuracy due to relaxation of the criterion in "random" viseme sets are taken to represent the amount of improvement expected due to chance. Changes in accuracy due to relaxation of the criterion in "real" sets are taken to represent the amount of improvement due to a scoring system that more adequately captures the perceptual confusability of a particular talker's utterances. Thus, the *difference* between the transitionals calculated on the "real" and "random" sets for a particular relaxation represents the amount of improvement across a transition due solely to a more accurate model of the underlying perceptual confusability of a particular talker's articulations, and not to improvements that would be expected with any haphazard relaxation of the criterion for a correct score.

In all the analyses of the transitional measures reported below, these difference scores were used as the dependent measure. As with the first-order measures, the FL class of transitionals has a different interpretation from those of the ORD and SGM classes, and so will be discussed separately.

FL Class Transitionals. Because a FL score can only have one of two possible values (1 for correct, 0 for incorrect), the second order transitional FL scores are limited to one of three possible values. If the target-response pair was incorrect at the more strict criterion involved in the transition and is correct at the more lax criterion, then the transitional score will be 1. If the target-response pair *remains* incorrect across the transition, then the transitional score will be 0. Transitional scores for target-response pairs that remain *correct* across the transition are undefined and not used during final analysis. If the target-response pair was correct at the more strict criterion, but is incorrect at the more lax criterion, then it will receive a transitional score of -1. Note that this last score is only possible when using randomly assigned visemes, since "real" visemes are subsets of one another (if a target-response pair is scored as correct using "real" visemes at one criterion, then it will be scored as correct at every more lax criterion).

These properties yield an interesting interpretation of the transitional FL scores. The average transitional FL score for a given transition is essentially the proportion of target-response pairs that become

correct across the transition. However, it must be noted that gains in accuracy would be made due to the relaxation of *any* criterion for accuracy, and so the transitional scores computed using randomly assigned visemes are subtracted from the "real" viseme transitional scores. This yields a measure of the proportion of target-response pairs for which the transition is informative, irrespective of the gains due to chance.

Now, recall that the segmental confusion patterns for a particular talker can be described as perceptual similarity spaces. These are assumed in the Confusability Continuum Heuristic to be subspaces of an underlyingly common segmental confusion space, translated with respect to the particular criterion for perceptual equivalence that describes a particular talker-speechreader interaction. The criterion is represented as a number on some continuum. High values of the criterion represent perceptual spaces in which segments are highly perceptually distinct from one another. Low values of the criterion represent perceptual spaces in which segments collapse onto one another to form perceptual equivalences. The six viseme sets used here are assumed to represent sampling this continuum at various points. Transitional scores, then, represent the gains in accuracy expected from shifting along the continuum between two sampling points. For example, if a particular transition yields a high proportion of target-response pairs that are correct, then is very likely that the criterion point that most adequately describes the talker is contained somewhere within the transition between the two sampling points. In fact, the transitional FL score for a particular transition can be said to represent the degree of support for the notion that the criterion point that describes a talker lies within that transition.

All ten talkers were ranked at each transition by the difference between their "real" and "random" FL accuracy improvement. Table 7 shows these ranks for Easy words. Table 8 shows the same data for Hard words. The "P_28" column represents the change in accuracy when the criterion is shifted from the strictest, full phoneme set to the 28 viseme set. The "28_19" column represents the change in going from 28 visemes to 19 visemes. The "19_12" column is the change in accuracy between 19 and 12 visemes. The "12_10" and "10_2" columns represent changes in accuracy across the 12 to 10 viseme set transitions and across the 10 to 2 viseme set transitions, respectively. None of the ranked scores were negative. This means that transitions computed using "real" viseme sets were higher on average than transitions computed using "random" viseme sets. Once again, we see that improvements in accuracy across these transitions were not simply due to an increased chance of being correct, but rather due to the fact that the relaxations of the criterion for being correct were conducted along dimensions that were perceptually difficult to resolve.

Easy	**P_28**	**28_19**	**19_12**	**12_10**	**10_2**
F1	5	1	3	1	7
F2	9	5	8	4	9
F3	2	6	5	2	4
F4	1	4	1	5	2
F5	3	2	2	8	8

M1	10	10	10	10	6
M2	6	7	7	9	5
M3	8	9	9	7	10
M4	4	3	5	3	1
M5	7	8	6	6	3

Table 7. The ten talkers ranked by the average difference between "real" and "random" FL accuracy improvement for Easy words at each transition.

Hard	**P_28**	**28_19**	**19_12**	**12_10**	**10_2**
F1	6	2	8	1	8
F2	5	7	3	7	7
F3	2	6	1	2	5
F4	3	1	4	3	3
F5	1	3	5	8	10
M1	10	10	10	10	2
M2	8	4	2	9	6
M3	4	9	6	6	9
M4	7	8	7	5	4
M5	9	5	9	4	1

Table 8. The ten talkers ranked by the average difference between "real" and "random" FL accuracy improvement for Hard words at each transition.

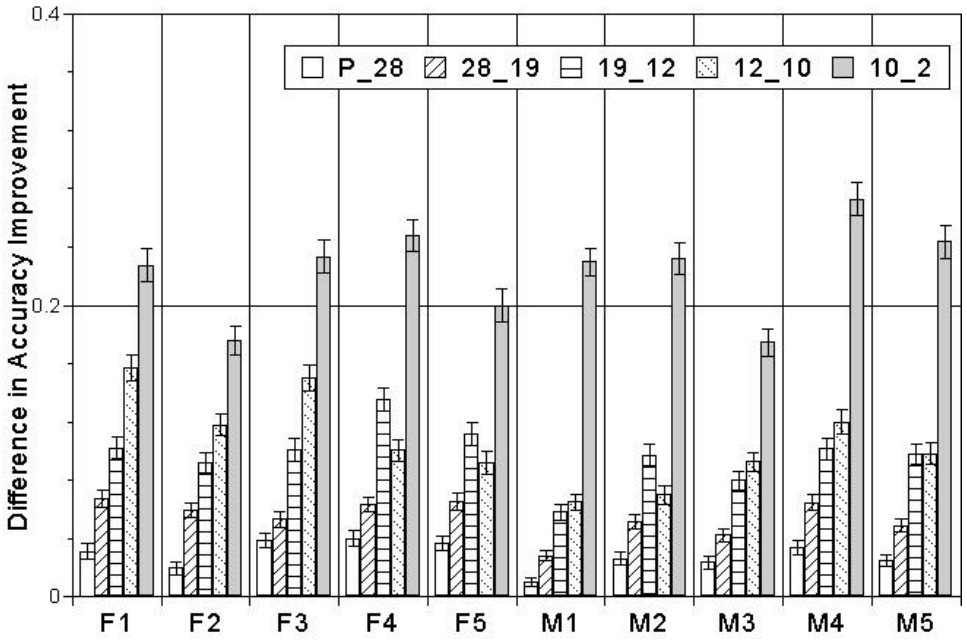


Figure 7. Average difference in FL transitional score between real and random viseme sets for Easy words. Each panel represents data from a different talker, and each column type represents a different transition.

Some talkers benefited more from some transitions than others. M1 benefited the least among all the talkers from all but the most lax transition from 10 to 2 visemes, for both Easy and Hard words. In contrast, with Hard words, F5 benefited the most among all the talkers from the phoneme to 28 viseme transition, but benefitted the least from the last 10 viseme to 2 viseme transition. For both the Easy and Hard words, F1 and F3 were helped the most by the 12 to 10 viseme transition. In addition, the ranks for some talkers remained relatively constant, like M1 and M4, while the ranks for other talkers vary wildly.

Although the rank tables are easy to understand, they do not capture the full extent of talker-based differences in the transitional scores. Figures 7 and 8 show the average difference in the transitional FL scores computed using real and random viseme sets for each talker. Figure 7 shows the scores for Easy words and Figure 8 shows the data for Hard words. Each panel in each graph represents the transitional scores computed for a particular talker. Within each panel, each bar represents the difference between the transitional scores computed using “real” and “random” viseme sets for a particular transition. The “P_28” bar represents the change in accuracy in shifting the criterion from the strictest, full phoneme set to the 28 viseme set. The “28_19” bar represents the change in going from 28 visemes to 19 visemes. The “19_12” bar is the change in accuracy between 19 and 12 visemes. The “12_10” and “10_2” bars represent changes in accuracy across the 12 to 10 viseme set transitions and across the 10 to 2 viseme set transitions, respectively.

It is apparent that for all talkers a higher proportion of target-response pairs became correct at the 10 to 2 transition than for any other transition, although the precise number varied across talkers. This is true for both Easy and Hard words. However, closer inspection of the figures shows substantial variation

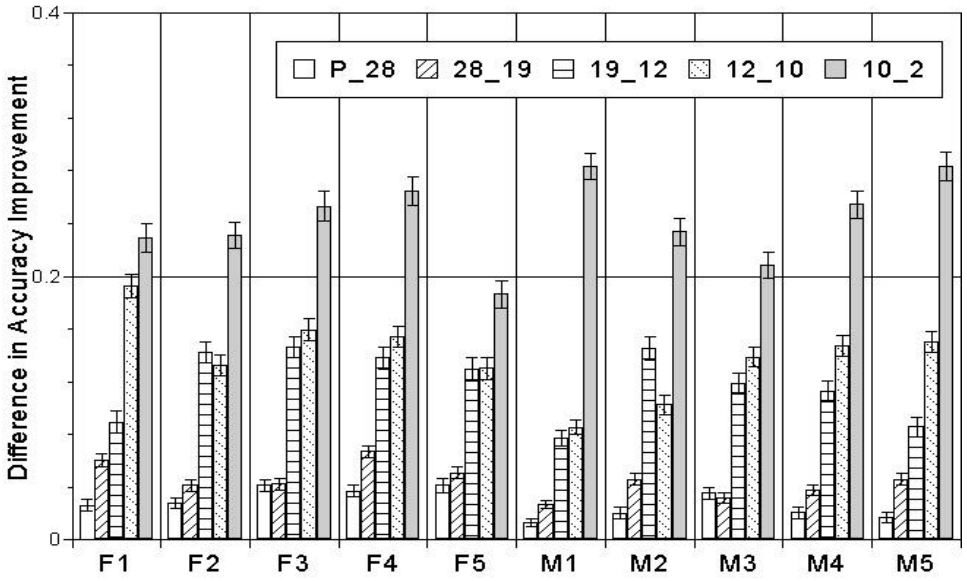


Figure 8. Average difference in FL transitional score between real and random viseme sets for Hard words. Each panel represents data from a different talker, and each column type represents a different transition.

in the degree to which each transition helped the responses for each talker. For example, for easy words (Figure 7) the transitional scores computed for M1 indicate that relatively little advantage was gained across the more strict transitions (phonemes to 28 visemes, and 28 visemes to 19 visemes). However, for other talkers, stricter transitions were informative for a greater proportion of words. In addition, steady improvements in the transitional score are not always observed as the criteria are relaxed. F4, F5 and M2 are all helped more by the 19 to 12 viseme transition than the 12 to 10 viseme transition.

ORD and SGM Class Transitionals. The segmentally-based transitional scores have a different interpretation from the FL class transitionals. The ORD and SGM class transitionals represent the amount by which the transition affected accuracy for a particular target-response pair. Averages across these measures, then, will represent the average amount by which accuracy for target-response pairs changed at a particular transition. Table 9 shows the ranked difference in ORD and SGM transitional accuracy for Easy words. Table 10 shows the same data for Hard words. As before, the "P_28" column represents the change in accuracy in shifting the criterion from the strictest, full phoneme set to the 28 viseme set. The "28_19" column represents the change in going from 28 visemes to 19 visemes. The "19_12" column is the change in accuracy between 19 and 12 visemes. The "12_10" and "10_2" columns represent changes in accuracy across the 12 to 10 viseme set transitions and across the 10 to 2 viseme set transitions, respectively. None of the ranked scores were negative (i.e., transitions computed using "real" viseme sets were higher on average than transitions computed using "random" viseme sets).

As with the FL transitional scores, there is substantial variation in the talkers most aided by the various transitions. In fact, there is even less consistency for these segmentally based transitionals than there was for the FL transitional scores. Although M1 is ranked lowest or close to the lowest for many of the transitions, all the other talkers are ranked somewhat differently at each transition. F2 is ranked highest for the 19 to 12 viseme transition using both ORD and SGM scores, but is ranked at or near the lowest for the phoneme to 28 viseme transition and the 10 viseme to 2 viseme transition.

	ORD					SGM				
Easy	**P_28**	**28_19**	**19_12**	**12_10**	**10_2**	**P_28**	**28_19**	**19_12**	**12_10**	**10_2**
F1	3	1	7	1	3	6	2	8	1	1
F2	10	4	1	6	10	9	7	1	6	9
F3	4	7	3	2	5	4	6	3	2	6
F4	6	3	4	9	1	5	1	5	8	3
F5	2	2	5	8	8	1	4	6	7	7
M1	9	10	10	5	4	8	10	10	4	2
M2	5	8	2	10	2	2	9	2	10	4
M3	7	9	8	7	9	8	8	7	9	10
M4	1	5	6	4	7	3	3	4	5	8
M5	8	6	9	3	6	10	5	9	3	5

Table 9. The ten talkers ranked by the average difference between "real" and "random" ORD and SGM accuracy improvement for Easy words at each transition.

In order to more fully capture and represent this variation, Figure 9 shows the average difference between "real" and "random" ORD accuracy improvement for Easy words. Figure 10 shows the same differences for Hard words. Each panel in each graph represents the transitional scores computed for a particular talker. Within each panel, each bar represents the difference between the transitional scores computed using "real" and "random" viseme sets for a particular transition. The "P_28" bar represents the change in accuracy in shifting the criterion from the strictest, full phoneme set to the 28 viseme set. The "28_19" bar represents the change in going from 28 visemes to 19 visemes. The "19_12" bar is the change in accuracy between 19 and 12 visemes. The "12_10" and "10_2" bars represent changes in accuracy across the 12 to 10 viseme set transitions and across the 10 to 2 viseme set transitions, respectively.

It is evident from inspection of these figures that there are marked differences in the degree to which each transition helped or hindered the scores given for a particular talker. With the Easy words, there is no consistently more advantageous transition for all the talkers, although the transition from 28 to 19 visemes tends to provide a big boost to ORD scores. However, M1 receives the most gain in accuracy at the 12 to 10 viseme transition. Furthermore, although one might expect that large gains in accuracy would be made in going from 10 to 2 visemes, it is clear that the amount by which this relaxation improves scores varies substantially across the ten different talkers.

	ORD					SGM				
Hard	**P_28**	**28_19**	**19_12**	**12_10**	**10_2**	**P_28**	**28_19**	**19_12**	**12_10**	**10_2**
F1	6	1	10	1	3	4	3	7	1	10
F2	4	4	3	8	4	6	6	3	9	5
F3	2	6	2	2	7	5	9	1	2	7
F4	1	7	7	3	2	1	2	8	3	2
F5	3	3	4	7	10	2	4	5	8	9
M1	9	10	8	10	1	9	10	9	7	1
M2	7	9	1	9	8	8	5	2	10	6
M3	5	8	5	6	9	3	7	4	5	8
M4	8	5	6	5	6	7	8	6	6	4
M5	10	2	9	4	5	10	1	10	4	3

Table 10. The ten talkers ranked by the average difference between "real" and "random" ORD and SGM accuracy improvement for Hard words at each transition.

When the ORD transitional scores are computed for Hard words, a different pattern emerges. For the most part, all talkers display their biggest gain in accuracy between the 12 and 10 viseme sets, however, the degree to which this transition is better than the next best transition differs across the talkers. For most of the talkers, the next best transition is the 28 to 19 viseme set. M2 diverges from this pattern, showing relatively equal gains in accuracy for the middle three transitions.

Although the ranks of the transitional SGM scores are similar to those of the ORD scores, an important difference between the ORD and SGM transitional scores does exist. With the ORD transitional data, one common pattern for both the Easy and Hard is that improvements in the "real" scores are always better than improvements in the "random" scores, as indicated by the consistently positive value of the differences graphed in Figures 9 and 10. This is not true, however, for the transitional SGM scores graphed in Figures 11 (which shows scores for Easy words) and 12 (which shows scores for Hard words). Both figures show that "random" transitional SGM scores were better than their "real" counterparts at the final 10 viseme to 2 viseme transition, as indicated by the negative values at the "10_2" bars for all talkers.

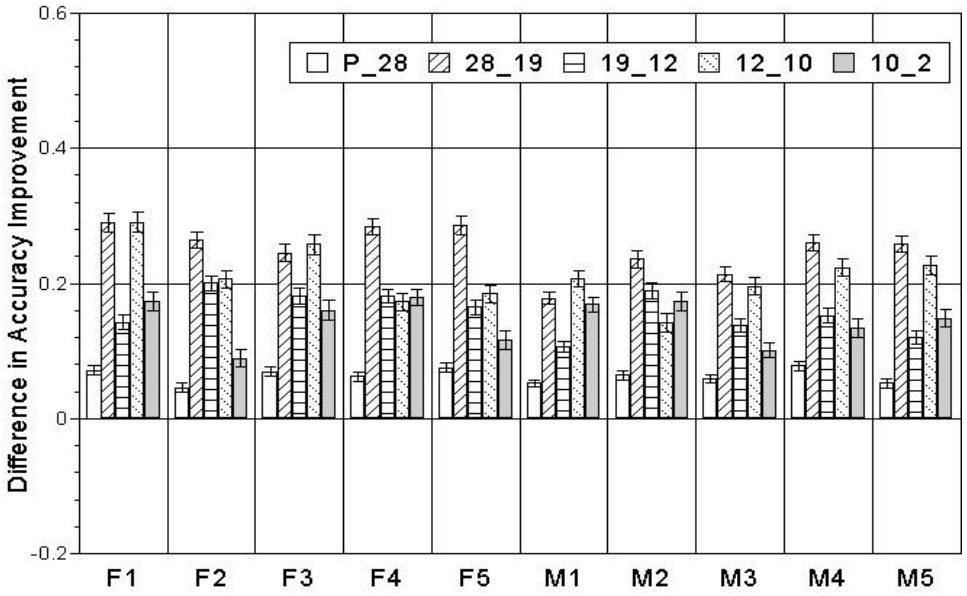


Figure 9. Average difference in ORD transitional score between real and random viseme sets for Easy words. Each panel represents data from a different talker, and each column type represents a different transition.

This finding most clearly illustrates the difference between the ORD and SGM classes of measures. Why should SGM scores computed using random viseme sets be better than those using "real" viseme sets, when the opposite is true for the ORD (and FL) class? Because ORD accuracy takes into account the order in which target segments occur in the response, it reveals the extent to which responses have captured the phonotactic and lexical constraints presumably available to all normal-hearing speechreaders. SGM accuracy, which ignores the order of occurrence, does not pick up on this aspect of response structure. The discrepancies between the patterns of ORD and SGM transitional scores must arise, then, because speechreaders are able to capitalize on the internal structure of words and the patterning of sound segments when making their responses based on lipreading.

Indeed, when using SGM scoring at the randomly-assigned 2 viseme level, the chances of any response segment being correct are roughly 50:50. The odds are not as good for the "real" 2 viseme set, because consonants and vowels are distributed into the 2 visemes less equally: there are more consonants than vowels. With ORD scoring at this level, the constraint of order narrows even further the chances that a particular segment will be correct. Keep in mind, however, that this does not necessarily mean that *first order* SGM scores will be higher for random sets than for real sets, because participants *are* capable of capitalizing on the robust perceptual distinction between consonants and vowels. In fact, the data reviewed above in the section on first order accuracy confirms this fact very clearly.

So, why is the difference observed for the second order transitional measures? At stricter viseme criteria, the random assignment of segments to visemes suppresses accuracy because it ignores the structure of perceptual confusability that speechreaders evidently use to constrain their guesses. Thus, on average, transitional scores will remain low. However, at the 2 viseme level, the random assignment suddenly becomes advantageous; any response strategy will be virtually indistinguishable from that of

random guessing, since the probability of getting a particular segment correct will always be close to 50%. This drastic shift in the utility of the random viseme structure yields, on average, high transitional

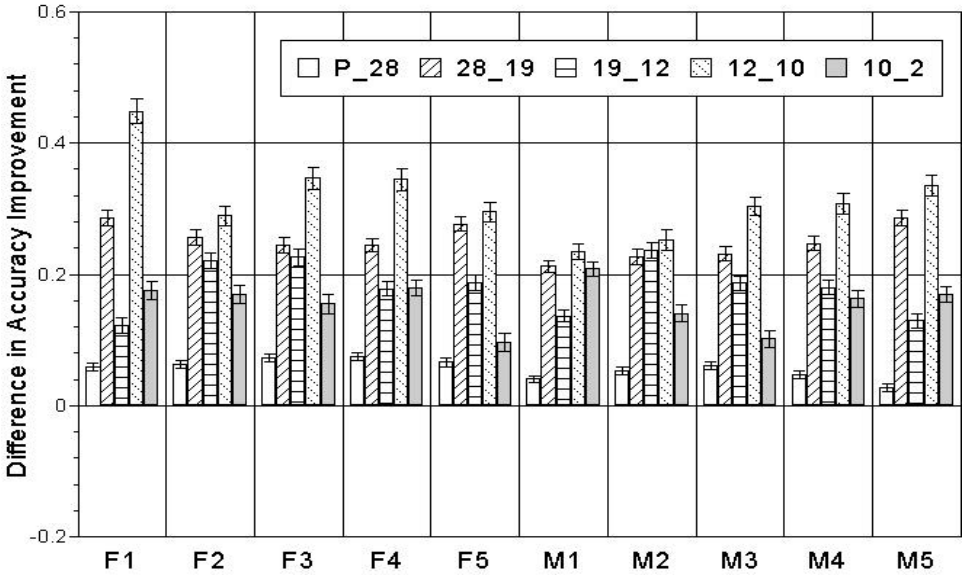


Figure 10. Average difference in ORD transitional score between real and random viseme sets for Hard words. Each panel represents data from a different talker, and each column type represents a different transition.

scores. No such shift, however, occurs for the scores computed using "real" visemes. The utility of the "real" perceptual confusion structure increases, by definition, gradually, because it is based on gradually relaxing criteria for perceptual equivalence. Consequently, the transitional scores using real viseme sets will tend to be lower than the transitional scores using random viseme sets, but only for the 10 to 2 transition.

Summary of Second Order Transitional Measures of Accuracy. Overall, the data from the second order measures of accuracy show that the extent to which a particular relaxation of the viseme criterion increases accuracy is dependent on the person speaking. If we accept the Confusability Continuum Heuristic and the notion that the underlying perceptual confusion spaces associated with talkers lie along a continuum of perceptual confusability, then we can interpret the current differences as indicating that the talkers in the HAVMD can be described by a range of points along that hypothetical continuum. Future studies will be able to capitalize on this variability in order to further elucidate those characteristics that make some talkers easier to speechread than others.

The Role of Lexical Factors

Because the Easy/Hard distinction was so robust for all measures under all viseme criterion levels, further analysis of the role of lexical factors in accuracy was conducted. For each target word, the Hoosier Mental Lexicon (Nusbaum, Pisoni, & Davis, 1984) database was searched for an entry that matched in transcription. All target words were found in the HML. When it was found, the log frequency, neighborhood density, and mean log neighborhood frequency were recorded. In contrast to the Easy/Hard distinction, which assigns ordinal values to these variables, the HML data provided a ratio scale measurement of these three important lexical factors, enabling regression analyses on the various dependent measures.

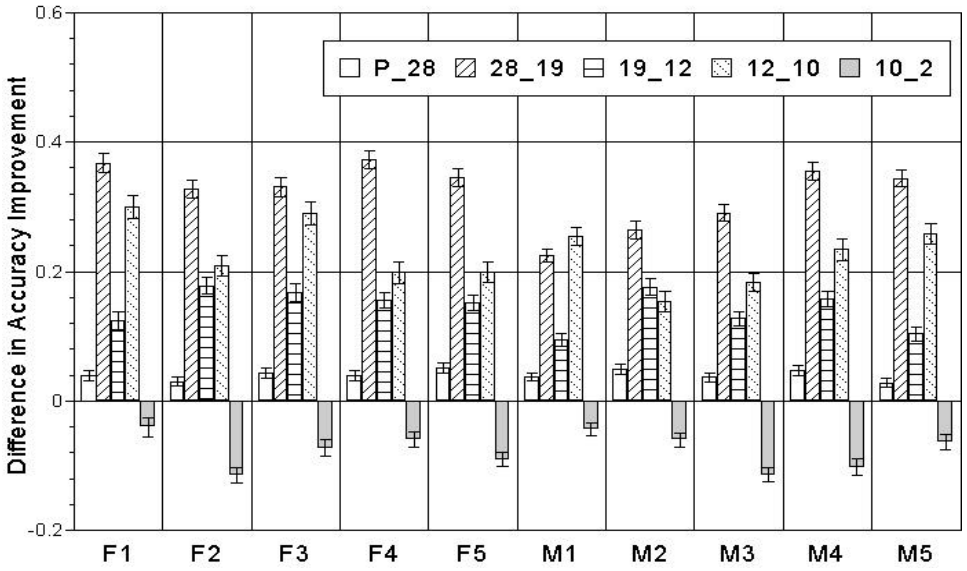


Figure 11. Average difference in SGM transitional score between real and random viseme sets for Easy words. Each panel represents data from a different talker, and each column type represents a different transition.

Stepwise multiple regression analyses were carried out using the log frequency, the neighborhood density and the mean log frequency of the target words as predictor variables. Predictor variables were entered into the regression equation with a p value of 0.05, and removed with a p value of 0.10. A separate regression analysis was performed using each viseme criterion level in each of the ORD and SGM classes. Figure 13 shows the R^2 values associated with the best fitting regression line for each viseme criterion level using ORD scores and SGM scores. The fact that the values are so low reiterates an important point about these data; they are extremely variable and noisy. Although all differences are significant with the powerful sample size used here, the speechreading process is extremely complex and the number of relevant factors is enormous. A simple picture of this phenomenon is unlikely. In addition, it is clear that, as the criterion for perceptual equivalence is relaxed, the picture becomes more complicated. Relatively little variance is accounted for by the lexical variables when a lax criterion for accuracy is used. It is only when the observer is required to make fine phonetic distinctions that lexical variables play a role in identification performance.

One interesting property of the regression analyses emerges between the 19 and 12 viseme criteria. As mentioned above, this transition seems to be special in some way. As Figure 13 shows, it is at this point that there seems to be a drastic drop in the R^2 values associated with each equation. Similarly, this change is reflected in the predictive qualities of the lexical variables in the equations, as well. The β weights associated with the lexical factors entered into the best-fitting regression equation for predicting the various measures are shown in Table 11. Up to the 19 viseme level, the effect of these variables is consistent with those predicted by the NAM (Luce & Pisoni, 1998): neighborhood density of the target word is negatively related to accuracy. Targets from large similarity neighborhoods are identified less accurately than those from small similarity neighborhoods. In addition, the frequency of the item is positively related to accuracy. Higher frequency words are recognized with higher accuracy than those of lower frequency.

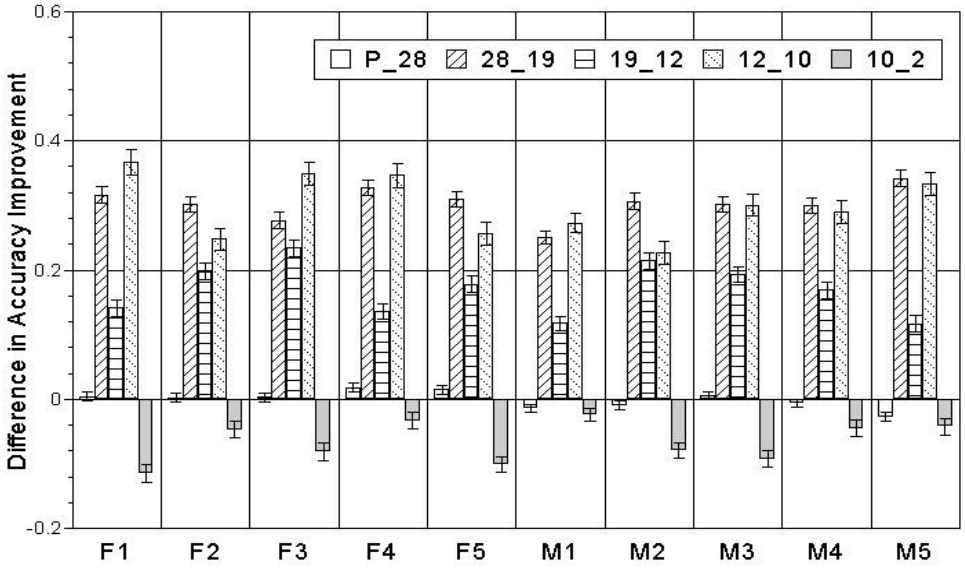


Figure 12. Average difference in SGM transitional score between real and random viseme sets for Hard words. Each panel represents data from a different talker, and each column type represents a different transition.

After the 19 to 12 viseme transition, however, these relationships reverse direction. In both the ORD and SGM classes of measures, density reverses its effect, if it had an effect at all, so that it is directly related to accuracy. Even more surprisingly, frequency becomes indirectly related to accuracy. The same trend, at least, is shown for the mean log neighborhood frequency.

Although further investigation needs to be conducted to determine why these changes happen at this transition point, the basic pattern can be explained in terms of the properties of the theoretical lexical similarity space in which these targets reside as it is collapsed across visually indistinguishable dimensions (Auer & Bernstein, 1997). Recall that expected class size is the number of words expected to be visually equivalent to a target after visemic transcription. Words that are visually equivalent form equivalence classes. Although Auer and Bernstein do not discuss the issue *per se*, words in dense neighborhoods are going to be more likely to collapse into equivalence classes with high class sizes, because there are simply more words with the potential to be visually equivalent. This is because neighbors differ from a target word in only one phoneme. With more neighbors comes the increased possibility that, after visual transcription, the last, differing phoneme will also be considered visually equivalent. This likelihood increases as the viseme criterion is relaxed. Thus, at the most lax criteria, simply picking a response from the neighborhood of the target word will usually be sufficient for a correct response.

This pattern holds at stricter criteria as well but is not as straightforward because subtle differences in the neighborhood structure of specific words will determine whether density plays a role or not. A target word might have one neighbor that doesn't collapse into the target's equivalence class until the most lax criterion, while another might only have neighbors that collapse into the target's equivalence class at the strictest criterion. For example, the word “bat” has at least one neighbor (“cat”) that remains visually

distinct at all but the most lax criterion point, whereas the word “bomb” has at least three neighbors that collapse at the very first relaxation of the criterion (“mom”, “mob”, “bum”).

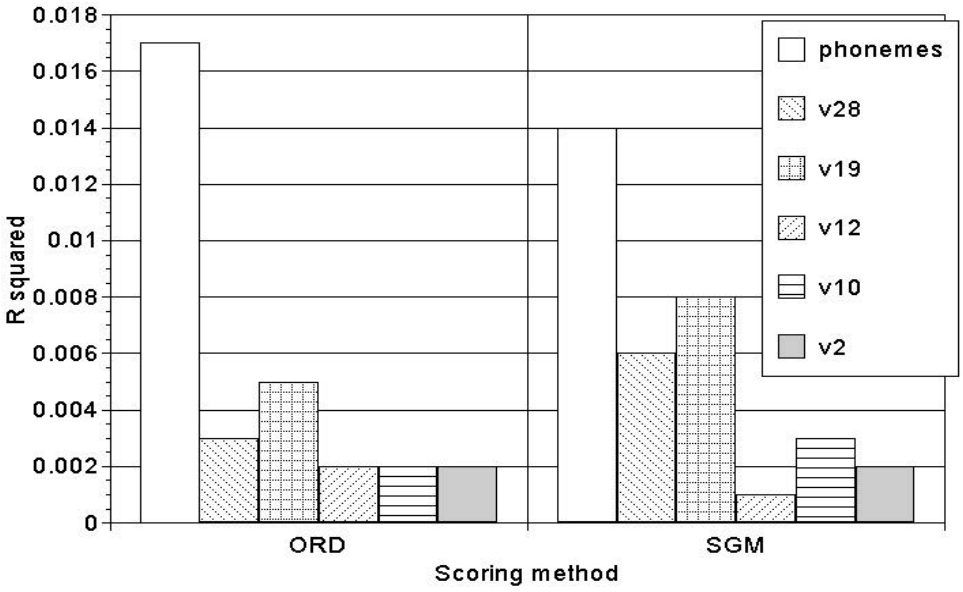


Figure 13. R^2 or percent variance accounted for by the best-fitting regression equation for accuracy using ORD and SGM scores at each of the visemic criterion levels. The predictor variables were log frequency, neighborhood density and mean log frequency of the target word.

But why would word frequency reverse its relationship with accuracy beyond the 19 to 12 viseme transition? One possible explanation of this effect may lie with the talker's implicit knowledge of word frequency and the way in which this knowledge shows up in the production of spoken words. Wright (1997) showed that vowels were more hypoarticulated (centralized) in Easy words than in Hard words. If vowels tend to be hypoarticulated in Easy words, then it may be the case that, in these words, the CVC structure of a target word is less perceptually robust than they are in Hard words. Scores at the 2-viseme level reflect the accuracy of a speechreader in perceiving only this kind of structure. The negative relationship between frequency and accuracy at the 2-viseme level may be due to the decreased distinction between consonants and vowels in Easy words. Of course, the current study does not address this problem specifically, and more work will be needed in order to confirm the validity of this hypothesis.

Measurement Class	Criterion Level	Neighborhood Density	Log Frequency	Mean Log Neighborhood Frequency
ORD	Phonemes	-0.070	0.494	-0.480
	28 visemes	-0.054	n/a	n/a
	19 visemes	-0.069	n/a	n/a
	12 visemes	n/a	0.056	-0.093
	10 visemes	0.019	-0.050	n/a
	2 visemes	0.018	-0.050	n/a
SGM	Phonemes	-0.058	0.441	-0.439
	28 visemes	-0.032	0.272	-0.0289
	19 visemes	-0.038	0.305	-0.326
	12 visemes	n/a	0.069	-0.103
	10 visemes	0.036	-0.055	n/a
	2 visemes	n/a	-0.231	0.226

Table 11. β-weights associated with lexical variables in the various regression analyses. Cells marked "n/a" were not entered into the best fitting regression equation.

Summary and Conclusions

The data reported here described in fuller detail the VO properties of the Hoosier Audiovisual Multitalker Database, a collection of isolated audiovisual spoken words. Several methods of scoring speechreading accuracy for isolated words have been proposed and the ways in which these measures are related was explored in detail. In addition, it was shown how each measure is useful for revealing different aspects of the structure contained in lipread responses to spoken words.

From the analyses above, it is evident that the stimuli in the HAVMD represent a useful sample of tokens with properties that have relevance to current issues in the perception of spoken language in visual only contexts. The 10 talkers filmed for these stimuli differ in intelligibility and their tokens can be used for further investigations into the way in which talker-specific properties of lipread speech affect the performance of speechreaders. In addition, differences in the lexical properties of the target words of the

HAVMD were shown to have effects on lipreading performance, a feature that will be valuable in future studies on the way in which lexical sources of knowledge interact with perceptual processes during VO spoken word recognition.

In the course of fully describing the VO responses to identification of the tokens in the HAVMD, I have made several assumptions referred to here as the Confusability Continuum Heuristic. The CCH proposes that the underlying perceptual confusability of speech segments in VO environments is constant. The idiosyncratic confusion patterns associated with a particular talker, in a particular environment, and viewed by a particular speechreader are transformations of these underlying confusion patterns, collapsed across dimensions that do not remain distinct in such settings. Furthermore, the CCH assumes that these dimensions collapse in a hierarchical way, such that certain dimensions must collapse before others can, *in all situations*. Using the CCH, it is possible to describe the segmental confusion space associated with speechreading in a particular context as a point on the continuum implied by the hierarchy. In order to evaluate the data set presented here, I have assumed that the empirically motivated set of visemes presented by Auer and Bernstein (1997) represent sampling points along this hypothetical continuum.

Further work must be conducted in order to determine whether the CCH provides a workable metaphor for describing the ways in which visual confusions across settings relate to each other. A major study involving the examination of segmental confusability using multiple talkers, speechreaders and viewing conditions may be able to do this. For now, however, it seems like a reasonable heuristic, given that certain distinctions, like those between [b], [p], and [m], are nearly always indistinguishable in the majority of viseme sets proposed in the literature, while others, like those between [b] and [g], are not.

In any event, the assumptions of the CCH have revealed interesting aspects of the structure of identification responses to VO stimuli that warrant further investigation. First, it is abundantly clear that much more information is available to speechreaders in VO environments than would otherwise be thought after examining accuracy using a strict, phoneme-based metric. Granting accuracy “credit” for responses that differ from the target in ways that are imperceptible when presented VO reveals that many “incorrect” responses are in fact extremely close to the intended target. In fact, as the criterion for perceptual confusability is relaxed, accuracy scores improve steadily. Furthermore, this improvement is not due to a simple improvement in the odds of getting particular segments correct. First order scores computed using the random viseme sets, which were matched with the real visemes in the number of classes available, were consistently lower than first order scores computed using real viseme sets. Thus, improvement with relaxation of the visemic criterion reflected the fact that the relaxed viseme sets were better approximations to the “actual” information available to speechreaders. In addition, the amount by which real viseme scores were better than random scores increased as the relaxation of the visemic criterion increased, indicating that relaxations of the real viseme set were *increasingly* better approximations than the random viseme sets.

Another interesting result that emerged from the first order measures was the fact that the differences in accuracy observed for Easy and Hard words decreased with increased relaxation of the visemic criterion. As the requirement for fine-grained phonetic detail in responses was dropped, lexical factors did not play as much of a role in determining the accuracy of responses. This property was also observed in the regression analyses, where it was shown that, with respect to accuracy, the predictive value of word frequency, neighborhood density and mean neighborhood frequency dropped off steeply as more and more coarsely grained information was accepted as correct. In fact, some of these lexical variables reverse their effects at the more relaxed visemic criteria.

Of particular interest was the apparently critical transition in spoken word recognition between 19 and 12 visemes. The data from Auer and Bernstein's (1997) study of visually transcribed lexicons implies

that some major changes occur between these two levels, with the number of unique words dropping rapidly after the 19 viseme level. It will be interesting to test in the future whether the effects found here over this transition are based on the properties of visually transcribed lexicons, and whether these properties have other effects on speechread spoken word recognition.

The different classes of measures also showed the use of partial information. The differences between accuracy calculated using the FL ("full") class and accuracy calculated using the ORD ("ordinal") and SGM ("segmental") classes demonstrated that even when the *entire word* wasn't perceived accurately, some proportion of the word was recognized. Furthermore, as the visemic criterion was relaxed, differences in the perceptibility of segments in a word became less and less influenced by lexical factors. This was not due to relaxation of the criterion for correct segmental identification, since the difference between real and random scores changed at different rates across the visemic criterion points for Easy and Hard words.

Finally, inter-talker differences were evident with all scoring methods, but the transitional scoring method showed most clearly that relaxations of the visemic criterion affected accuracy on different talkers to different extents. However, all of the effects mentioned above varied from talker to talker, indicating that the CCH is not a perfect metaphor for the variability inherent in various talker-speechreader systems. Clearly, no one set of visemes will suffice to describe the perceptual phonetic confusion spaces of all talkers, being observed by all speechreaders, in all conditions. The properties of different talkers and the ways in which those properties affect segmental confusability will be of great value in determining the precise nature of the information available to speechreaders.

The study of spoken word recognition in visual only contexts is still relatively unexplored and many aspects of this phenomenon await discovery. As demonstrated here, even a simple identification task can reveal much about the relevant factors involved. It is hoped that the HAVMD and the analyses above will provide useful tools in future investigations of spoken word recognition in both uni- and multi-modal contexts.

References

Auer, E. T. & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of refuced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America, 102(6),* 3704 - 3710.

Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment. *Journal of the Acoustical Society of America, 95*, 3617 - 3622.

Boothroyd, A. (1988). Linguistic factors in speechreading. *The Volta Review, 90(5),* 77 - 87.

Demorest, M. E., & Bernstein, L. E. (1991). Computational explorations of speechreading. *Journal of the Academy of Rehabilitative Audiology, 24*, 85 - 96.

Dodd, B. & Burnham, D. (1988). Processing speechread information. *The Volta Review, 90(5)*, 45 - 60.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11,* 796 - 804.

Gagné, J-P, Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in audiotry, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology, 27*, 135 - 158.

Gagné, J-P, Tugby, K.G., & Michaud, J. (1991). Development of a speechreading test on the utilization of contrxtual cues (STUCC): Preliminary findings with normal-hearing subjects. *Journal of the Academy of Rehabilitative Audiology, 24*, 157 - 170.

Goh, W. D. & Pisoni, D. B. (1998). Effects of lexical neighborhoods on immediate memory span for spoken words: A first report. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 195 - 213). Bloomington, IN: Speech Research Laboratory, Indiana University.

Huttenlocher, D. P., & Zue, V. W. (1984). A model of lexical access from partial phonetic information. Paper presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA. 1- 4.

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90(5),* 99 -115.

Jeffers, J. & Barley, M. (1971). *Lipreading (Speechreading).* Springfield, IL: Charles C. Thomas.

Kricos, P. B. & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review, 84*, 219 - 225.

Kricos, P. B. & Lesner, S. A. (1985). Effect of talker differences on the speechreading of hearing-impaired teenagers. *The Volta Review, 87,* 5 - 16.

Lachs, L. and Hernàndez, 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.

Lansing, C. R., and Helgeson, C. L. (1995). Priming the visual recognition of spoken words. *Journal of Speech and Hearing Research, 38*, 1377 - 1386.

Lesner, S. A. (1988). The talker. *The Volta Review, 90(5)*, pp. 89 - 98.

Lesner, S. A. and Kricos, P. B. (1981). Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology, 14*, 252 - 258.

Lively, S.E., Pisoni, D.B., & Goldinger, S.D. (1994). Spoken word recognition: Research and theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, New York: Academic Press. Pp. 265-301.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1 – 36.

Lyxell, B. & Rönnberg, J. (1989) Information-processing skill and speechreading. *British Journal of Audiology, 23*, 339 - 347.

Massaro, D. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

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

Miller, G. & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338 - 352.

Montgomery, A. A. & Jackson, P.L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America, 73*, 2134 - 2144.

Montgomery, A. A., Walden, B. E., & Prosek, R. H. (1987). Effects of consonantal context on vowel lipreading. *Journal of Speech and Hearing Research, 30*, 50 -59.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115(1)*, 39 -57.

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 - 377). Bloomington, Indiana University, Department of Psychology, Speech Research Laboratory.

Sheffert, S., Lachs, L. & Hernandez, L. R. (1996-1997). 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.

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. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye* (pp. 5 - 31). London: Erlbaum.

Walden, B.E., Prosek, R.H., 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.

Wright, R. (1996). Lexical competition and reduction in speech: A preliminary report. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 471 - 485). Bloomington, IN: Speech Research Laboratory, Indiana University.

Appendix A

Real viseme sets:

28 visemes: {u ju}{ʊ}{ə˞}{oʊ}{aʊ}{ɪ i}{ei ɛ}{æ}{ɔi}{ɔ}{ai}{ə ɑ ʌ j}{b p m m̩}
{f v}{l syll. l}{n n̩}{k}{ŋ g}{h}{d}{t}{s z}{w ɹ}{ð ɵ}{ʃ}{t͡s}{ʒ}{d͡ʒ}

19 visemes: {u ju ʊ ə˞}{oʊ aʊ}{ɪ i}{ei ɛ}{æ}{ɔi}{ɔ}{ai ə ɑ ʌ j}{b p m m̩}
{f v}{l syll. l}{n n̩ k}{ŋ g}{h}{d}{t s z}{w ɹ}{ð ɵ}{ʃ t͡s ʒ d͡ʒ}

12 visemes: {u ju ʊ ə˞}{oʊ aʊ}{ɪ i ei ɛ æ}{ɔi}{ɔ}{ai ə ɑ ʌ j}{b p m m̩}{f v}
{l syll. l n n̩ k ŋ g h}{d t s z}{w ɹ}{ð ɵ}{ʃ t͡s ʒ d͡ʒ}

10 visemes: {u ju ʊ ə˞}{oʊ aʊ}{ɪ i ei ɛ æ}{ɔi ɔ ai ə ɑ ʌ j}{b p m m̩}{f v}
{l syll. l n n̩ k ŋ g h d t}{s z}{w ɹ}{ð ɵ}{ʃ t͡s ʒ d͡ʒ}

2 visemes: {u ju ʊ ə˞ oʊ aʊ ɪ i ei ɛ æ ɔi ɔ ai ə ɑ ʌ j}
{b p m m̩ f v l syll. l n n̩ k ŋ g h d t s z w ɹ ð ɵ ʃ t͡s ʒ d͡ʒ}

Random viseme sets:

28 visemes: {h}{p t ɛ ʊ}{j}{k g ə˞}{syll. l}{u ju}{n n̩ w i ɪ}{ŋ}{l}{aʊ oʊ}
{ɔ ə}{b ʒ m ɹ ai ɔi}{v æ}{d͡ʒ ð}{z}{f ɵ m̩}{d t͡s ʃ ʌ}{ɑ}{s}{ei}

19 visemes: {h t͡s}{k}{ŋ ʃ}{ə˞}{m ɔi z}{ syll. l ɔ ð m̩}{g}{l ɪ}{oʊ ʒ d}{ə}
{n aʊ d͡ʒ j}{p ɹ s}{i ei}{ʊ u w ɵ}{ai ɑ}{ju b v æ}{t ɛ f}{n̩ ʌ}

12 visemes: {k ə˞ d n æ}{ŋ syll. l ʒ d͡ʒ i ʊ u w ɵ ai ju ɛ}{oʊ j ɹ v t}
{t͡s ɔi ə f}{h p}{ɔ aʊ}{ʃ}{m g}{ɪ l b ʌ}{ɑ n̩}{s}{z ei}{ðm̩}

10 visemes: {ɹ ð}{n d͡ʒ ɔi m}{i ɛ f h l s ei}{u j ə ɑ n̩}{ə˞ d w ʃ}{k oʊ ɪ b ʌ z}
{ai ju v}{ŋ syll. l p ɔ}{ʒ ʊ ɵ t͡s m̩}{æ aʊ g t}

2 visemes: {ʊ ɪ i ei æ ɔ ai ə b p m̩ n̩ k ŋ d s w ɹ ʒ}
{u ju ə˞ oʊ aʊ ɛ ɔi ɑ ʌ j m̩ f v l syll. l n g h t z ð ɵ ʃ t͡s ʒ}

Appendix B

FULL	segments	F1	F2	F3	F4	F5	M1	M2	M3	M4	M5
Easy	phoneme	0.259	0.203	0.247	0.281	0.266	0.126	0.257	0.228	0.231	0.221
	v28	0.314	0.240	0.301	0.337	0.316	0.144	0.296	0.264	0.278	0.261
	v19	0.385	0.303	0.361	0.396	0.377	0.178	0.352	0.309	0.340	0.315
	v12	0.465	0.384	0.448	0.494	0.467	0.240	0.430	0.387	0.422	0.402
	v10	0.567	0.484	0.556	0.559	0.531	0.308	0.480	0.466	0.512	0.483
	v2	0.726	0.639	0.727	0.740	0.681	0.564	0.674	0.629	0.699	0.690
Hard	phoneme	0.121	0.070	0.090	0.110	0.135	0.052	0.103	0.074	0.096	0.074
	v28	0.176	0.123	0.146	0.172	0.202	0.079	0.160	0.120	0.145	0.111
	v19	0.241	0.172	0.197	0.245	0.260	0.115	0.223	0.171	0.193	0.168
	v12	0.345	0.311	0.345	0.382	0.394	0.210	0.373	0.308	0.314	0.273
	v10	0.508	0.427	0.474	0.503	0.506	0.300	0.466	0.438	0.449	0.413
	v2	0.676	0.618	0.674	0.693	0.654	0.590	0.649	0.626	0.665	0.653

Table B.1 Average FULL score accuracy for Easy and Hard words for each of the talkers at each of the viseme criterion levels. The "segments" column denotes the number of segments used in scoring.

ORD	segments	F1	F2	F3	F4	F5	M1	M2	M3	M4	M5
Easy	phoneme	0.479	0.436	0.462	0.499	0.480	0.305	0.476	0.448	0.433	0.422
	v28	0.591	0.536	0.574	0.598	0.588	0.413	0.572	0.540	0.544	0.525
	v19	0.663	0.612	0.643	0.664	0.658	0.478	0.635	0.604	0.623	0.600
	v12	0.753	0.726	0.746	0.760	0.761	0.588	0.742	0.701	0.715	0.694
	v10	0.808	0.783	0.802	0.790	0.797	0.649	0.771	0.753	0.774	0.749
	v2	0.939	0.909	0.935	0.936	0.913	0.878	0.915	0.898	0.924	0.922
Hard	phoneme	0.382	0.316	0.330	0.373	0.395	0.277	0.368	0.330	0.353	0.331
	v28	0.516	0.474	0.474	0.507	0.527	0.405	0.496	0.457	0.476	0.450
	v19	0.591	0.548	0.548	0.583	0.596	0.469	0.566	0.530	0.553	0.536
	v12	0.713	0.703	0.702	0.718	0.728	0.606	0.728	0.681	0.691	0.661
	v10	0.801	0.763	0.770	0.771	0.788	0.671	0.771	0.758	0.759	0.734
	v2	0.933	0.920	0.918	0.929	0.907	0.892	0.912	0.905	0.917	0.916

Table B.2 Average ORD score accuracy for Easy and Hard words for each of the talkers at each of the viseme criterion levels. The “segments” column denotes the number of segments used in scoring.

SGM	segments	F1	F2	F3	F4	F5	M1	M2	M3	M4	M5
Easy	phoneme	0.543	0.502	0.521	0.553	0.539	0.363	0.535	0.506	0.496	0.471
	v28	0.661	0.612	0.643	0.665	0.650	0.485	0.640	0.608	0.617	0.593
	v19	0.739	0.692	0.717	0.742	0.725	0.554	0.704	0.683	0.704	0.681
	v12	0.830	0.803	0.820	0.832	0.824	0.673	0.802	0.780	0.805	0.776
	v10	0.884	0.856	0.880	0.872	0.863	0.739	0.839	0.832	0.867	0.838
	v2	0.977	0.951	0.977	0.965	0.942	0.912	0.936	0.929	0.963	0.952
hard	phoneme	0.465	0.405	0.413	0.453	0.469	0.346	0.446	0.403	0.424	0.402
	v28	0.595	0.559	0.561	0.589	0.594	0.471	0.569	0.533	0.553	0.519
	v19	0.675	0.642	0.640	0.674	0.672	0.548	0.650	0.619	0.641	0.620
	v12	0.802	0.784	0.786	0.795	0.798	0.690	0.791	0.761	0.769	0.748
	v10	0.894	0.853	0.866	0.859	0.856	0.769	0.840	0.837	0.845	0.830
	v2	0.991	0.986	0.981	0.969	0.939	0.928	0.942	0.949	0.955	0.968

Table B.3 Average SGM score accuracy for Easy and Hard words for each of the talkers at each of the viseme criterion levels. The “segments” column denotes the number of segments used in scoring.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Use of Gap Duration Identification in Consonant Perception by Cochlear Implant Users[1]

Adam R. Kaiser,[2] Mario A. Svirsky,[2] and Ted A. Meyer

DeVault Otologic Research Laboratory
Department of Otolaryngology-Head & Neck Surgery
Indiana University School of Medicine
Indianapolis, IN 46202

[1] This work was supported by NIH/NIDCD Training Grant DC-00012 and by grants from the Deafness Research Foundation, National Organization for Hearing Research, American Academy of Otolaryngology – Head and Neck Surgery Foundation. Earlier versions of this work were presented at the 1999 ARO MidWinter Meeting. We are grateful to Cara Lento and Julie Wirth for their assistance in data collection and processing.
[2] Also, Speech Research Laboratory, Department of Psychology, Bloomington, IN 47405.

Use of Gap Duration Identification in Consonant Perception by Cochlear Implant Users

Abstract. Cochlear implant (CI) users show substantial individual differences in their ability to understand speech. Even the most successful CI users do not understand speech as well as normal hearing listeners. One explanation for these individual differences and limited speech perception abilities lies in their difficulty in discriminating or identifying spectral cues (e.g., formant frequencies; Tong & Clark, 1985). In contrast to their discrimination of spectral cues, CI users are quite proficient in perceiving temporal cues. It has been demonstrated that postlingual adult CI users can detect the presence or absence of a gap in an acoustic stimulus as well as listeners with normal hearing (Shannon, 1989). Although detecting silent gaps within speech sounds is known to be important in consonant identification, the ability to discriminate between gaps of different duration may also be important. For example, the gap in "acha" is longer than the gap in "aba" and may be a cue for differentiating these phonemes. The present study examined the ability of listeners who use CIs and listeners with normal hearing to identify gaps of varied duration. After a short period of training, nine postlingual adult CI users and five adult listeners with normal hearing were presented with one of seven one-second synthetic vowel-like stimuli generated with Klatt 88 software (Klatt & Klatt, 1990). One stimulus was continuous, while the other six contained an intervening gap ranging in length from 15 to 90 ms. This range was chosen to encompass the range of silent gaps found in English consonants. Each listener was asked to identify the stimulus that was presented. This task was repeated until performance reached plateau. A *d'* was calculated for each pair of adjacent stimuli from the identification scores. All listeners, normal hearing and CI users, were able to perfectly discriminate the continuous stimulus from the stimulus with the 15 ms gap. However, the ability to differentiate among stimuli with gaps ranging from 15 to 90 ms varied between these two groups. The cumulative *d'* for the six gapped stimuli ranged from 6.0 to 6.6 for the normal hearing listeners and from 0.1 to 6.8 for the cochlear implant users. These cumulative *d'* scores were significantly correlated with consonant identification scores on the Iowa Consonant Test for the CI users. In addition, analyses of consonant confusion matrices suggested that CI users do not make optimal use of the temporal information they receive through their implant.

Introduction

Cochlear implants (CIs) are electronic devices that have enabled individuals with severe to profound hearing losses to regain some hearing. Nearly all of those who receive cochlear implants regain the sensation of sound. Some patients receive enough perceptual benefit that they can even communicate using a standard telephone (Dorman et al., 1991). This is a difficult task because there are no visual cues and the signal itself is not optimal. There is, however, substantial variability in speech perception performance among users of cochlear implants (Staller et al., 1997). Unfortunately, there is much we do not know about the exact mechanisms used by listeners with cochlear implants to understand speech. By studying psychophysical mechanisms, researchers have sought to more precisely determine the cues that patients with cochlear implants use for speech perception. A better understanding of these cues may eventually guide the development of better speech processing algorithms, and will also lend insight into the peripheral and central processes involved in speech perception by users of cochlear implants. It is generally accepted that one of the reasons why CI users have more difficulty understanding speech sounds than listeners with normal hearing is that the former are limited in formant perception ability (for a comparison see: McDermott & McKay, 1994; Kewley-Port & Watson, 1994). It is also generally accepted that, in contrast to their discrimination of spectral cues, CI users are quite proficient at

perceiving temporal cues. For example, it has been demonstrated that users of cochlear implants can detect the presence or absence of a gap in an acoustic stimulus as well as listeners with normal hearing (Shannon, 1989).

Gap detection, the detection of a silent period within an acoustic or electric stimulus, in contrast to gap identification, has frequently been used to evaluate the temporal ability of users of cochlear implants. Using gap detection, several researchers have studied the ability of postlingualy deafened adults with cochlear implants to detect gaps. They have found that the minimum gap detection threshold for these listeners (Shannon, 1989; Preece & Tyler, 1989; Moore & Glasberg, 1988) is similar to that of normal hearing listeners (Fitzgibbons & Gordon-Salant, 1987; Fitzgibbons 1984, 1983; Fitzgibbons & Wightman, 1982; Florentine and Buus, 1984) , ranging from 2 to 15 ms when sensation level and stimulus characteristics are taken into account. In general, studies of this type have seldom found a strong correlation between gap detection abilities and measures of speech perception. Some have reported a statistically significant correlation between psychophysical and speech perception measures (Hochmair-Desoyer et al., 1985) whereas others have argued against using a gap detection threshold measurement as a predictor of speech recognition (Shannon, 1989). This argument is not surprising because the speech perception measures of subjects with normal hearing and of those with cochlear implants are widely divergent, particularly among users of cochlear implants, yet their gap detection abilities are nearly the same. If gap detection ability is fundamental to speech perception, this would not be expected. It should be noted, however, that the similarity of thresholds for postlingual adult CI users and normal hearing adults may be specific to postlingually deafened CI users: a recent study of prelingually deafened users of cochlear implants has shown substantial variability in gap detection thresholds for congenitally deaf participants but did not find a significant correlation between detection thresholds and their scores on several speech perception measures (Busby & Clark, 1999).

The ability of listeners to detect the presence of acoustic gaps is potentially important in speech perception. For example, some intervocalic consonants, such as in *acha* and *aba*, contain acoustic gaps, while others such as *aza* and *ama* do not. Thus, detection of an acoustic gap may be used to distinguish between sounds. Additionally, the average length of the gap may be a useful cue for phoneme identification. For example, the gap in *acha* is, on average, longer than the gap in *aba* and thus could be a potential cue for identification of these phonemes. Temporal cues such as gap duration may be particularly important for CI users because their perception of spectral cues is significantly worse than that of normal hearing listeners.

In the present study listeners had to identify stimuli containing gaps of lengths ranging from 0 ms (i.e., no gap) to 90 ms. Using this absolute identification task, we evaluated the ability of CI users and normal hearing listeners to identify silent gaps spanning the range of gaps that occur in the English language. One motivation for obtaining these data was to calculate just noticeable differences (JNDs) for gap duration based on *d'*. These measurements are important input to our multidimensional phoneme identification (MPI) model, a quantitative framework that has been proposed to explain the mechanisms employed by CI users to understand speech sounds (Svirsky, 1991; Svirsky & Meyer, 1998; Svirsky, in press). In addition, we aimed to assess the possible relation between performance in this temporal processing task and speech perception by CI users.

Method

Subjects

Fourteen adult listeners were tested in this study. Nine of the listeners were postlingually deafened adult users of cochlear implants who were recruited from the clinical population at Indiana University (Table 1). All of these subjects had profound bilateral sensorineural hearing losses and greater

than one year of experience with their device. Each subject was reimbursed for travel to and from testing sessions and for the time of participation. Five of the subjects used the Nucleus 22 device, one used the Nucleus 24 device, and four others used the Clarion device. The comparison group consisted of five adults with normal hearing.

Table 1

Demographics of subjects with cochlear implants (S) and normal hearing (NH)

Subject	Age (Years)	Age at Onset Of Deafness	Age at Implantation	Implant Use (Years)	Implant Type
S1	67	*	65	2	Nucleus 22
S2	41	*	38	2	Clarion 1.0
S3	62	56	57	5	Nucleus 22
S4	69	55	66	2	Nucleus 22
S5	73	27	71	2	Nucleus 22
S6	58	*	52	5	Clarion 1.0
S7	41	32	34	7	Nucleus 22
S8	66	43	61	5	Clarion 1.0
S9	45	42	43	1.7	Clarion S
S10	37	34	36	1.1	Nucleus 24
S11	35	29	31	3	Clarion 1.2
NH1	35	-	-	-	-
NH2	30	-	-	-	-
NH3	25	-	-	-	-
NH4	22	-	-	-	-
NH5	22	-	-	-	-
NH6	32	-	-	-	-

* S1, Progressive; S2, Progressive since childhood; S6, Progressive since childhood

Stimuli and Equipment

Seven stimuli were created using the Klatt 88 speech synthesizer software (Klatt & Klatt, 1990). The first stimulus in the continuum was a synthetic three-formant steady state vowel with a duration of 1 second. Formant frequencies were 500, 1500, and 2500 Hz. Onset and offset of the vowel envelope occurred over a 10 ms period, and this transition was linear (in dB). The other six stimuli were similar to the first one, with the exception that they contained an intervening silent gap of length: 15, 30, 45, 60, 75, or 90 ms. The transitions in volume from full volume to silence (the silent gap) and from silence back up to full volume were made over a period of 10 ms., and they were linear in dB. The gap length was specified as the interval between the midpoint of the upward and downward slopes of the envelope represented in dB. The total amount of energy in each of the seven stimuli was identical, because the stimuli with longer gaps had correspondingly longer duration (see Fig. 1). The stimuli were digitally stored using a sampling rate of 11025 Hz at 16 bits of resolution. They were presented from an Intel® based PC equipped with a SoundBlaster compatible sound card. Stimuli were presented at a level of at least 70 dB C weighted SPL over an Acoustic Research loudspeaker. Custom software was used to present stimuli and record responses.

Figure 1: Graphical representation of psychophysical stimulus waveforms arranged by stimulus number.

Procedure

Subjects were tested using a seven-alternative absolute identification task. Each of the seven stimuli was randomly presented ten times during each block of testing (for a total of 70 presentations per block). Prior to each block of testing, the subjects were familiarized with each stimulus and its corresponding number. Subjects were allowed to listen to the stimuli at will prior to running each testing block. During the testing proper, subjects verbally responded with the corresponding stimulus number following stimulus presentation. After each response, feedback was provided on the computer monitor before moving on to the presentation of the next stimulus. Normal hearing subjects were tested until they reached asymptotic performance, determined by failure to improve their scores. The number of testing blocks for these subjects was from 6 to 10. We determined that the average of the best two scores during the first eight blocks was essentially the same as asymptotic performance for all the normal hearing listeners. Therefore, we decided to administer eight testing blocks to the CI users, which can be accomplished in one experimental session lasting three hours. This was done to minimize testing time for CI users, allowing them to participate in several other studies. The results presented below represent the best two of the first eight blocks, both for CI users and normal hearing listeners.

CI users were asked to return at a later date to complete a speech perception battery. Tests in this battery included CNC word lists, a 50 item monosyllabic open-set word identification task (Peterson & Lehiste, 1962), and 16 consonants from the Iowa Consonant Identification Task (female speaker) (Tyler et al., 1987). Both of these tests used natural speech and were presented in an auditory only condition. The consonant identification task is a closed-set 16 alternative task that uses 16 consonants in an /a/consonant/a/ format. For example, the consonant *m* would be presented as *ama*. All subjects performed at least 5 repetitions of the consonant identification task for a total of fifteen presentations of each consonant except for S2 and S7 who performed 4 and 2 repetitions respectively. Most subjects were administered three CNC word lists.

Analysis

A discrimination index or d' was calculated for each pair of adjacent stimuli (1 vs. 2, 2 vs. 3, . . .) for each testing block, using equation 1 (Levit, 1972). d' is a parameter that indicates the discriminability between two normal distributions (or, equivalently, between two stimuli associated with percepts that follow a normal distribution). For example, a d' of 0 indicates total lack of discriminability: an optimal

observer trying to identify one of two stimuli that had a *d'* of 0 would give correct responses 50% of the time, i.e., random performance. A *d'* of 1 would result in 69% correct responses in the same two-stimulus task, and a *d'* of 3 would result in 93% correct responses. Given that a *d'* of 3 indicates near perfect discrimination, it is customary to assign a value of 3 to any calculated *d'* that is greater than that.

$$d_n' = \frac{\bar{x}_{n+1} - \bar{x}_n}{\frac{s_{n+1} + s_n}{2}}$$

Equation 1: *d'* Calculation: where *d'* is the discriminability index calculated for the stimulus pair n. $\bar{x}_{n+1}, s_{n+1}, \bar{x}_n, \sigma_n$ are the means and standard deviations of the responses to stimulus n and stimulus n+1 respectively.

This index was calculated for each adjacent set of stimuli, 1vs 2, 2 vs. 3, 3 vs. 4 etc. The results from these individual comparisons were then summed to arrive at a cumulative *d'* which is a more global measure of the subjects' discrimination abilities across all of the gap lengths. The best two cumulative *d'* scores were averaged to arrive at the final score for this task.

The performance on the CNC wordlists was calculated as a percent correct. The scores for the Iowa Consonant Identification task, however, underwent a more detailed analysis. First, a percent correct score was calculated. In addition, since the ability of the subjects to perceive acoustic gaps in the speech tokens was of particular interest, the responses to consonants that either contained or did not contain gaps were also examined. Specifically, information transfer analysis as described by Miller and Nicely (1955) was used to calculate the percent information received by the listener in the task of identifying gapped vs. non-gapped stimuli. For example, a listener who never confuses a gapped stimulus with a non-gapped one is said to receive 100% of the "gap" information.

The gaps in all the consonant tokens in the Iowa 16-consonant tests were measured using a special purpose program called Scilab (Bögli et al., 1995). This program captures the stimulation parameters delivered by the cochlear implant speech processor in response to a given incoming sound. All three repetitions of the 16 consonants in the test were played at 70 dB SPL (measured at the level of the speech processor's microphone) and the corresponding stimulation patters coming out of the speech processor were recorded to disk. Silent gaps during the consonants were measured using the display capabilities of the Scilab software. For the purpose of calculating gap length we disregarded stimulation due to voicing during vocal tract closure, which showed up as stimulation pulses delivered to the lowest frequency channel.

Results

Table 2 shows cumulative *d'* scores for all listeners as well as speech perception scores for the CI users. Results in the speech perception tests showed the typically wide performance range found in CI users. In contrast, virtually all normal hearing listeners would be expected to score close to 100% in all these speech perception tests under the same conditions. Cumulative *d'* scores for normal hearing listeners fell within a relatively small range (7.9 to 11.6). However, scores for CI users again fell in a very wide range, 3.1 to 9.8. In general, subjects with normal hearing had better cumulative *d'* scores. However, this was not always the case. S9, S10 and S11 all performed within the range of the group with normal hearing.

Table 2

Performance in the gap duration identification task (all subjects) and in the speech perception tasks (CI users).

Subject Number	Cumulative *d'*	Information Transfer For Gap (%)	Iowa Consonants (% Correct)	CNC Wordlists (% Correct)
S1	3.1	0	6	1
S2	3.3	7	28	18
S3	3.4	6	20	11
S4	3.8	49	63	32
S5	4.0	16	32	25
S6	4.9	4	15	0
S7	5.4	65	54	30
S8	6.3	59	68	52
S9	8.0	35	56	58
S10	8.7	53	50	46
S11	9.8	70	57	68
NH1	7.9	-	-	-
NH2	8.7	-	-	-
NH3	9.7	-	-	-
NH4	10.0	-	-	-
NH5	10.3	-	-	-
NH6	11.9	-	-	-

Note: CI users are listed first (S), and normal hearing listeners follow (NH). Within each group, subjects are listed in order of increasing ability to identify silent gap duration.

Figure 2 graphically illustrates the both the cumulative *d'* scores (i.e., the first column of numbers in Table 2) as the total bar height as well as the fraction due to the comparison between the first two stimuli (in black). The total cumulative *d'* is broken down into two parts: black bars represent a *d'* comparing responses to stimuli 1 and 2 (i.e., the non-gapped stimulus and the stimulus with a 15 ms gap), and white bars represent cumulative *d'* for stimuli 2-7 (i.e., stimuli with gaps ranging from 15 ms to 90 ms). It is very clear from the figure that all listeners, CI users and normal listeners alike, had perfect or near perfect ability to discriminate the non-gapped stimulus from the stimulus with a 15 ms gap, because the corresponding *d'* values are all equal to 3. In contrast, the ability to identify gaps of different lengths (cumulative *d'* for stimuli 2-7) from near zero for S1 to a substantial 6.8 for S11.

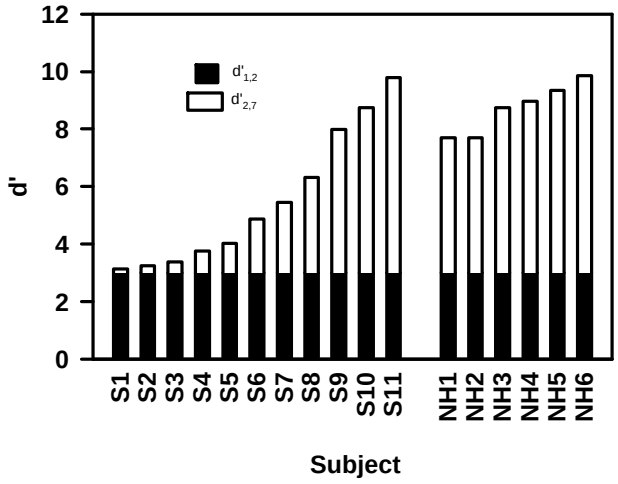


Figure 2: *d'* by subject. The black part of the bar represents the portion of the cumulative *d'* that is due to the subject's ability to discriminate between stimulus 1 and stimulus 2. S=subjects with cochlear implants, NH=subjects with normal hearing

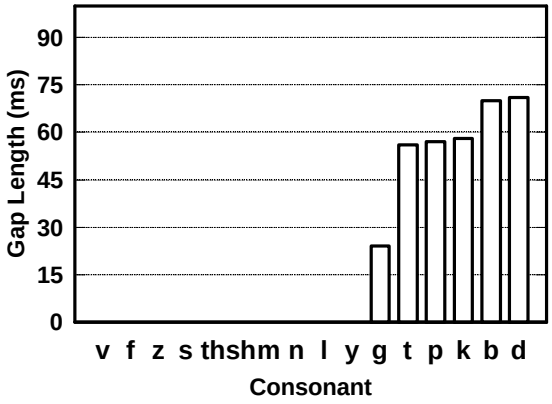


Figure 3: Consonant gap length as measured from the stimulus generated by a Spectra 22 cochlear implant speech processor and visually analyzed using Scilab software.

Figure 3 shows the measured silent gaps for each one of the 16 consonants used in this test. Each bar represents the average of three tokens. Based on these measurements, one would expect that all of the CI subjects would be able to perceive the difference between a consonant with a gap and a consonant without a gap. This is because all of the gapped consonants contain gaps that are at least 15ms in length, and all of the subjects showed that they were able to perfectly distinguish continuous sounds from those containing a 15 ms gap. However, information transmission scores for the gap-no gap feature (shown in the y-axis if the left panel in Fig. 4) were less than 100% for all CI users.

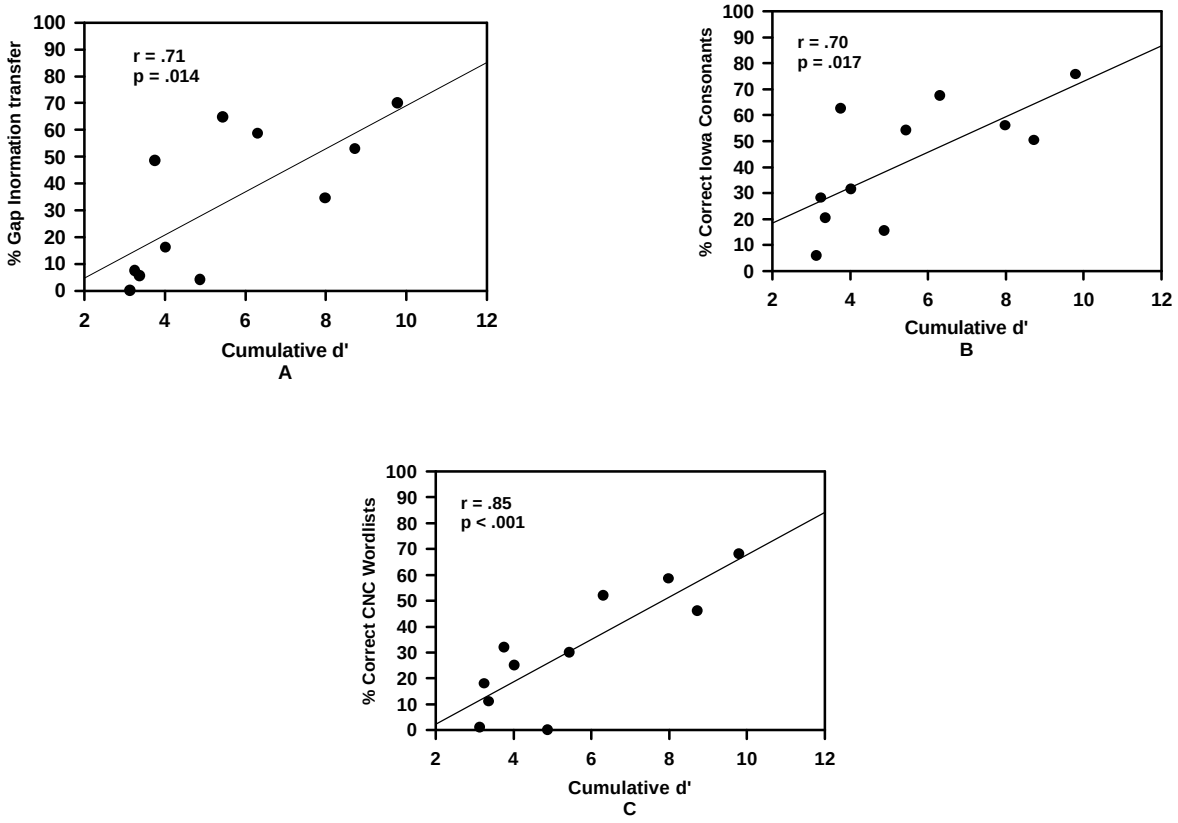


Figure 4: Correlation between cumulative *d'* and (A) gap information transfer, (B) performance on the Iowa Consonant Identification (female) task, (C) and CNC wordlist scores. Information transfer analysis was performed using gapped and gapless consonants as informational categories.

Finally, figures 4A and 4B show scatter-plots of cumulative *d'* scores against percent information transmitted for the gap-no gap contrast, and against the percentage of correctly identified consonants respectively. Figure 4C is a scatter-plot of cumulative *d'* against CNC wordlist percent correct. All speech perception scores were significantly correlated with cumulative *d'*.

Discussion

Speech perception performance by CI users is highly variable from one patient to the next and is also, on average, lower than that of normal hearing listeners. The question remains whether there any psychophysical capabilities that might explain these two facts. The consensus in the cochlear implant field is that CI users suffer from poor frequency discrimination, although their temporal processing skills are comparable to those of normal hearing listeners (Shannon, 1989). The latter conclusion is supported by results from gap detection threshold experiments, and is consistent with our finding that all of our subjects (normal hearing and CI users) were clearly able to label and differentiate a continuous stimulus from one with a 15 ms gap. However, the ability of CI users to identify silent gaps of different lengths was extremely variable, ranging from almost nil to normal. In addition, this ability (as measured by cumulative *d'* scores for stimuli 2-7) was strongly correlated with speech perception scores as shown in 4B and 4C. These results suggest that CI users may use silent gap duration as a cue to consonant identity,

and that their differing abilities to identify such gaps may help explain their individual differences in speech perception. This hypothesis receives additional, indirect support from the modeling studies of Svirsky et al. (1999). The MPI model developed by Svirsky et al. can fit consonant confusion data much better when they include a temporal dimension in the model. The temporal dimension used by Svirsky et al. in the MPI model is the duration of a silent gap. In any case, it seems clear that the temporal processing abilities of poorer CI users lag substantially behind those of normal hearing listeners or those of more successful CI users. What is the physiological reason behind this variability in gap duration identification? In our view, the auditory periphery is an unlikely locus for individual differences in temporal processing, because the electrically stimulated auditory nerve is quite capable of encoding the gross temporal differences that were present in our stimuli. Instead, it may be that the differences in cumulative *d'* that we observed have their origin in more central differences in auditory processing and categorization of sensory input into stable perceptual units.

Another observation is that the CI users' identification of gapped and non-gapped consonants was much lower than would be expected based on their psychophysical performance. We say this because their discrimination of continuous sounds from sounds with a 15 ms gap was virtually perfect in the psychophysical task (where they had to perform absolute identification of synthetic stimuli), but their ability to identify consonants with and without gaps was far from perfect (as evidenced by their information transfer scores for the gap-no gap feature). One possible account for this result is that listeners may have been paying more attention to acoustic cues other than gap duration, or at least weighting those other cues more heavily. In other words, they may have known that a given stimulus contained a silent gap, but other acoustic cues led them to identify the stimulus as a non-gapped consonant. Alternatively, it may be that the cumulative *d'* estimates obtained in this study can only be achieved by most subjects under relatively ideal conditions, that is, with carefully synthesized acoustic stimuli that are identical to each other except for the silent gap. When subjects have to process gap information in conjunction with other acoustic cues (spectral and amplitude cues, for example) their processing of temporal cues may suffer to some extent.

Perception of speech sounds by human listeners in general and by CI users in particular is a very complex phenomenon. In the case of CI users, we hope to develop a comprehensive quantitative model of speech perception that is based, in part, on the individual listener's discrimination abilities (Svirsky & Meyer, 1998). The present study is a first step in that direction. It is our hope that the existence of a theoretical framework for speech perception by CI users will help guide the search for improved signal processing and aural rehabilitation strategies for this population.

References

Bögli, H., Dillier, N., Lai, W. K., Rohner, M., & Zillus, B. A. (1995). Swiss Cochlear Implant Laboratory (Version 1.4) [Computer software]. Zürich, Switzerland: Authors.

Busby, P. A. & Clark, G. M. (1998). Gap detection by early-deafened cochlear-implant subjects. *Journal of the Acoustical Society of America,* 105, 1841-1852.

Dorman, M., Dove, H. J., Parkin, J., Zacharchuk, S., & Dankowski, K. (1991). Telephone use by patients fitted with the Ineraid cochlear implant. *Ear and Hearing,* 12, 368-369.

Fitzgibbons, P. J., (1983). Temporal gap resolution in masked normal ears as a function of masker level. *Journal of the Acoustical Society of America,* 76, 67-70.

Fitzgibbons, P. J., (1984). Temporal gap detection in noise as a function of frequency bandwidth, and level *Journal of the Acoustical Society of America*, 74, 67-72.

Fitzgibbons, P. J., & Gordon-Salant, S. (1987). Temporal gap resolution in listeners with high-frequency sensorineural hearing loss. *Journal of the Acoustical Society of America*, 81, 133-137.

Fitzgibbons, P. J., & Wightman F. L., (1982). Gap detection in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 72, 761-765.

Florentine, M., & Buus, S. (1984). Temporal gap detection in sensorineural and simulated hearing impairments. *Journal of Speech and Hearing Research*. 27, 449-455.

Hochmair-Desoyer, I. J., Hochmair, E. S., & Stiglbrunner H. K. (1985). Psychoacoustic temporal processing and speech understanding in cochlear implant patients. In Schindler, R. A. & Merzenich, M. M. (Eds.), *Cochlear Implants* (pp. 291-303). New York: Raven Press.

Klatt, D. H., & Klatt L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America.* 87, 820-57.

Levit, H. (1972). Decision Theory, Signal-detection Theory, and Psychophysics. In David, E. E., Denes P. B. (Eds.), *Human Communication a Unified View*. (pp. 114-174). New York: McGraw-Hill.

Miller, G. A. & Nicely P. A. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America,* 27, 338-352.

Moore, C. B. J., & Glassberg, B. R. (1988). Gap detection with sinusoids and noise in normal, impaired, and electrically stimulated ears. *Journal of the Acoustical Society of America.* 83, 1093-1101

Peterson, G., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders*, 27, 62-70.

Preece, J. P., & Tyler, R. S. (1989). Temporal-gap detection by cochlear prosthesis users. *Journal of Speech and Hearing Research*, 32, 849-856.

Shannon, R. V. (1989). Detection of gaps in sinusoids and pulse trains by patients with cochlear implants. *Journal of the Acoustical Society of America,* 85, 2587-2592.

Staller, S., Menapace, C., Domico, E., Mills, D., Dowell, R. C., Geers, A., Pijl, S., Hasenstab, S., Justus, M., Bruelli, T., Borton, A.A., & Lemay, M. (1997). Speech perception abilities of adult and pediatric Nucleus implant recipients using spectral peak (SPEAK) coding strategy. *Otolaryngology-Head & Neck Surgery*. 117, 236-242.

Svirsky, M. A. & Meyer, T. A. (1998). A mathematical model of consonant perception in adult cochlear implant users with the SPEAK strategy. *Proceedings of the 16th International Congress on Acoustics* and *135th Meeting of the, Acoustical Society of America*, Vol. III: 1981-2.

Svirsky, M. A., Kaiser, A. R., Meyer, T. A., Shah, A., & Simmons, P. M. (1999, October). Psychophysical and cognitive limitations to speech perception by cochlear implant users. Paper presented at the 138th meeting of the *Acoustical Society of America*, Columbus, OH.

Svirsky, M. A. (1991). A mathematical model of vowel perception by users of pulsatile cochlear implants. Presented at the *22nd Annual Neural Prosthesis Workshop*, Bethesda, MD.

Svirsky, M. A. (In press). Mathematical modeling of vowel perception by cochlear implantees who use the "compressed-analog" strategy: Temporal and channel-amplitude cues. *Journal of the Acoustical Society of America.*

Tong, Y. C., & Clark, G. M., (1985). Absolute identification of electric pulse rates and electrode positions by cochlear implant patients. *Journal of the Acoustical Society of America.* 77, 1881-1888.

Tyler, R. S., Preece, J. P., & Lowder, M. W. (1987). The Iowa audiovisual speech perception laser video disc. *Laser Videodisc and Laboratory Report,* University of Iowa at Iowa City, Department of Otolaryngology – Head and Neck Surgery.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Neighborhood Density, the Tip-of-the-Tongue Phenomenon, and Aging[1]

Michael S. Vitevitch and Mitchell S. Sommers[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This research was supported in part by NIH-NIDCD training grant T32 DC 00012 to Indiana University and the Brookdale Foundation (Washington University). The authors would like to thank Emily McCutcheon for her assistance in data collection and analysis, and Luis R. Hernandez for his programming assistance.
[2] Department of Psychology, Washington University, Campus Box 1125, St. Louis, MO 63130.

Neighborhood Density, the Tip-of-the-Tongue Phenomenon, and Aging

Abstract. A tip-of-the-tongue (TOT) elicitation task was used with younger adults in Experiment 1 and with older adults in Experiment 2 to examine the influence of word frequency, neighborhood density, and neighborhood frequency on the retrieval of phonological word forms from the lexicon. The results of Experiment 1 replicated the results of Harley and Bown (1998): More TOT states were elicited for words with low frequency and sparse neighborhood density (i.e., words with few similar sounding words). However, the results from Experiment 2 showed that in addition to word frequency and neighborhood density, neighborhood frequency, the mean frequency of phonological neighbors, also influenced lexical retrieval for speech production in older adults. Specifically, neighborhood-frequency interacted with word frequency and with neighborhood density. More TOT states were elicited for words with low neighborhood frequency and low word frequency. In addition, more TOT states were elicited for words with low neighborhood-frequency and sparse neighborhoods. These results demonstrate that the number of similar sounding words affects lexical retrieval in production as well as perception. Furthermore, the influence of these lexical characteristics on the process of retrieving word-forms during speech production changes with age. The results are discussed within the context of the Node Structure Theory (MacKay, 1987), a model of cognitive processing.

Introduction

Speech production is a rapid and efficient process.[3] However, there are instances in which the fluent retrieval of a lexical item fails to occur. One example of failed retrieval occurs in tip-of-the-tongue (TOT) states. Tip-of-the-tongue states occur when only partial information associated with a word can be retrieved: information regarding the meaning or syntactic class of the word may be accessible, but not the complete phonological form of the word. The ability to access partial information often results in one having a "feeling of knowing" the word, despite being unable to retrieve all the information associated with the word.

Factors that Affect Lexicalization and TOTs

The process of retrieving a word form from lexical memory during speech production is called lexicalization. Several factors affect the speed and accuracy of lexicalization. The factors that affect the speed and accuracy of normal, unimpaired, lexicalization also affect instances of incomplete lexicalization, or TOT states. One factor that affects the speed and accuracy of normal lexicalization, as well as TOT states, is word frequency. Using a picture-naming task, Oldfield and Wingfield (1965) demonstrated that pictures of high-frequency words were named more quickly than were pictures of low-frequency words. Stemberger (1985; Stemberger & MacWhinney, 1986; see also Dell, 1990) found that phonological speech errors, such as spoonerisms (switching the initial phoneme of two nearby words, such as "darn bore" instead of "barn door"), perseverations (carrying-over a phoneme from a word produced earlier, such as "Ladies and lentlemen..." instead of "Ladies and gentlemen..."), and anticipations (producing a phoneme

[3] This is about the only point that researchers agree on. Debates continue about the processes involved in speech production. For example, Garrett (1976) and Levelt (1989) take a modular approach to processing, whereas Dell (1986) and Harley (1984; 1993) argue for an interactive approach. There are also debates about where certain types of information are represented in the speech production system. For example, Roelofs, Meyer, and Levelt (1998) argue that syntactic information is accessed at the lemma level, whereas Caramazza and Miozzo (1998) claim that syntactic information is accessed at the lexeme level.

from an upcoming word, such as “pig pig” instead of “big pig”), occurred more often for low-frequency words than for high-frequency words. Finally, more TOT states occur and are experimentally induced in low frequency words than in high frequency words (Brown & McNeill, 1966; Burke, MacKay, Worthley, & Wade, 1991; Harley & Bown, 1998; cf. Yaniv & Meyer, 1987). Taken together, these findings suggest that the well-documented perceptual disadvantage of low-frequency words (i.e., poorer identification and slower response times) is paralleled by a similar disadvantage in production, as evidenced by greater difficulty with lexicalization.

Another factor that affects the speed and accuracy of both perception and lexicalization is neighborhood density. Neighborhood density refers to the number of words that are phonologically similar to a given target word (Luce & Pisoni, 1998). A rough measure of phonological similarity can be obtained by determining the number of new words that are created by the addition, deletion, or substitution of a phoneme in a target word.[4] For example, the word “cat” has as neighbors the words “scat,” “at,” “hat,” “cut,” and “cap,” as well as other words. Words with many similar sounding words are said to have dense neighborhoods, whereas words with few similar sounding words are said to have sparse neighborhoods. Previous research has shown that words with sparse neighborhoods are recognized more quickly and more accurately than words with dense neighborhoods (Luce & Pisoni, 1998).

Previous investigations (Goldinger & Summers, 1989; Harley & Bown, 1998; Vitevitch, 1997a; 1997b) have also demonstrated that neighborhood density can affect speech production. For example, Goldinger and Summers (1989) found that neighborhood density influenced the voice onset time (VOT) for spoken words. VOT refers to the point in time at which vocal fold vibration starts, following the release of a closure (Crystal, 1992). Participants repeated word pairs that differed in the voicing of the initial consonants within the pair (e.g., dutch-touch) and that varied in neighborhood density across pairs. An acoustic analysis showed that the differences in VOT between the first word and the second word of the pairs were larger for word pairs with dense neighborhoods than for word pairs with sparse neighborhoods. These differences decreased across sessions for word pairs with sparse neighborhoods, but increased across sessions for word pairs with dense neighborhoods. Furthermore, Goldinger and Summers found that the interword interval, or the time between the offset of the first word and the onset of the second word within each minimal pair, varied with neighborhood density. The interword interval was greater for dense neighborhood word pairs than for sparse neighborhood word pairs. These results demonstrate that neighborhood density influences certain aspects of timing in speech production.

The accuracy of lexical retrieval in speech production is also affected by neighborhood density. Vitevitch (1997a) used tongue twisters containing words that had either dense or sparse neighborhoods to elicit phonological speech errors from participants. He found that more errors occurred in tongue twisters that contained words with sparse neighborhoods than with dense neighborhoods. These results suggest that neighborhood density influences speech production in demonstrable ways (see also Vitevitch, 1997b; Wright, 1997; cf. Jescheniak & Levelt; 1994). Specifically, neighborhood density appears to produce “supportive” or facilitative effects among words. That is, words with many similar sounding words (a dense neighborhood) are produced more accurately than words with few similar sounding words (a sparse neighborhood). These findings contrast with the competitive effect of neighborhood density typically observed in spoken word recognition: Words with dense neighborhoods are recognized more slowly and less accurately than words with sparse neighborhoods (Luce & Pisoni, 1998).

[4] An alternate method used to measure "similarity" is to use phoneme confusion matrices as in Luce and Pisoni (1998) in the calculation of Neighborhood Probability Rules (NPRs). Both methods have been successfully used to demonstrate effects of neighborhood density on spoken word recognition.

Vitevitch (1997a; see also MacKay & Burke, 1990) speculated that the difference in neighborhood density effects found in speech production and spoken word recognition were due to the differences in the flow of information during speech production and spoken word recognition. That is, in spoken word recognition, acoustic-phonetic input activates many similar sounding words in memory (e.g., Luce & Pisoni, 1998). This candidate set must be winnowed down to a single item that will then retrieve semantic and syntactic information related to that word from the lexicon. Thus, more time will be required to winnow down the candidate set if there are many competitors (Luce & Pisoni, 1998). In contrast, speech production begins with a single conceptual representation that proceeds to activate a single lexical item and a single phonological word form (Levelt, 1989). That single phonological word form then activates the many sub-lexical units that it contains, such as syllables, phonemes, features, etc. Thus, a word that has components shared by many other words (a word with a dense neighborhood) will be able to spread activation along pathways between those components that are well traversed. A word that has components shared by few other words (a word with a sparse neighborhood) will have difficulty spreading activation along the pathways between components that are less traveled.

Additional evidence of neighborhood density affecting speech production can be found in the work of Harley and Bown (1998). They recently reported that more TOT states were elicited for words from sparse neighborhoods than for words from dense neighborhoods, and also suggested that neighborhood density played a "supportive" role in speech production (Harley & Bown, 1998). Although Harley and Bown (1998) accounted for their results in the context of extant interactive models of lexicalization (i.e., Dell, 1986; Harley, 1993), their results are difficult to clearly interpret because of confounding variables in their stimulus set.

Specifically, in two experiments that manipulated word frequency and neighborhood density, Harley and Bown (1998) attempted to induce TOT states experimentally using words that varied in length from one syllable (e.g., "act") to five syllables (e.g., "chronological"). Word length was a variable that was not stringently controlled in their stimuli, and, unfortunately, proved to be a confounding variable. The results of their first experiment showed that more TOT states were reported for words that were low in frequency and that had few neighbors as defined by Coltheart-*N* (Coltheart, Davelaar, Jonasson, & Besner, 1977). Although the tip-of-the-tongue phenomenon is often described as an inability to retrieve a sound-based representation from the lexicon, Harley and Bown (1998) constructed their stimulus set using a metric of similarity based on orthographic similarity (Coltheart-*N*) instead of a metric based on phonological similarity. It should be noted, however, that when Harley and Bown analyzed the results from a reduced set of their stimuli based solely on phonological neighborhoods, their findings remained relatively unchanged.

However, when Harley and Bown performed a regression analysis on the data in Experiment 1, they found a significant effect of word length on TOT states: TOT states were more likely to occur with longer words than shorter words. Across the lexicon, short words tend to have denser neighborhoods than longer words (Bard & Shillcock, 1993; Pisoni, Nusbaum, Luce & Slowiaczek, 1985). Their results are further complicated by other relationships among word frequency, word length, and neighborhood density in the lexicon. For example, Zipf (1965) observed an inverse relationship between word length and word frequency in English: Short words are more common in English than long words. Also, Landauer and Streeter (1973) found a positive correlation between word frequency and neighborhood density: High frequency words tend to have denser phonological neighborhoods than low frequency words. Thus, it is unclear whether the results in Experiment 1 of Harley and Bown (1998) were due to neighborhood density or another related variable.

Harley and Bown (1998) attempted to control word length more precisely in their second experiment by using monosyllabic and disyllabic words (however, the trisyllabic word "audience" appears as a stimulus item in a low *N* condition) to examine the effects of word frequency and neighborhood density on TOT states. Although the word frequency and neighborhood density effects from Experiment 1 were replicated, a close examination of the stimuli in Experiment 2 reveals that word length was not entirely controlled. An analysis of the stimuli in appendix B of Harley and Bown (1998) shows that words with dense neighborhoods were still shorter than words with sparse neighborhoods. This is true when word length is measured in number of phonemes (dense words, mean = 3.17 phonemes; sparse words, mean = 5.07 phonemes; F (1,58) = 54.15, $p < .001$) and in number of syllables (dense words, mean = 1.06 syllables; sparse words, mean = 1.83 syllables; F (1,58) = 8.82, $p < .001$). Given the complex relationships among word length, word frequency, and neighborhood density, it is unclear how each of these individual factors affected TOTs in Harley and Bown (1998).[5]

Accounts of TOTs

Several hypotheses have been advanced to account for the occurrence of TOT states. One hypothesis states that similar sounding words interfere with— or "block"—the retrieval of the phonological word-form (Jones, 1989, Jones & Langford, 1987; Maylor, 1990; Woodworth, 1929). For example, Jones (1989) presented definitions to participants and primed them with a word that was semantically, phonologically, or both semantically and phonologically related to the target word. Jones (1989; see also Jones & Langford, 1987, and Maylor, 1990) found that more TOT states were elicited when a phonologically related prime was presented after hearing the definition of the target word. Jones (1989) interpreted these results as being consistent with the hypothesis that phonologically related words block the retrieval of the desired word-form.

An alternative explanation of TOTs claims that insufficient activation results in incomplete retrieval of the target word (Brown, 1991; Burke, MacKay, Worthley & Wade, 1991). According to this hypothesis, similar sounding words should act to aid rather than block the retrieval of word-forms. Evidence for this hypothesis comes from the work of Meyer and Bock (1992) and Perfect and Hanley (1992). Meyer and Bock (1992) and Perfect and Hanley (1992) showed that the targets used by Jones (1989) differed across conditions in the susceptibility to TOT states. When targets with equal susceptibility to TOT states were used across conditions, phonological primes did not interfere with the retrieval of the target word form; rather, phonological primes aided in the retrieval of the target word-form (Meyer & Bock, 1992; Perfect & Hanley, 1992).

The results of Harley and Bown (1998) also support the hypothesis that phonological similarity can serve to support lexicalization. They found more TOTs for words with sparse neighborhoods than for words with dense neighborhoods, suggesting that the more neighbors a word has, the more "support" it receives, and the more likely it will be correctly and completely retrieved. Harley and Bown (1998) accounted for their results by hypothesizing that the representation of an intended word was not fully activated because of insufficient amounts of supportive feedback between the lexeme level (which contains phonological information) and the lemma level (which contains semantic and/or syntactic information). They state that "...[l]emmas corresponding to phonological forms that have no or few close neighbours can receive little or no supporting activation from feedback between the phonological and lemma levels from

[5] As in Experiment 1, Harley and Bown (1998) performed a regression analysis, but failed to find a relationship between word length and number of TOT states. However, the restricted range of word length in Experiment 2 (mostly mono- and disyllabic words, with one trisyllabic word) compared to the broader range of word length in experiment 1 (words with one to five syllables) may account for the non-significant regression.

related forms." (Harley & Bown, 1998; pp. 163-164). Harley and Bown (1998; see also Harley & MacAndrew, 1992) further hypothesize that weak representations or random noise in the connections between the lemma and phonological representations may also contribute to TOT states.

Rather than being at the interface between the lemma and lexeme, as postulated by Harley and Bown (1998), the locus of TOT states may instead be at the interface between the lexeme (i.e., the phonological representation of the whole word) and sub-lexical representations. This hypothesis was postulated by Burke et al. (1991) within the context of the Node Structure Theory (NST), an interactive model of cognitive processing (see MacKay, 1987). Specifically, they state that "...TOTs result when phonological feature nodes receive insufficient priming to become activated." (Burke et al., 1991, pp. 547). Insufficient activation between "word" and "phoneme" representations as postulated in NST can also explain the results of Harley and Bown (1998).

Node Structure Theory and TOTs

NST consists of a network of processing units, or nodes, organized hierarchically into semantic, phonologic, and motoric levels. Nodes are localist representations and are connected symmetrically —both bottom-up and top-down connections. The same network of nodes is involved in the perception and production of language (MacKay, 1987).

Two processes operate in NST: priming and activation. Priming is the sub-threshold excitation of a node that prepares it for activation. Activation is an all-or-none state in which the node has crossed a certain threshold.

Priming has several characteristics. It spreads in parallel to all connected nodes higher and lower in the hierarchical structure of nodes. Nodes can sum the priming that they receive simultaneously from several other nodes, or that they receive temporally across a single connection. Finally, transmission of priming becomes less efficient when a node has been satiated after prolonged and repeated activation.

In NST, activation is different from priming. For example, activation does not "spread" as it does in other network theories (e.g., McClelland & Rumelhart, 1981). Instead, activation proceeds sequentially and hierarchically through the network in a top-down and left-to-right manner. Activation must occur (i.e., the threshold must be crossed) in order to consciously retrieve the information associated with a node.

In addition to the nodes representing information at various levels, there are sequence nodes that connect nodes that share the same syntactic function or sequential privilege of occurrence in words and sentences. (This collection of similar nodes is referred to as a domain.) When a sequence node is activated, it multiplies the priming of all the nodes in a domain. The consequence of this multiplication of priming is that the node that initially had the most priming in that domain reaches threshold first and becomes activated. This domain specific activation accounts for the regularity found in many types of substitution errors. For example, nouns often substitute for nouns rather than verbs in a sentence, initial consonants often substitute for initial consonants rather than vowels or final consonants, etc. (e.g., Dell, 1986; Stemberger, 1985; MacKay, 1979).

Burke et al. (1991) suggest that TOT states arise in NST due to a deficit in transmission of priming across certain connections that are crucial for producing a target word. In a TOT state, semantic nodes become activated giving access to semantic information associated with that word. However, priming

does not spread sufficiently among phonological nodes, resulting in some phonological information not being activated and retrieved.

Transmission deficits may be caused by three factors: frequency of use, recency of use, and aging. The frequent activation of a node results in an increase in the rate and amount of priming that is transmitted across the connections of that node. Connections between less frequently activated nodes weaken with time, making the transmission of priming less efficient. This factor accounts for the frequency effects commonly found in speech error corpora: word and phoneme substitutions occur more often among low frequency items than among high frequency items (e.g., Stemberger, 1985, Stemberger & MacWhinney, 1986). Furthermore, object naming is faster for high frequency items than low frequency items (e.g, Oldfield & Wingfield, 1965), and TOT states tend to occur more often for low frequency than high frequency words (e.g, Burke et al., 1991).

Over time, connection strength between nodes decreases. If the connections become too weak, transmission of priming becomes less efficient. The longer a node has been "inactive," the more decay occurs to the connections of that node. Evidence for this factor comes from a diary study by Burke et al. (1991) in which young and older participants recorded naturally occurring TOT states. Burke et al. (1991) found that a TOT state would more likely occur for a proper name the longer the duration since that acquaintance was contacted.

Finally, age weakens the strength of connections within the entire network (MacKay & Burke, 1990), reducing the rate and amount of priming being transmitted. This factor accounts for the general slowing of cognitive processes often associated with aging (e.g., Salthouse, 1985) and also the increased rate of TOT states seen for older compared with younger adults (Burke et al., 1991).

The neighborhood density effects observed by Harley and Bown (1998)—more TOTs for words with sparse neighborhoods—can also be accounted for within NST if one views neighborhood density as a frequency effect among the phonological constituents of a word. For example, a word with a dense neighborhood has many similar sounding neighbors, and, therefore, is comprised of segments that are common, or very frequent in the language. A word with a sparse neighborhood has fewer similar sounding neighbors, and, therefore is comprised of segments that are less frequent in the language. (See Vitevitch, Luce, Pisoni, & Auer (1999) for evidence of a positive correlation between neighborhood density and the frequency of segments and sequences of segments, also know as phonotactic probability.) Phonemes (represented by phonological nodes) that constitute words with dense neighborhoods receive more priming than phonological nodes that constitute words with sparse neighborhoods. The greater amount and rate of priming received by phonological nodes that constitute words in dense neighborhoods further strengthens those connections, whereas the connections between phonological nodes and words with sparse neighborhoods become weaker over time. Thus, the phonological information associated with a word with a dense neighborhood is more efficiently retrieved than the phonological information associated with a word with a sparse neighborhood. When the number of similar sounding words (i.e., neighborhood density) is viewed in terms of sub-lexical constituents, NST can also account for the neighborhood density effects observed by Harley and Bown (1998) without invoking additional feedback mechanisms between lexemes and lemmas.

NST, Neighborhoods, and Aging

In the current set of experiments, we examined the influence of neighborhood density on the number of TOT states elicited experimentally with stimuli that were controlled for word length. To

unambiguously demonstrate that neighborhood density, independent of word frequency and word length, affects TOT states, we used monosyllabic words with a consonant-vowel-consonant syllable structure that varied in word frequency, neighborhood density, and neighborhood frequency—a variable not manipulated by Harley and Bown (1998). (Neighborhood frequency is the mean frequency of the neighbors of a given target word). Also, given that the TOT phenomenon is the inability to retrieve phonological word forms from the lexicon, we used a metric of similarity based on phonological representations rather than orthographic representations as in Harley and Bown (1998).

A second goal of our investigations was to examine age differences in the effects of neighborhood density on lexicalization by eliciting TOT states from older adults in Experiment 2. Previous research (Sommers, 1996) reported that neighborhood density has greater effects on older than on younger adults. Therefore, Experiment 2 examined whether this age difference in the effects of neighborhood density on perception would have parallels in production. Work by Burke et al. (1991; see also Burke & Laver, 1990; MacKay & Burke, 1990; and Rastle & Burke, 1996) suggests that older adults have more problems than younger adults retrieving items from memory during language production. For example, older adults report more TOT states than younger adults. In Experiment 2, we examined how age modulates the effects of neighborhood density on the frequency of TOT states. In summary, we examined the influence of several lexical characteristics on speech production with a stringently controlled set of stimuli, and we also investigated how the influence of these characteristics may change over the life span by eliciting TOT states from young and older adults.

Experiment 1

The principle goal of Experiment 1 was to determine whether the effects of neighborhood density on TOT states reported by Harley and Bown (1999) could be replicated using a set of stimuli that were controlled for word length. Experiment 1 used a TOT elicitation task similar to that developed by Brown and McNeill (1966). In this task, participants are presented with a definition and must retrieve from memory the word that best matches the definition. Participants indicate whether they know the word (and produce it), don't know the word, or know the word but can't retrieve it (i.e., they are in a TOT state). The target words in this experiment were monosyllabic words that varied in word frequency, neighborhood density, and neighborhood frequency. Monosyllabic words were used to control for effects of word length on TOT susceptibility observed in Harley and Bown (1998).

Several predictions were made based on NST (Burke et al., 1991, MacKay, 1987) and the results of Harley and Bown (1998). Insufficient activation of phonological nodes would result in more TOT states for low frequency words than high frequency words. Based on the view that neighborhood density can be considered a frequency effect among the phonological constituents of a word, this same mechanism would also result in more TOT states for words with sparse neighborhoods than for words with dense neighborhoods. Finally, insufficient activation of phonological nodes would result in more TOT states for words with low-frequency neighbors than for words with high-frequency neighbors.

Method

Participants

Twenty-four native English-speaking adults were recruited from the Washington University community. None of the participants reported a history of a speech or hearing disorder and all received partial credit towards an Introductory Psychology class for their participation. The mean age of these

participants was 20.7 years (SD =1.9). Mean years of education for the participants was 14.1 years (SD = 1.4). Data from one participant was excluded from all analyses because of failure to comply with experimental instructions.

Materials

One hundred twenty monosyllabic words consisting of a consonant-vowel-consonant syllable pattern were used as targets in the TOT elicitation task. Eight conditions, each containing fifteen words, were formed by orthogonally combining two factors of word frequency (HIGH and LOW), neighborhood density (SPARSE and DENSE), and neighborhood frequency (HIGH and LOW). The familiarity ratings (based on a scale from 1 "Don't know the word" to 7 "Know the word"; Nusbaum, Pisoni, & Davis, 1984) did not differ across conditions (F (1,112) = 2.02, $p > .10$). The means and standard deviations for familiarity, word frequency, neighborhood density, and neighborhood frequency for the words in each condition are listed in Table 1.

High-frequency words (mean = 38.96 occurrences per million) had significantly higher frequencies of occurrence (based on values from the Kucera and Francis (1967) word counts) than low-frequency words (mean = 2.52 occurrences per million; F (1,112) = 462.08, $p < .001$). Neighborhood density was calculated by determining the number of words that could be created from a target word by adding, deleting, or substituting a phoneme. Words in the SPARSE NEIGHBORHOOD conditions (mean = 13.23 words) had significantly fewer neighbors, than the words in the DENSE NEIGHBORHOOD conditions (mean = 24.40 words; F (1,112) = 247.17, $p < .001$). Neighborhood Frequency, defined as the mean word frequency of all the neighbors of a target word, was also calculated using an on-line database. Words in the HIGH NEIGHBORHOOD FREQUENCY conditions (mean = 217.88 words per million) had neighbors with significantly higher values of word frequency than the neighbors of words in the LOW NEIGHBORHOOD FREQUENCY conditions (mean = 40.98 words per million; F (1,112) = 255.36, $p < .001$).

The questions for inducing TOT states were based on the definitions for each word found in the Webster's New Collegiate Dictionary (1979). A pilot study using another group of participants (a young and an old adult group) determined if the target word was an appropriate answer to the question. Each question and associated target word was presented to participants for a rating of how well the word

CONDITION	Familiarity	Word Frequency	Neighborhood Density	Neighborhood Frequency
High Frequency Dense Neighborhood High Neigh. Freq.	6.92 (.19)	52.33 (67.88)	25.6 (4.51)	273.05 (403.49)
High Frequency Dense Neighborhood Low Neigh. Freq.	6.81 (.32)	31.73 (31.92)	22.6 (2.47)	51.05 (51.95)
High Frequency Sparse Neighborhood High Neigh. Freq.	6.89 (.17)	37.86 (27.98)	14.0 (2.59)	170.35 (180.30)
High Frequency Sparse Neighborhood Low Neigh. Freq.	6.80 (.42)	33.93 (26.71)	14.0 (3.25)	36.32 (38.92)
Low Frequency Dense Neighborhood	6.51	3.86	26.3	325.66

High Neigh. Freq.	(.56)	(2.79)	(5.09)	(273.09)
Low Frequency Dense Neighborhood Low Neigh. Freq.	6.68 (.37)	3.80 (2.62)	23.1 (3.63)	40.99 (48.21)
Low Frequency Sparse Neighborhood High Neigh. Freq.	6.20 (.94)	1.14 (.36)	12.3 (4.75)	102.47 (368.20)
Low Frequency Sparse Neighborhood Low Neigh. Freq.	6.60 (.37)	1.26 (.45)	12.6 (4.15)	35.55 (85.26)

Table 1. Mean familiarity, word frequency, neighborhood density, and neighborhood frequency values for the eight conditions of target words in the TOT elicitation task (standard deviations are in parenthesis).

answered the question (1 = "Does not answer the question at all," 4 = "Answer acceptable, but there is a better word, " 7 = "Answers the question very well"). Any question that participants rated as not being appropriately answered by the target word (a mean rating of 5 or below) was modified until additional pilot study deemed the target word an appropriate response to the question.

Each target word and question pair had three additional foils that were of the same word class and were semantically similar. The foils were derived from the same source as the questions.

Procedure

The procedure followed that used by Burke et al. (1991). Participants were given the vocabulary sub-test of the Weschler Adult Intelligence Scale. They then heard a description of the TOT state from Brown and McNeill (1966), and were guided through a practice session. The practice session consisted of four questions that were similar to those used in the experimental session of the TOT elicitation task. The TOT-inducing questions were presented on an IBM-compatible computer. For each question, participants typed their responses on the computer keyboard.

A flow-chart description of the TOT elicitation task, adapted from Burke et al. (1991), is presented in Figure 1. For each question, three options were initially presented to the participants: K if they knew the answer, D if they didn't know the answer and T if the answer was on the tip of their tongue. After providing the initial response (K, D or T) for each TOT-inducing question, participants were asked to rate how familiar they were with the word in question on a scale from 1 ("unfamiliar") to 7 ("very familiar"). As in Burke et al. (1991), participants then rated how certain they were that they could recall the word in question on a scale from 1("uncertain") to 7 ("certain").

If participants had initially responded K (they knew the answer), they were asked to type the response to the question. If they were correct, appropriate feedback was given and the next trial was initiated with the presentation of a new TOT-inducing question. If the participant responded with an incorrect answer, they were given multiple choices from which to select a response. If they selected the correct option, appropriate feedback was given. If they selected an incorrect option, they were provided with the correct response and a new trial began.

If the initial response was D (they didn't know the answer), participants immediately received multiple options from which to select (after answering the questions regarding familiarity and likelihood of recall for the word). Appropriate feedback was again provided for each response before a new trial was initiated.

If the participant indicated that they were in a TOT state by initially selecting T, a number of other questions followed the two rating estimates. Participants were asked to provide, if possible: the initial sound of the word, the final sound of the word, the number of syllables in the word, and any similar sounding words that persistently came to mind. As in the "Don't Know" response, they were given multiple options to select from with the additional option of "None of the above," and received appropriate feedback.

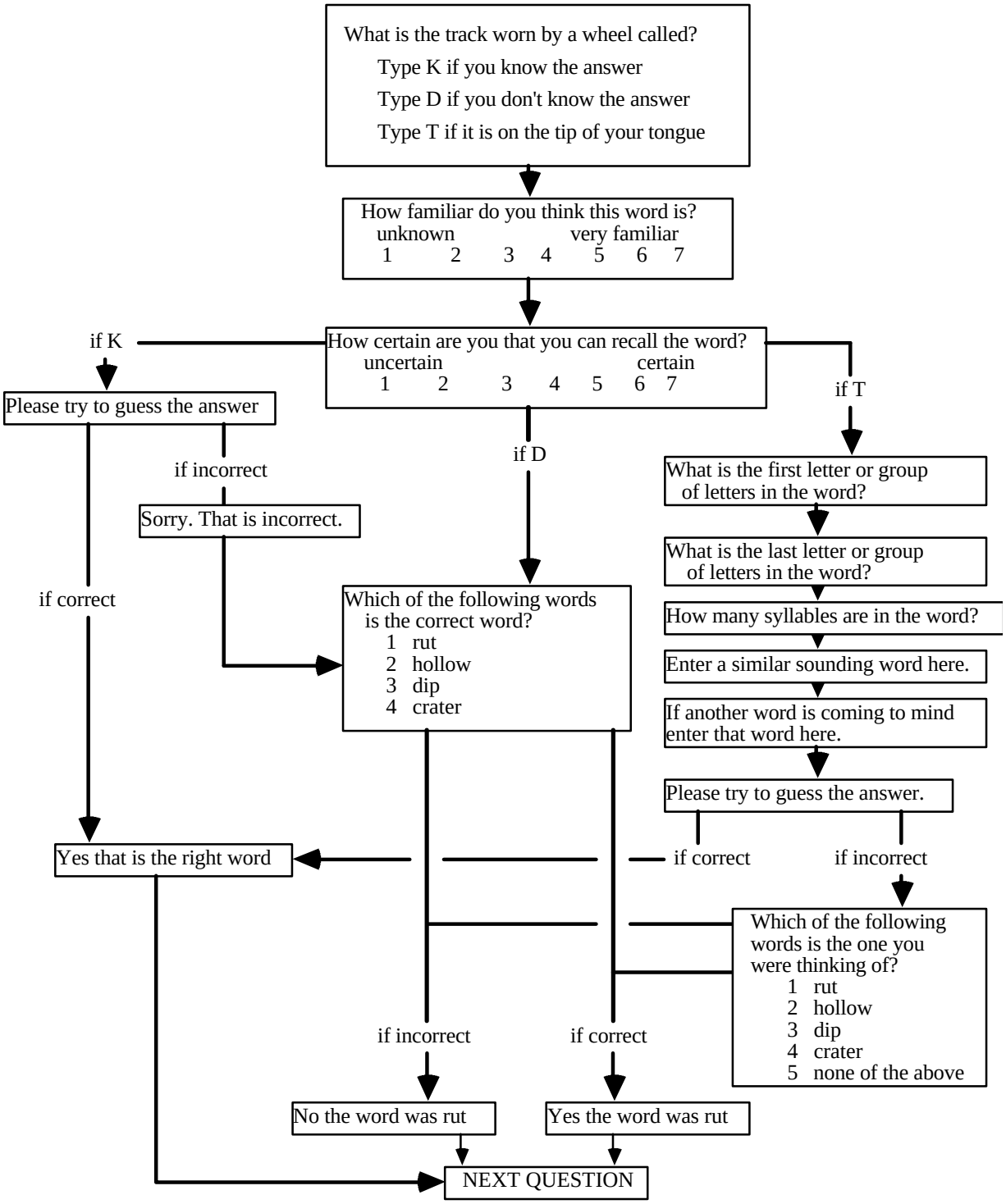


Figure 1. Flow chart description of the TOT elicitation task.

A brief practice session preceded the experiment. For the first question, participants were told to select K and to answer the questions that followed. For the second question, participants were told to select D and to answer the questions that followed. For the third question, participants were told to select T and to answer the questions that followed. For the last practice question, participants were allowed to select the option that was appropriate for their present state. This was done to familiarize the participants with all the possible types of questions that they might encounter. Upon completion of the practice session, participants began the experimental session of the TOT elicitation task and proceeded at their own pace. Participants were tested individually and received the 120 TOT-inducing questions in a different random order. As in Burke et al. (1991), participants could not backtrack to earlier questions, and the computer scored only the first three letters of an answer to minimize errors due to misspellings.

Results

Separate repeated-measures Analyses of Variance (ANOVAS) were performed on each type of response across the eight conditions formed by the orthogonal combination of two levels of three variables (word frequency, neighborhood density, and neighborhood frequency). Analyses by participants (F_1) and items (F_2) are reported. Additional analyses were also performed on the "interlopers," (i.e., words that persistently came to mind instead of the desired word), and the other types of partially retrieved information.

Familiarity Rating

Participants were asked to rate how familiar they were with each word they were being questioned about on a scale from 1 ("Unknown") to 7 ("Very Familiar"). There were no significant differences in either of the analyses (by F_1 and F_2) across the eight conditions in the familiarity ratings. The mean familiarity rating across the eight conditions (formed by crossing two levels of three variables) was 4.6 (range of 4.3 to 5.1 across conditions), suggesting that all the words were at least somewhat familiar to the participants.

Recall Rating

Participants were also asked to rate how certain they were that they could recall the word they were being questioned about on a scale from 1 ("Uncertain") to 7 ("Certain"). Again, there were no significant differences in either of the analyses (by F_1 and F_2) across the eight conditions in the certainty ratings. The mean certainty rating across the eight conditions was 4.4 (range of 4.1 to 4.8 across conditions). This finding suggests that none of the stimulus variables (word frequency, neighborhood density, or neighborhood frequency) affected listeners' confidence in their ability to recall appropriate answers.

Total "Know" Responses

"Know" responses were analyzed in several ways. The first analysis examined the total number of "know" responses made. The second analysis examined the number of "know" responses that were actually correct. The final analysis examined the number of "know" responses that were initially incorrect when the participant did not have any options to choose from, but which were correctly selected once response alternatives were provided.

For the total number of "know" responses, a main effect of word frequency was found by participants (F_1 (1,22) = 5.74, $p < .05$), but not by items ($F_2 < 1$), such that slightly more "know" responses were made to high frequency words (mean = 10.40) than to low frequency words (mean = 9.81). A significant interaction between neighborhood density and neighborhood frequency was also found by participants (F_1 (1,22) = 16.57, $p < .001$) but not by items (F_2 (1,112) = 2.93, $p > .10$). Slightly more "know" responses were made if the word was high in both neighborhood density and neighborhood frequency (mean = 10.41) or low in both neighborhood density and neighborhood frequency (mean = 10.87) than if one variable was high and the other variable was low (high neighborhood frequency/sparse neighborhood mean = 9.63; low neighborhood frequency/dense neighborhood mean = 9.52). Finally, a significant interaction was found by participants (F_1 (1,22) = 6.72, $p < .05$) but not by items ($F_2 < 1$) between neighborhood density, word frequency, and neighborhood frequency. The means for this interaction can be found in the top portion of Table 2.

Younger Adults

	High Frequency				Low Frequency			
	Dense Neighborhood		Sparse Neighborhood		Dense Neighborhood		Sparse Neighborhood	
	Hi NHF	Lo NHF	Hi NHF	Lo NHF	Hi NHF	Lo NHF	Hi NHF	Lo NHF
Know Responses	11.0 (2.2)	9.2 (2.6)	10.0 (2.0)	11.3 (1.5)	9.7 (2.0)	9.8 (2.2)	9.2 (9.2)	10.4 (2.3)
Responses that were correct	6.0 (2.0)	6.1 (1.3)	4.7 (2.3)	7.0 (1.9)	4.9 (1.7)	6.5 (2.1)	4.6 (2.0)	3.5 (2.2)
Correct selection from multiple options	4.2 (1.7)	2.0 (1.2)	4.6 (2.2)	3.3 (1.8)	3.6 (1.9)	3.1 (2.0)	4.0 (1.9)	4.2 (1.9)
Don't know response	3.6 (2.3)	5.6 (2.6)	4.7 (2.1)	3.5 (1.5)	5.1 (2.0)	4.9 (2.1)	4.9 (2.1)	4.0 (2.1)
Correct response	2.6 (1.8)	3.7 (2.0)	3.4 (1.9)	2.0 (1.1)	3.7 (1.7)	4.2 (2.1)	3.6 (1.8)	3.0 (1.6)
TOT Reported	.26 (.54)	.13 (.34)	.21 (.67)	.17 (.38)	.13 (.45)	.22 (.42)	.83 (.93)	.57 (1.12)
Resolved	3	1	0	0	1	3	6	5

Table 2. Means for each type of response for young adults. Standard deviations are in parenthesis. *Note*. NHF = neighborhood frequency. The value for "TOT Reported" is the mean across 23 participants, whereas the value for "Resolved" is the raw number of resolved TOT states per condition.

Correctly Answered "Know" Responses

For the number of "know" responses that were actually correct, participants responded with the correct word more often for words with dense neighborhoods (6.0 correct responses) than for words with sparse neighborhoods (4.9 correct responses; F_1 (1,22) = 24.68, $p < .001$; F_2 (1,112) = 4.65, $p < .05$). No other differences were significant in the participant or items analyses (all F's < 1). The means across each condition can be found in the top portion of Table 2.

"Know" Responses with Options

Of the "Know" responses that participants did not initially provide the correct word for, participants were relatively accurate in selecting the correct word from among the options displayed. There were no significant differences among the eight conditions in the participant or items analyses (all F's < 1) for the number of correct selections made when participants selected from the multiple options presented. These means are also displayed in the top portion of Table 2.

Total "Don't know" Responses

The means for the number of "don't know" responses are shown in the middle of Table 2. Similar to the "know" responses, the "don't know" responses were analyzed in several ways. First, the total number of "don't know" responses were analyzed. When a "don't know" response was made, participants were presented with several choices to select from. We then analyzed the number of correct selections made from the multiple choices.

The main effect of neighborhood density for the number of "don't know" responses made was significant by participants (F_1 (1,22) = 5.06, $p < .05$) but not by items (F_2 < 1). Participants were slightly more likely to respond "don't know" for target words from dense neighborhoods (mean = 4.8) than for target words from sparse neighborhoods (mean = 4.2). Significant (by participants only) interactions were also observed between neighborhood density and neighborhood frequency (F_1 (1,22) = 17.04, $p < .001$) and between neighborhood density, word frequency, and neighborhood frequency (F_1 (1,22) = 10.06, $p < .01$). For the two-way interaction, slightly more "don't know" responses were made to words with dense neighborhoods and either low neighborhood frequency (mean = 5.3) or high neighborhood frequency (mean = 4.3) than for words from sparse neighborhoods and either low neighborhood frequency (mean = 3.7) or high neighborhood frequency (mean = 4.8). The means for the three-way interaction are displayed in Table 2. However, none of these effects were significant by items (all F's < 2.6, $p > .10$).

Correct "Don't know" Responses

When participants made a "Don't know" response, they were presented with multiple options from which to choose. No significant differences (by either participants or items) were found for the number of "Don't know" responses for which the correct choice was selected (all F's < 1).

Total TOT Responses

The bottom portion of Table 2 displays the mean number of TOT responses as a function of word frequency, neighborhood density, and neighborhood frequency. Several analyses were performed on the TOT responses. First, we analyzed the total number of TOT responses. A main effect of word frequency

was found (F_1 (1,22) = 11.30, $p < .01$; F_2 (1,112) = 4.98, $p < .05$) such that more TOT states were elicited for low frequency words (mean = .43) than for high frequency words (mean = .20). A main effect of neighborhood density was also found (F_1 (1,22) = 10.40, $p < .01$; F_2 (1,112) = 5.93, $p < .05$) such that more TOT states were elicited for words with sparse neighborhoods (mean = .45) than for words with dense neighborhoods (mean = .18). No main effect of neighborhood frequency was found ($F < 1.2$ for both participant and item).

The number of reported TOTs as a function of word frequency and neighborhood density are displayed in Figure 2. A significant interaction between word frequency and neighborhood density (F_1 (1,22) = 16.72, $p < .001$; F_2 (1,112) = 5.93, $p < .05$) was found. This interaction was due to more TOT states being elicited for words that had low frequency and sparse neighborhoods (the striped bar on the right; mean = .70) than for words in the other conditions. Pair-wise comparisons between words that had low frequency and sparse neighborhoods and the other conditions support this finding: 1) words that had low frequency and dense neighborhoods (the clear bar on the right; mean = .17; F (1,22) = 33.49, $p < .001$), 2) words that had high frequency and sparse neighborhoods (the striped bar on the left; mean = .20; F (1,22) = 30.70, $p < .001$), 3) words that had high frequency and dense neighborhoods (the clear bar on the left; mean = .20; F (1,22) = 30.71, $p < .001$). No other differences nor interactions were significant (all F's < 1).

Across participants, the range of TOTs reported went from a high of 18 to a low of 1 reported TOT. The average rate of TOT responses in Experiment 1 was 3%. This value is considerably lower than the 13.2% and 19.9% reported in Harley and Bown (1998) and the 10.9% reported in Burke et al. (1991). The lower percentage of elicited TOTs in the current study may be due to the shorter length of the words used in this experiment. All of the stimuli in the present experiment were monosyllabic words, whereas other studies (e.g., Jones, 1989; Jones & Langford, 1987; Burke et al., 1991; Harley & Bown, 1998) have typically used bisyllabic or multisyllabic words. Recall that Harley and Bown (1998) found that the frequency of TOTs varied as a function of word length such that TOTs were more likely among longer words.

"Subjective" versus "Objective " TOTs

Additional analyses were conducted on the reported TOT states to examine participants' ability to resolve the TOT states and determine how much partial information was available in those TOTs that were not resolved. A distinction is often made between "subjective" TOT states, defined as the items self-reported by participants to be TOTs (Jones & Langford, 1987) and "objective" TOT states, defined as those TOT states that were correctly resolved. Nineteen of the 58 reported TOT states were resolved correctly without having to receive multiple options to select from (25 were guessed correctly when the participants received multiple options to select from, and 14 were guessed incorrectly by participants when presented with multiple options). A Chi-square analysis was used to compare the distribution of resolved TOT states to the distribution of reported TOT states. No difference between the two distributions was found (X^2 4.12, p = .76). Thus, the pattern of "subjective" TOT states was similar to the pattern of "objective" TOT states.

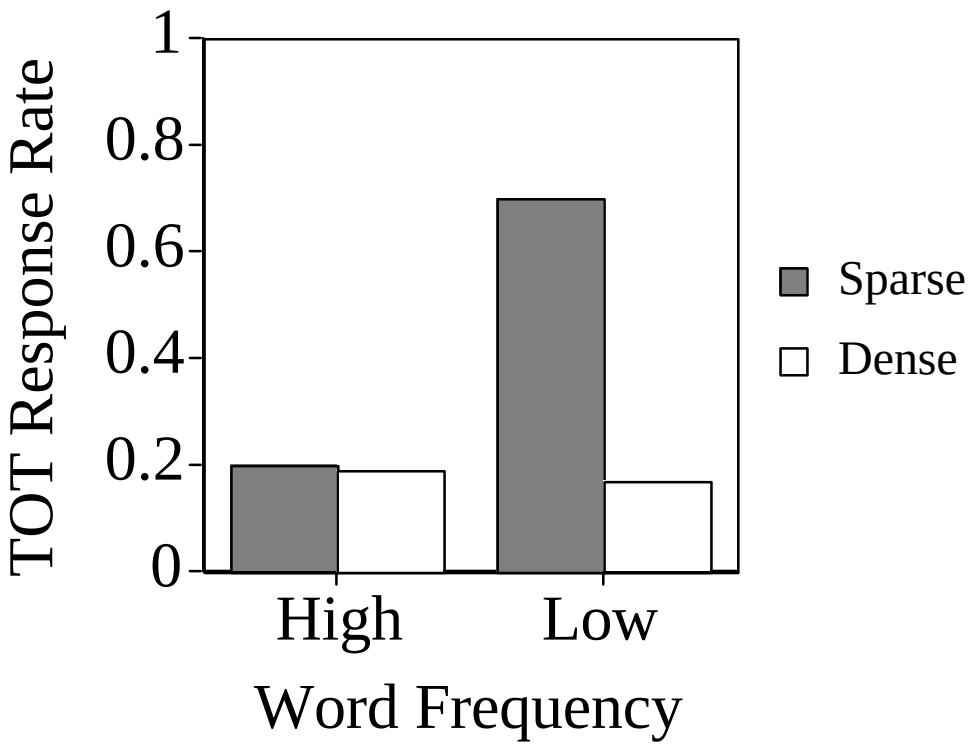


Figure 2. The number of TOTs for young adults as a function of word frequency and neighborhood density.

Analysis of the Interlopers

Among the 19 correctly resolved TOT states 10 interlopers (interfering words) were reported and all were either semantically or phonologically related to the target: Seven of the words were semantically related, 1 word was phonologically related, and 2 words were both semantically and phonologically related. We compared the frequency counts, familiarity ratings, neighborhood density, and neighborhood frequency of the interlopers and the targets. Information on 6 of the 10 interlopers was obtained from the same lexical database used to create the stimulus set. A paired *t*-test found no difference in the familiarity of the targets (mean = 6.4) and interlopers (mean = 6.8; t (1,10) = 1.32, p = .21) and no difference in the neighborhood frequency counts of the targets (mean = 73.56) and interlopers (mean = 18.53; t (1,10) = 1.48, p = .17). There was also no difference in the frequency counts (from Kucera & Francis, 1967) of the targets (mean = 12.50) and interlopers (mean = 18.83; t (1,10) < 1). This finding failed to replicate the results of Reason and Lucas (1983) who found that interlopers tended to be rated more frequent than target words.

One possible account of the differences between our experiment and the diary study of Reason and Lucas (1983) may be due to the methodology used in each study. As Reason and Lucas (1983) discuss, cognitive diary studies are subject to several types of biases in the selection and recording of events in the

diary. Perhaps only those interlopers that were more frequent were the interlopers that were reported by participants in the Reason and Lucas (1983) diary studies.

Another possible explanation for the differences between our findings and those of Reason and Lucas (1983) is the number of interlopers analyzed in each study. In our study we examined six interlopers, whereas Reason and Lucas (1983) examined 40 interlopers in their first diary study and 22 interlopers in their second diary study. The sample size of interlopers in our experiment may have been too small to provide adequate power to detect a difference in our statistical analysis.

There are also differences between the two studies in the way that word frequency was assessed. Reason and Lucas (1983) asked for subjective ratings of word frequency, whereas we used objective word counts (i.e., Kucera & Francis, 1967). Moreover, we used a *t*-test to quantitatively assess the difference in word frequency, whereas Reason and Lucas (1983) interpreted the differences in word frequency ratings qualitatively. (See Vitevitch, 1997 for a comparison of qualitative (e.g., chi-square) vs. quantitative (e.g., ANOVA) assessments of a speech error corpus, and the differences that may result in interpretation from the different assessment methods.) However, qualitative (chi-square) analysis performed on the current data failed to reveal significant effects for word frequency and neighborhood frequency (all $p > .10$), arguing against differences in the type of analyses as the primary factor responsible for the differences between the two studies.

However, a significant difference was found in the neighborhood density values for the interlopers (mean = 7.2 neighbors) and the targets (mean = 25.1 neighbors; $F(1,10) = 18.56$, $p < .01$). This finding suggests that when an interloper interfered with a participant's ability to correctly retrieve a target item, the interloper was likely to reside in a lower-density neighborhood than the target. Harley and Bown (1998) reported a similar finding in their study. It is difficult to clearly interpret this result in the current experiment because the interlopers were longer (3 words were 1 syllable long, 6 words were bisyllabic, and 1 word was trisyllabic) than the target words (all were monosyllabic words). Recall that Pisoni, Nusbaum, Luce and Slowiaczek (1985) found an inverse relationship between word length and neighborhood density: Short words tend to have many similar sounding items, whereas longer words tend to have fewer similar sounding items. The difference in word length between the targets and the interlopers may reflect a conscious search strategy adopted to resolve the TOT: if one is having a difficult time retrieving a lexical item (i.e., one is in a TOT state), it must be because the item is extremely "unusual" in some way. Such conscious awareness may result in the search for and retrieval of longer words, which are less common (i.e., "unusual") than monosyllabic words in English (Zipf, 1965). Conscious search strategies may also over-ride any affects of automatic processes, such as frequency-based biases, which may account for the non-significant difference in word-frequency between targets and interlopers observed in the present experiment.

Reports of Partial Information

Participants were also asked to report any partial information regarding the word they were attempting to retrieve. Participants correctly reported the number of syllables in the target word 5 out of 6 times, correctly reported the first letter of the target word 6 out of 7 times, and correctly reported the last letter of the target word 6 out of 6 times.

Discussion

The results from the current experiment replicate and extend the results of several previous studies examining factors that may influence the rate of TOTs (Brown & McNeill, 1966; Burke et al., 1991; Harley & Bown, 1998). Specifically, our findings indicated that TOT states were more likely for words with low-frequency and sparse neighborhoods. Several hypotheses have been advanced to account for the role of similar sounding words in TOT states. One hypothesis states that similar sounding words interfere with the retrieval of the lexeme, or phonological word form (Jones, 1989, Jones & Langford, 1987; Maylor, 1990; Woodworth, 1929). However, the current results—more TOT states for words with few rather than many similar sounding words—do not support this hypothesis. Instead, our results support the hypothesis that similar sounding words aid in the retrieval of intended representations (Brown, 1991; Burke et al., 1991; Meyer & Bock, 1992; Perfect & Hanley, 1992). Moreover, the finding of significantly more correctly answered "Know" responses for words with dense neighborhoods (6.0 correct answers) than for words with sparse neighborhoods (4.9 correct answers) further suggests that the number of similar sounding neighbors aids the retrieval of the intended word-form.

Our results are also consistent with predictions derived from NST regarding the influence of frequency and density on the probability of TOTs (Burke et al., 1991; MacKay, 1987). In NST, TOT states are the result of insufficient activation of the intended representations, particularly phonological nodes. Representations fail to be sufficiently activated because the connections between nodes transmit priming less efficiently. Transmission deficits result from the nodes not being used very often (frequency of use), not being used lately (recency of use), or because of aging. Because less common (low frequency) words do not receive priming as efficiently as more common (high frequency words), we predicted that more TOT states would occur for low frequency words than for high frequency words. The results of the current experiment support that prediction, replicating several other studies examining the role of word-frequency in lexicalization (e.g., Brown & McNeill, 1966; Burke, MacKay, Worthley, & Wade, 1991; Harley & Bown, 1998).

We also described neighborhood density in terms of the frequency of the sub-lexical constituents of a word. Specifically, words with dense neighborhoods are comprised of phonemes that are frequent and shared by many words, whereas words with sparse neighborhoods are comprised of phonemes that are less frequent and shared by few words (Vitevitch, Luce, Pisoni, & Auer, 1999). Phonemes (represented by phonological nodes in NST) that constitute words with dense neighborhoods receive priming more frequently and more recently than phonological nodes that constitute words with sparse neighborhoods. The difference in the frequency and recency of priming further strengthens the connections between words with dense neighborhoods and their constituent phonemes over time, whereas the connections between phonological nodes and words with sparse neighborhoods become weaker over time. Thus, the phonological information associated with a word with a dense neighborhood is more efficiently retrieved than the phonological information associated with a word with a sparse neighborhood. This hypothesis is supported by the finding that more TOT states were elicited for words with sparse neighborhoods than for words with dense neighborhoods (see also Harley & Bown, 1998). Additional support for the hypothesis that phonological information associated with dense rather than sparse neighborhoods is more readily available comes from the interesting observation that more of the interlopers that were either phonologically related or were both phonologically and semantically related occurred for target words with dense (66%) rather than sparse neighborhoods (33%).

Given the frequency-based effects (i.e., word-frequency and the frequency of the phonemes making up those words) on lexicalization that were observed in the present experiment, it is somewhat surprising that the frequency of the neighbors—defined as neighborhood frequency—did not appear to have an influence on the number of TOT states elicited. The effects of neighborhood frequency may have been

overshadowed by the more dominant influences of word frequency and neighborhood density. Although neighborhood frequency effects were not significant in the current experiment, the manipulation of this lexical characteristic is an important extension of Harley and Bown (1998). Both experiments by Harley and Bown failed to take this variable into consideration and neither manipulated nor controlled neighborhood frequency in their stimuli.

Of further interest is how the effects of these lexical characteristics on the process of lexicalization change over the life span. Work by Sommers (1996) has shown that the influence of neighborhood density on lexical retrieval in speech perception changes with age. Specifically, Sommers (1996) found that older adults were less accurate than younger adults at identifying words with many similar sounding neighbors, even under conditions producing similar identification performance for words with few similar sounding neighbors. One question that we wanted to address in the present study was whether the age-related changes in neighborhood density found in speech perception would also be found in speech production?

The findings of Burke et al. (1991) suggest that aging does affect speech production: older adults have more TOT states than younger adults. They attributed this increase in TOT states with age to an age-linked transmission deficit that decreased the availability of partial (phonological) information. Maylor (1990) found similar difficulties in accessing phonological information in older adults. Is this increase in TOT states with age differentially influenced by neighborhood density? Can this change with age also inform us about the nature of the decline associated with aging? To examine these questions we used the same stimuli from Experiment 1 to elicit TOT states in elderly adults.

Experiment 2

Previous studies have found more TOT states among elderly adults than among younger adults (Burke et al., 1991; Burke & Laver, 1990; MacKay & Burke, 1990; Rastle & Burke, 1996). Other research (Sommers, 1996; Sommers & Danielson, 1999) has demonstrated age-related declines in the ability to access the lexicon in spoken-word recognition. In Experiment 2, we wanted to examine how differential access to the lexicon in younger and older adults would affect the frequency of TOTs in these two populations.

NST proposes that transmission of priming between nodes becomes less efficient as a function of age if "...other characteristics of the nodes are assumed to be equal...especially their history of prior practice and their recency of activation." (pg. 222, MacKay & Burke, 1990). This principle generally explains why young adults can retrieve information more efficiently than can older adults. However, what happens if the history of the nodes in young and older adults is not equal, as in the case of words with dense and sparse neighborhoods? Recall that neighborhood density can be viewed as a frequency effect among the sub-lexical components of a word (Vitevitch, Luce, Pisoni & Auer, 1999). Thus, words with dense neighborhoods have sub-lexical components that have been activated more frequently than the sub-lexical components of words with sparse neighborhoods.

Furthermore, the difference in the frequency with which components of words with dense neighborhoods versus those with sparse neighborhoods are activated increases with age. To illustrate this point imagine that at time t_1 word *x* has a frequency of 2 occurrences per million and word *y* has a frequency of 100 occurrences per million, a difference of 98 occurrences per million. If one is exposed to a million words a year, in ten years, time t_{10}, word *x* will have been heard 20 times and word *y* will have been heard 1,000 times. The difference between words *x* and *y* from time t_1 to time t_{10} is now 980 occurrences per million, much greater than the original difference. Thus, even though overall vocabulary, as measured

by the vocabulary sub-test of the WAIS, for example, may not have increased, the same person at time t_{10} will have been exposed to more instances of a particular word than at time t_1. This difference in exposure to words may also be observed in different people that are matched on all relevant variables, such as IQ or vocabulary size, but that differ in age.

The connections to the components of words that have a higher exposure will also have been activated and strengthened over time. With the increased linguistic exposure as a function of age, words with dense neighborhoods and strong connections to sub-lexical components may become at least partially "insulated" from the effects of age-related transmission deficits. In contrast, words with sparse neighborhoods will have much weaker connections to sub-lexical components and will be more susceptible to age-related transmission deficits. We predict that age-related transmission deficits will result in more TOT states for older adults than for younger adults, replicating the results of Burke et al. (1991). We further predict that the major source of TOT states in older adults will be for words with sparse neighborhoods. We hypothesize that the differences between older and younger adults is a result of differential exposure to words as a function of age, and the differential influence of age-related transmission deficits on the sub-lexical components of words with dense and sparse neighborhoods. To examine how lexicalization may be affected by normal aging, we used the same stimuli that were used in Experiment 1 and presented them to elderly participants.

Method

Participants

Twenty-four native English-speaking adults were recruited from the Washington University community. All subjects reported no history of a speech or hearing disorder and received $20 for their participation. The mean age of these participants was 70.3 years (SD = 4.9). Mean WAIS vocabulary scores for this group of older adults did not differ significantly from vocabulary scores for younger adults ($F < 1.2$).

Materials and Procedure

The same materials and procedure used in Experiment 1 were used in the current experiment.

Results

Analyses similar to those in Experiment 1 were performed in the current experiment. In addition, the data from the two experiments were combined to compare the performance of younger and older adults. In analyses comparing the performance of younger and older adults Experiment (Experiment 1 versus Experiment 2) was treated as a between-participants factor.

Familiarity Rating

As with the young adults, there were no significant differences in familiarity ratings ($F < 3.14$, $p > .08$ for both participants and items) across the eight conditions. The mean familiarity rating across all conditions for older adults was 5.2 (range of 4.8 to 5.4 across conditions), suggesting that all the words were at least somewhat familiar to the participants. There was a significant difference in the familiarity ratings given by the younger and older adults by participants (F_1 (1, 45) = 5.45, $p < .05$) but not by items

(F_2 < 1). Overall, older adults tended to give higher familiarity ratings than younger adults. No other main effects or interactions were significant (all *F*'s< 1).

Recall Rating

The mean certainty rating for older adults regarding their ability to recall the word they were being questioned about was 5.2 (range of 4.7 to 5.5 across conditions). Certainty ratings did not vary significantly across the eight conditions. A main effect of age was found in the recall ratings by participants (F_1 (1, 45) = 15.50, p < .001) but not by items (F_2 < 1). Overall, older adults gave higher certainty ratings than younger adults. No other main effects or interactions were significant (all *F*'s< 1).

Total "Know" Responses

As with the younger adults, "Know" responses from the older adults were analyzed in several ways. The first analysis examined the total number of "know" responses. The second analysis examined the number of those "know" responses that were actually correct. The final analysis examined the number of "know" responses that were initially incorrect when the participant did not have any options to choose from, but which were correctly selected when options were provided. Mean number of "Know" responses as a function of word frequency, neighborhood density, and neighborhood frequency are displayed in the top panel of Table 3.

For the older adults, a main effect of word frequency was found for the total number of "Know" responses by participants (F_1 (1,23) = 5.72, p < .05), but not by items (F_2 < 1). There were slightly more "Know" responses to high-frequency words (mean = 11.61) than to low-frequency words (mean = 11.21). No other main effects or interactions were significant (all F < 1) for the number of "Know" responses.

In comparing the younger and older adults, a main effect of age was found for the total number of "know responses" (F_1 (1,45) = 3.94, p < .05; F_2 (1,224) = 23.07, p < .001) such that older adults responded "know" more often (mean = 11.42) than younger adults (mean = 10.10), replicating one of the findings of Burke et al. (1991).

Correctly Answered "Know" Responses

For older adults, there were no main effects or interactions that were significantly different in both the participants and item analyses for the number of "Know" responses that were actually answered correctly. There were also no main effects nor interactions significant by both participants and items when the number of "Know" responses that were actually answered correctly by the older adults was compared to the number of "Know" responses that were actually answered correctly by the younger adults (all *F*'s <1).

"Know" Responses with Options

Of the "Know" responses that the older participants did not initially provide the correct word for, they were relatively accurate in selecting the correct word from among the options displayed. There were no significant differences among the eight conditions for the number of correct selections made from the multiple options presented to the participants (all *F*'s < 1). Furthermore, there were no differences between the younger and older adults in the number of "Know" responses that they answered correctly when presented with multiple options (all *F*'s < 1).

Total "Don't know" Responses

The mean number of "don't know" responses as a function of word frequency, neighborhood density, and neighborhood frequency are displayed in the middle panel of Table 3. "Don't know" responses for older adults did not vary across the eight conditions (all F's< 1). A main effect of age was found for the number of "Don't know" responses (F_1 (1,45) = 5.65, $p < .05$; F_2 (1,224) = 15.09, $p < .001$) such that older adults responded "Don't Know" fewer times (mean = 3.0) than younger adults (mean = 4.6), also replicating one of the findings of Burke et al. (1991). No other main effects or interactions across the age groups were significant by both analyses.

Older Adults

	High Frequency				Low Frequency			
	Dense Neighborhood		Sparse Neighborhood		Dense Neighborhood		Sparse Neighborhood	
	Hi NHF	Lo NHF	Hi NHF	Lo NHF	Hi NHF	Lo NHF	Hi NHF	Lo NHF
Know Total	11.7 (3.3)	11.5 (3.2)	11.1 (2.8)	12.0 (3.0)	11.2 (3.2)	11.4 (3.1)	11.5 (3.0)	10.6 (2.7)
Responses that were correct	6.9 (2.8)	7.4 (2.3)	5.1 (2.5)	7.8 (1.7)	6.8 (2.1)	6.6 (2.2)	7.6 (1.9)	4.1 (2.6)
Correct selection from multiple options	4.1 (2.4)	2.8 (1.9)	4.9 (2.2)	2.5 (1.3)	3.6 (1.8)	3.9 (1.9)	3.1 (1.6)	4.1 (2.3)
Don't know Total	2.6 (3.1)	3.0 (3.0)	3.4 (2.9)	2.5 (3.0)	3.2 (3.2)	2.8 (3.1)	3.2 (2.9)	3.2 (2.4)
Correct response	1.9 (2.2)	2.1 (2.4)	2.5 (1.9)	1.8 (1.8)	2.7 (2.9)	2.0 (2.5)	2.2 (2.8)	2.3 (1.8)
TOT Reported	.58 (1.02)	.42 (.65)	.38 (.57)	.46 (1.10)	.50 (.78)	.67 (.87)	.29 (.69)	1.0 (1.51)
Resolved	4	8	5	8	8	9	3	8

Table 3. Means for each type of response for old adults. Standard deviations are in parenthesis. *Note*. NHF = neighborhood frequency. The value for "TOT Reported" is the mean across 24 participants, whereas the value for "Resolved" is the raw number of resolved TOT states per condition.

Correct "Don't know" Responses

For the older adults, there was no difference among the conditions for the number of "Don't know" responses answered correctly from among the multiple options presented (both F's < 1). However, when the number of "Don't know" responses answered correctly by the older adults was compared to the number of "Don't know" responses answered correctly by the younger adults, a significant difference was found by subjects (F_1 (1,45) = 4.81, p < .05) but not by items (F_2 < 1). Younger adults tended to answer more of the "Don't know" responses correctly (mean = 3.3) when presented with multiple options than the older adults (mean = 2.1). This marginal difference may reflect different criteria for selecting the "don't know" option, as is indicated by a reduced likelihood to select the "don't know" response for older adults. The adoption of different "error criteria" by young and older adults is also consistent with NST (MacKay & Burke, 1990).

Total TOT Responses for Older Adults

The mean number of TOT responses for older adults as a function of word frequency, neighborhood density, and neighborhood frequency are shown in the bottom panel of Table 3. The main effect of word frequency was marginally significant by participants but not by items (F_1 (1,23) = 3.54, p = .07; F_2 (1,112) = 2.00, p = .15). There tended to be more TOT states elicited for low-frequency words (mean = .62) than for high-frequency words (mean = .46). Main effects of neighborhood density and neighborhood frequency were not significant (all F < 1).

However, an interaction between neighborhood frequency and neighborhood density was significant by participants (F_1 (1,23) = 5.08, p < .05) and marginally significant by items (F_2 (1,112) = 3.13, p = .08). For words from sparse neighborhoods, more TOTs were observed for words with low neighborhood-frequency (mean = .75) than for words with high neighborhood-frequency (mean = .33; F (1,23) = 10.18, p < .01). For words from dense neighborhoods, however, neighborhood frequency did not influence the number of TOTs (means for each = .54). No other differences were significant (all F's < 1).

There was also an interaction between word frequency and neighborhood frequency (F_1 (1,23) = 8.12, p < .01; F_2 (1,112) = 4.51, p = .05). More TOT states were elicited for words that had low word- and neighborhood-frequency (mean = .85) than for words that had high word- and neighborhood-frequency (mean = .48; F (1,23) = 9.13, p < .01). These results are displayed in Figure 3. There was no difference between low frequency words (mean = .40) and high frequency words (mean = .48) with high neighborhood frequency (F < 1). No other differences nor interactions were significant (all F's < 1). The mean number of "TOT" responses for each condition is displayed in Table 3. TOT responses for older participants ranged from a high of 20 to a low of one.

"Subjective" versus "Objective" TOTs for Older Adults

Additional analyses were conducted to examine participants' ability to resolve TOT states and to determine how much partial information was available when TOT states could not be resolved. Fifty-three of the 106 reported TOT states ("subjective TOT") were resolved correctly ("objective TOT") without having to receive multiple options to select from (36 were guessed correctly when the participants received multiple options to select from, and 17 were guessed incorrectly by participants when presented with multiple options). As in Experiment 1, a Chi-square analysis was used to compare the distribution of resolved ("objective") TOT states in the older adults to the distribution of reported ("subjective") TOT states in the older adults. No difference between the two distributions was found (X^2 4.41, p = .73), suggesting that the pattern of "subjective" TOT states is similar to the pattern of "objective" TOT states.

Among the 53 correctly resolved TOT states, 18 interlopers (interfering words) were reported: 13 of the words were semantically related, 0 words were phonologically related, 4 words were both semantically and phonologically related, and 1 word bore no obvious relationship to the target word. We compared the frequency counts, familiarity ratings, neighborhood density, and neighborhood frequency of the interlopers and the targets. Information on 8 of the 18 interlopers was available from the same lexical database used to create the stimulus set. Independent t-tests found no difference in the familiarity of the targets and interlopers (t (1,14) < 1), no difference in the frequency counts (from Kucera & Francis, 1967) of the targets and interlopers (t (1,14) < 1), and no difference in the neighborhood frequency counts of the targets and interlopers (t (1,14) = 2.24, p = .15). As in experiment 1, a chi-square analysis was also done, but no significant differences were found (p > .10).

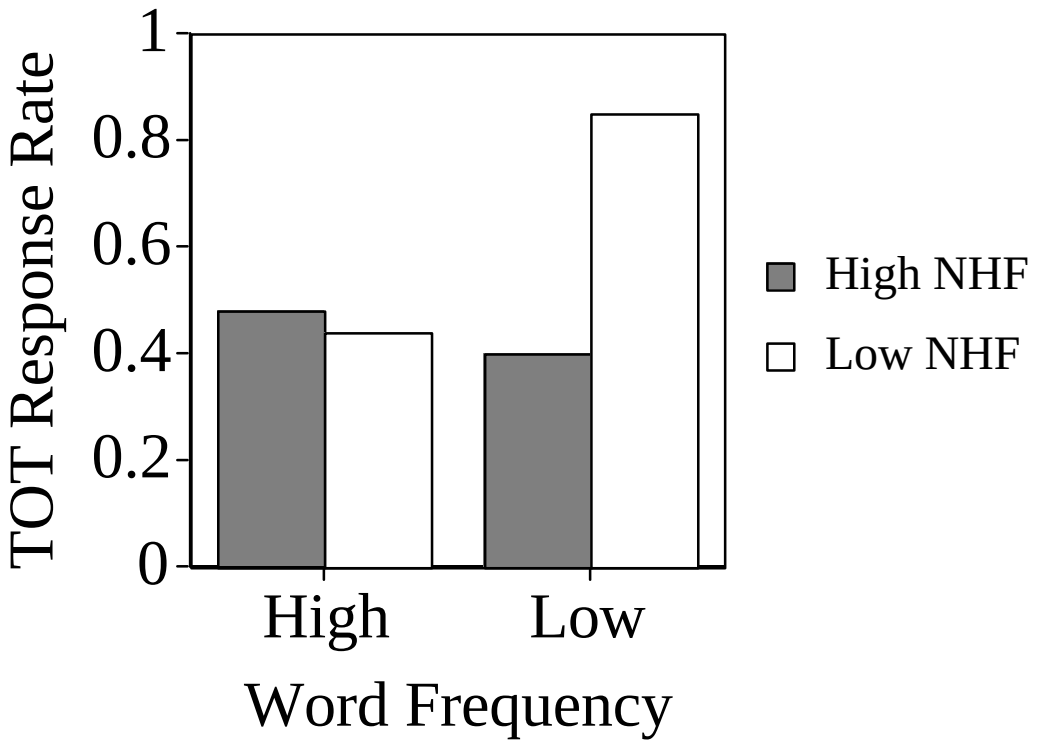


Figure 3. The number of TOTs for old adults as a function of word frequency and neighborhood frequency.

However, a significant difference was found in the neighborhood density values for the interlopers (mean = 18.0 neighbors) and the targets (mean = 29.6 neighbors; t (1,14) = 2.81, p < .05). This finding replicates the results of Harley and Bown (1998) and of Experiment 1, indicating that interlopers tended to be from sparser neighborhoods than their targets. Only one of the interlopers given by an older adult was a word with two syllables, and therefore would naturally have few similar sounding words (Pisoni, Nusbaum, Luce & Slowiaczek, 1985). The difference in neighborhood density between the interlopers and the targets was still significant (t (1,12) = 2.28, p < .05) when this item (and the related target word) was excluded

from the analysis. As in Experiment 1, the difference in neighborhood density between the targets and the interlopers may be due to a conscious search strategy adopted to resolve the TOT. That is, participants become consciously aware that they are having difficulty retrieving a word from the lexicon, decide the item must be "unusual" in some way, and search out "lexical hermits," or words with very few phonological neighbors. Adopting such conscious search strategies may mask natural biases used to select word-forms in normal processing.

Reports of Partial Information for Older Adults

Older participants were asked to report any partial information regarding the word they were attempting to retrieve. Participants correctly reported the number of syllables in the target word 8 out of 9 times, reported information regarding the first letter of the target word only once (and did so correctly), and did not report any information regarding the last letter of the target word. This result is interesting given the fact that younger adults always got the last letter correct.

Age differences in Total TOT states

To examine age-related differences in TOTs, we combined the TOT data from Experiments 1 and 2. A main effect of age was not significant by participants (F_1 (1,45) = 1.78, p = .19), but was significant by items (F_2 (1,224) = 9.41, p < .01). There was a tendency for older adults (mean = .54) to report more TOT responses than younger adults (mean = .32). This finding replicates the results of Burke et al. (1991). There was a marginally significant main effect of neighborhood density by participants (F_1 (1,45) = 4.08, p = .05) but not by items (F_2 (1,224) = 2.56, p = .11). There was a tendency for more TOT states to be reported for words with sparse neighborhoods (mean = .50) than for words with dense neighborhoods (mean = .36). A main effect of word frequency was found (F_1 (1,45) = 12.63, p < .001; F_2 (1,224) = 6.42, p < .05), such that more TOT states were reported for low frequency words (mean = .53) than for high frequency words (mean = .33).

A marginally significant interaction was found between age and neighborhood density by participants (F_1 (1,45) = 4.08, p = .05) but not by items (F_2 (1,224) = 2.56, p = .11). There was a tendency for more TOTs to be reported for words with sparse neighborhoods than for words with dense neighborhoods for young adults, and no difference between the two conditions for older adults.

A marginally significant interaction was found between age and neighborhood frequency (F_1 (1,45) = 4.04, p = .05; F_2 (1,224) = 3.49, p = .06). Older adults tended to report more TOTs for words that had low neighborhood frequency (mean = .64) than for words with high neighborhood frequency (mean = .43). This trend was reversed for younger adults; more TOTs were reported for words that had high neighborhood frequency (mean = .36) than for words with low neighborhood frequency (mean = .27).

A significant interaction was found between neighborhood density and word frequency (F_1 (1,45) = 5.62, p < .05; F_2 (1,224) = 4.55, p < .05). Figure 4 displays TOT response rates (combined across older and younger adults) as a function of word frequency and neighborhood density. As shown in the Figure, neighborhood density did not significantly influence the probability of TOTs for high frequency words. In contrast, for low frequency words significantly more TOTs were observed for words from sparse neighborhoods (mean = .68) than for words from dense neighborhoods (mean = .38; F (1,45) = 8.68, p < .01).

The interaction between neighborhood frequency and word frequency was significant by participants (F_1 (1,45) = 4.64, $p < .05$), and marginally significant by items (F_2 (1, 224) = 2.56, $p = .11$). There was a tendency for more TOTs to be reported for low frequency words than for high frequency words. This difference was greater for words with low neighborhood frequency than for words with high neighborhood frequency.

The interaction between neighborhood frequency, word frequency, and age was also significant by participants (F_1 (1,45) = 4.64, $p < .05$), and marginally significant by items (*F2* (1, 224) = 2.56, $p = .11$). Again, there was a tendency for more TOTs to be reported for low frequency words than high frequency words. For younger adults, this difference was approximately equal for words with both high and low neighborhood-frequency. However, for older adults the difference between the number of TOTs reported for high and low frequency words was much greater for words with low neighborhood frequency than high neighborhood frequency.

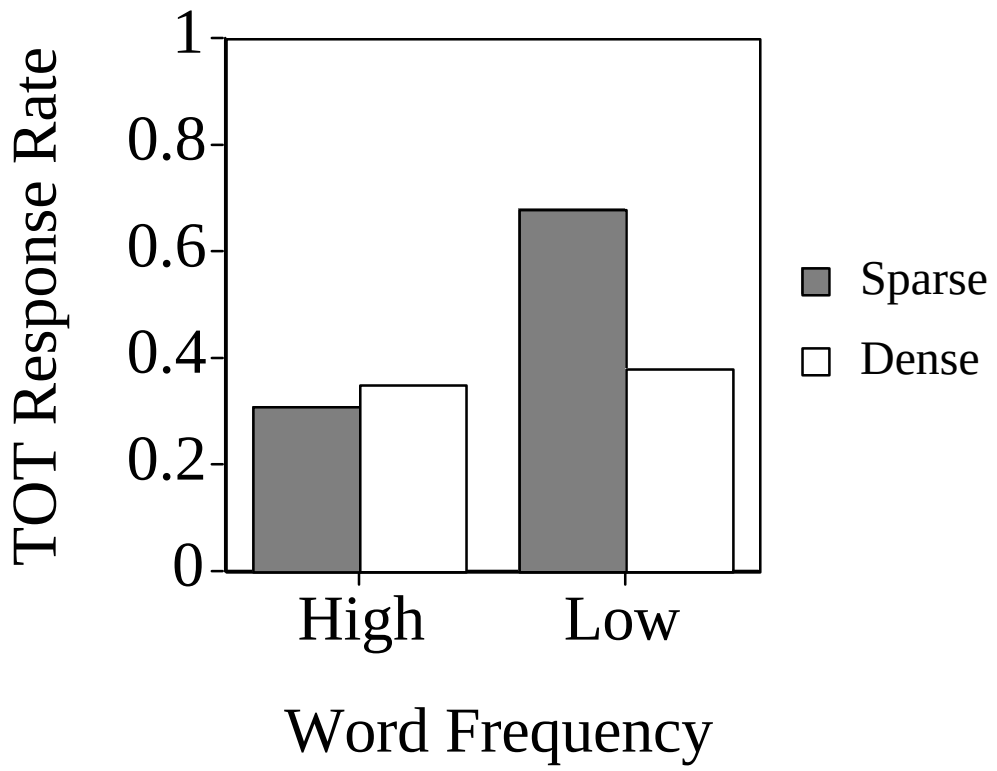


Figure 4. Overall mean number of TOT responses (younger and older adults combined) as a function of word frequency and neighborhood density.

Finally, an interaction was found between neighborhood density, neighborhood frequency, and age (F_1 (1,45) = 5.28, $p < .05$). This interaction was marginally significant by items (F_2 (1, 224) = 3.01, $p =$

.08). The trend shows that younger adults consistently reported more TOTs for words with sparse neighborhoods than for words with dense neighborhoods, regardless of neighborhood frequency. In contrast, older adults tended to report more TOTs for words with sparse neighborhoods only for words with low neighborhood frequency. This tendency was slightly reversed in older adults for words with high neighborhood frequency.

Discussion

The results of Experiment 2 replicate and extend findings of Burke et al. (1991; see also Burke & Laver, 1990; MacKay & Burke, 1990; and Rastle & Burke, 1996). For example, the current study found that more TOT states were elicited from older adults than from younger adults. These results are consistent with the transmission deficit account of TOT states (e.g., Bock & Levelt, 1994; Burke et al., 1991; Jescheniak & Levelt, 1994) within the NST (Burke et al., 1991). With age, the connections between nodes weaken, decreasing the rate and amount of priming transmitted between nodes. The decrease in the amount of priming between nodes in older adults results in the unsuccessful activation and retrieval of information associated with those nodes.

The "transmission deficit" hypothesis is also supported by the difference in availability of partial (phonological) information in younger and older adults. In the current experiment, older adults reported less partial information than did the younger adults. This observation is also consistent with the findings of Maylor (1990) who found that older adults had decreased availability of partial (phonological) information.

Of greater importance is the differential influence of neighborhood density with age. The results of Experiment 1, examining the influence of neighborhood density on TOT states in young adults, replicated the findings of Harley and Bown (1998): More TOT states were elicited for words with few neighbors than for words with many neighbors. However, older adults did not evidence significant main effects of neighborhood density as predicted. Instead, the process of lexicalization in older adults seemed to be differentially influenced by neighborhood frequency. Specifically, more TOT states were elicited in older adults for words with low neighborhood frequency and either sparse neighborhoods or low word frequency. No such interactions were observed in younger adults.

Although our predictions regarding the interaction between neighborhood density and aging were not observed in the results of Experiment 2, the findings of the present experiment can still be accounted for within NST (Burke et al., 1991; Burke & Laver, 1990; MacKay, 1987; MacKay & Burke, 1990). Within NST the efficiency of transmitting priming between nodes decreases with aging. Furthermore, the ability to access phonological information (i.e., phonological nodes) specifically decreases with age (Burke et al., 1991; Maylor, 1990). This accounts for the overall increase in the number of TOTs reported by older adults compared to younger adults.

Age-related transmission deficits may not be fully counter-acted by the stronger connections that arise from differential exposure rates as a function of age to words with dense neighborhoods versus words with sparse neighborhoods as we had initially predicted. Although phonemes shared by many words (i.e., phonemes in a word with a dense neighborhood) generally are "protected" from TOTs and other speech errors in younger adults, older adults activate and retrieve phonological information less successfully overall than younger adults. To compensate for the transmission deficits and weaker connections that result from aging, MacKay and Burke (1990) predicted that older adults rely on other sources of priming. Specifically they state that:

> "...age differences are likely to be pronounced when a node critical to a task receives priming from only a single source or connection within the network. Age-linked transmission deficits are very likely to affect performance in such a task because no other sources of priming will be able to offset the reduced priming across that critical connection." (MacKay & Burke, 1990, pg. 251).

One of the other sources of priming that older adults may use to compensate for age-related transmission deficits that affect a single source of priming is the frequency-based priming transmitted by the neighboring word nodes. A word with low neighborhood frequency has neighboring words with a mean word frequency that is low, whereas a word with high neighborhood frequency has neighboring words with a mean word frequency that is high. Neighboring words, like the target word, prime the phonological nodes they are connected to as a function of their frequency via the symmetrical connections in NST. Connections that are frequently primed are maintained and become stronger over time, whereas less frequently primed connections become weaker over time. Phonological nodes connected to a word with neighbors that have low frequency would, therefore, receive less priming from those neighboring words than nodes connected to a word with neighbors that have high frequency. The interaction of neighborhood frequency with word frequency and with neighborhood density in older but not younger adults supports the "single-source factor" hypothesis proposed by MacKay and Burke (1990) and the age-related transmission deficit hypothesis. That is, older adults will compensate for the decrements in priming from one source or connection by relying on additional sources or connections for priming. The current results suggest that older adults may be sensitive to neighborhood frequency as an alternative source of priming in speech production, verifying the predictions of MacKay and Burke (1990) regarding the "single-source factor."

General Discussion

The results from Experiment 1 replicated and extended the results of Harley and Bown (1998): More TOT states were elicited (from young adults) for words with sparse neighborhoods than for words with dense neighborhoods. More importantly, we demonstrated this effect independent of word length, a factor that was not controlled for in Harley and Bown (1998), and with stimuli that varied in neighborhood density based on a phonological rather than orthographic similarity metric. Furthermore, we manipulated another variable—neighborhood frequency—that Harley and Bown (1998) neither manipulated nor controlled. Although no significant effects of neighborhood frequency were observed in young adults, demonstrable influences of this variable were observed in Experiment 2 among older adults.

The results of the current set of experiments provide crucial insight into the process of lexicalization and how it may change over the life span. Specifically, older, but not younger adults showed a significant influence of neighborhood frequency on the number of TOT states elicited. The influences of a single source of priming interacting with age-related transmission deficits led MacKay and Burke (1990) to predict that older adults may adopt additional sources of priming to compensate for their less efficient processing. The results of Experiment 2 verify the prediction that older adults may adopt a form of compensation in order to maintain "normal" processing. That is, because of age-related transmission deficits, older adults are unable to obtain sufficient priming from neighborhood density and are therefore forced to rely on additional sources of priming such as those available from neighborhood frequency. The absence of an effect of neighborhood density for older adults may indicate a shift in weighting functions for these two sources of priming, with neighborhood frequency having greater weights for older than for younger adults. Although clearly speculative at this point, the proposal of age-related shifts in weighting functions for neighborhood density and neighborhood frequency are consistent with the current findings.

More definitive conclusions will need to await further research on factors that can influence the weighting functions for different sources of priming.

In addition to "adopting" alternative sources of priming, older adults (and to some extent, younger adults) may adopt conscious search strategies to resolve TOT states. Indeed, Reason and Lucas (1984) found that the only internal strategy used with any success to resolve TOT states was the generation of similar sounding words. Such conscious search-strategies through the lexicon may account for our finding (also observed by Harley & Bown, 1998) of interlopers having sparser neighborhood density, or fewer similar sounding words, than the targets. One might reason that if one is having a difficult time retrieving a lexical item (i.e., is in a TOT state), it must be because the item is extremely "rare," or "unusual" in some way. This may result in a conscious search strategy for unique items in memory. Indeed, Burke et al. (1991) found that TOTs commonly occurred for proper names, which are highly unique items. If a similar search strategy were adopted for words, "lexical hermits," or words that have very few phonological neighbors, may be retrieved, accounting for the finding that interlopers have sparser neighborhood density than targets. Harley and Bown (1998) provide a similar explanation for their findings. Such conscious search strategies may also over-ride biases that arise as a result of automatic processes, such as frequency and word-length biases.

Furthermore, the results from the current set of experiments adds to a growing body of literature suggesting that "interlopers," phonologically related words that continually come to mind instead of the intended target word, support rather than block the retrieval of the target word (Harley & Bown, 1998; Meyer & Bock, 1992; Perfect & Hanley, 1992). If interlopers were to block the retrieval of words, one would predict that a word with many similar sounding words, a word with a dense neighborhood, would elicit more TOT states than a word with few similar sound words, a word with a sparse neighborhood. Exactly the opposite was observed (see also Harley & Bown, 1998), suggesting that similar sounding words support the retrieval of the target word. A word with a dense neighborhood receives more support than a word with a sparse neighborhood and, therefore, will most likely be completely retrieved.

One of the conclusions from our data—the number of similar sounding words (i.e., neighborhood density) acts to support rather than block the retrieval of word forms from lexical memory—coincides with the conclusions of Harley and Bown (1998). However, our explanation for why this is observed differs from the explanation posited by Harley and Bown. We (along with Burke et al., 1991; MacKay & Burke, 1990) propose that TOT states result from insufficient activation at the interface between word-forms and sub-lexical representations. In contrast, Harley and Bown propose that insufficient feedback between the lemma level (which represents semantic and/or syntactic information) and the lexeme level (which represents the phonological word-form) is responsible for a TOT state.

Although the data from the present experiments (and from Harley & Bown, 1998) can not directly distinguish between these two accounts, we propose that TOTs are due to insufficient feedback between word-forms and sub-lexical representations. Identifying the locus of TOTs at the interface between word-forms and sub-lexical representations allows us to easily account for our results in an extant model of cognitive processing, namely NST (Burke et al. 1991, MacKay, 1987), without having to posit modifications or additional assumptions to a model of speech production. In postulating that the locus of TOTs is at the interface between lemmas and lexemes, Harley and Bown may have to propose an additional mechanism to account for the decreased ability of older adults to specifically retrieve phonological information as observed in Experiment 2 (see also Burke et al., 1991; Maylor, 1990). It is unclear how such a difference across the life span could be accounted for if the locus of TOTs is at the interface between lemmas and lexemes. Further empirical work must ultimately establish the locus of TOTs.

The results from the current experiment also address whether the speech production system could best be described as a modular or an interactive system. We accounted for our results in the context of an interactive model of cognitive processing, the Node Structure Theory (Burke et al. 1991, MacKay, 1987). It is not entirely clear how a modular system that posits independent processing could account for the current results. More TOT states for words with few rather than many similar sounding words seems counter-intuitive to a strict modular account of speech production. Such accounts argue that processing at one level must be completed before processing at another level can proceed (e.g., Garrett, 1976; Levelt, 1989). Competition among many similar sounding words at the word-form level would seem to slow rather than speed processing, much as it does in spoken word recognition (e.g., Luce & Pisoni, 1998). Furthermore, it is unclear how multiple candidates in a modular model might be activated at the word-form level given that many models of speech production, including interactive models, do not posit lateral connections between word-forms (e.g., Dell, 1986; Garrett, 1976; Harley, 1993; Levelt, 1989). It is dubious how multiple word-form candidates might be activated in a modular system if a lemma, or semantic representation, activates only the lexeme, or word-form representation, connected to it without allowing feedback from phonological units as in an interactive model of speech production.

A critical finding from Experiment 2 is that we demonstrated that word frequency and neighborhood density are not the only variables that influence lexicalization (see Harley & Bown, 1998). Rather, we have demonstrated that several variables—word frequency, neighborhood density, and neighborhood frequency—affect lexicalization differently at various points in the life span. Additional work on lexicalization must be done to further examine the locus of TOTs and how the processes that operate during speech production change across the life span. All of the results from the current experiments are consistent with a well-specified and tested model of cognitive processing, namely the Node Structure Theory (MacKay, 1987; MacKay & Burke, 1990). Our results also verified a prediction of NST regarding the interaction of priming from a single source and age-related transmission deficits, suggesting that NST is a useful model for understanding the changes in the speech production system as a function of age. Additional work in the context of NST may provide insight into how changes in speech production may differ from the changes that occur in speech perception across the life span and shed light on the differences between the two cognitive systems.

References

Bard, E. & Shillcock, R. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing* (pp. 235-275). Hove, Sussex: Erlbaum.

Brown, A.S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, **109**, 204-223.

Brown, R. & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, **5**, 325-337.

Burke, D.M. & Laver, G.D. (1990). Aging and word retrieval: Selective age deficits in language. In E.A. Lovelace (ed.) *Aging and Cognition: Mental Processes, Self-Awareness and Interventions*. (pp. 281-300) North Holland: Elsevier Science Publishers.

Burke, D.M., MacKay, D.G., Worthley, J.S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, **30**, 542-579.

Caramazza, A. & Miozzo,M. (1998). More is not always better: a response to Roelofs, Meyer, and Levelt. *Cognition*, **69**, 231-241.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.) *Attention and Performance, VI*, pp. 535-555. London: Academic Press.

Crystal, D. (1992). *An Encyclopedic Dictionary of Language and Languages*. London: Blackwell.

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283-321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, **27**, 124-142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, **5**, 313-349.

Garrett, M.F. (1976). Syntactic processes in sentence production. In R.Wales & E. Walker (Eds.) *New approaches to language mechanisms* (pp. 231-256). Amsterdam: North-Holland.

Goldinger, S.D. and Summers, V.W. (1989) Lexical neighborhood in speech production: A first report. In *Research on Speech Perception Progress Report No. 15* Bloomington, IN: Speech Research Laboratory, Indiana University.

Harley, T.A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science*, **8**, 191-219.

Harley, T.A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, **8**, 291-309.

Harley, T.A. & Bown, H.E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, **89**, 151-174.

Harley, T.A. & MacAndrew, S.B.G. (1992). Interactive models of lexicalization: Some constraints from speech error, picture naming, and neuropsychological data. In J. Levy, D. Bairaktaris, J. Bullinaria & D. Cairns (Eds.), *Connectionist Models of Memory and Language*, pp. 378-383. London; UCL Press.

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, & Cognition*, **20**, 824-843.

Jones, G.V. (1989). Back to Woodworth: Role of interlopers in the tip of the tongue phenomenon. *Memory & Cognition*, **17**, 69-76.

Jones, G.V. & Langford, S. (1987). Phonological blocking in the tip of the tongue state. *Cognition*, **25**, 115-122.

Kucera, H. & Francis, W.N. (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.

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, **19**, 1-36.

MacKay, D.G. & Burke, D.M. (1990). Cognition and aging: A theory of new learning and the use of old connections. In T.M. Hess (ed.) *Aging and Cognition: Knowledge Organization and Utilization*. (pp. 213-263) North-Holland: Elsevier Science Publishers.

MacKay, D.G. (1987) *The Organization of Perception and Action: A Theory for Language and other Cognitive Skills*. New York: Springer-Verlag.

Maylor, E. A. (1990). Age, blocking and the tip of the tongue state. *British Journal of Psychology*, **81**, 123-134.

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.

Meyer, A.S. & Bock, J.K. (1992). The tip of the tongue phenomenon: Blocking or partial activation? *Memory and Cognition*, **20**, 715-726.

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*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Oldfield, R.C. & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Psychology*, **17**, 273-281.

Perfect, T.J. & Hanley, J.R. (1992). The tip of the tongue phenomenon: Do experimenter-presented interlopers have any effect? *Cognition*, **45**, 55-75.

Pisoni, D.B. Nusbaum, H.C., Luce, P.A., & Slowiaczek (1985) Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, **4**, 75-95.

Rastle, K.G. & Burke, D.M. (1996). Priming the tip of the tongue: Effects of prior processing on word retrieval in young and older adults. *Journal of Memory and Language*, **35**, 586-605.

Reason, J. & Lucas, D. (1984). Using cognitive diaries to investigate naturally occurring memory blocks. In J.E. Harris & P.E. Morris (Eds.) *Everyday Memory Actions and Absent-Mindedness*. (pp. 53-70). London: Academic Press.

Roelofs, A., Meyer, A.S., & Levelt, W.J.M. (1998). A case for the lemma/lexeme distinction in models of speaking: comment on Caramazza and Miozzo (1997). *Cognition*, **69**, 219-230.

Salthouse, T.A. (1985). *A theory of cognitive aging*. Amsterdam: North-Holland.

Sommers, M.S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging*, **11**, 333-341.

Sommers, M.S. & Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging*, **14**, 458-472.

Stemberger, J.P. (1984) Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology*, **1**, 281-313.

Stemberger, J.P. (1985) The reliability and replicability of speech error data: A comparison with experimentally induced errors. In *Research on Speech Perception Progress Report No. 11*, Bloomington, IN: Speech Research Laboratory, Indiana University.

Stemberger, J.P. & MacWhinney, B. (1986). Frequency and the lexical storage or regularly inflected forms. *Memory and Cognition*, **14**, 17-26.

Vitevitch, M.S. (1997a). Tongue twisters reveal neighborhood density effects in speech production. In *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Vitevitch, M.S. (1997b). The neighborhood characteristics of malapropisms. *Language and Speech*, **40**, 211-228.

Vitevitch, M.S., Luce, P.A., Pisoni, D.B., & Auer, E.T. (1999). Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language*, **68**, 306-311.

Webster's New Collegiate Dictionary (1979). Springfield, MA: G. & C. Merriam Company.

Woodworth, R.S. (1929). *Psychology* (2nd rev. ed.) New York: Holt.

Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. In *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Yaniv, I. & Meyer, D.E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13**, 187-205.

Zipf, G.K. (1965). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, Mass., M. I. T. Press.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Talker-Specific Effects in Recognition Memory for Sentences[1]

Kipp McMichael

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by a grant from the NIH-NIDCD Research Grant DC00111. For his detailed, insightful, and exhaustive editing of this paper, I thank David Pisoni. Special thanks also goes to Luis Hernandez for his technical assistance during the completion of this research.

Talker Specific Effects in Recognition Memory for Sentences

Abstract. Speech perception is usually considered to be the process by which listeners change spoken sounds into a meaningful string of words and ideas. Yet information about the talker's age, gender, and socioeconomic status is also carried in parallel with the symbolic content of the linguistic message. The traditional view of speech perception posits that generic linguistic units such as phonemes and words are recovered from the speech signal while "extra-linguistic" talker-specific attributes such as gender, dialect, and emotional state are filtered from the signal during perceptual processing and before encoding into memory. The last decade has seen a dramatic increase in speech research specifically designed to measure the effects of stimulus variability. Most of this earlier work has used spoken word recognition to study speech processing, encoding and recall. The present series of four recognition memory experiments utilized sentences as stimuli to examine the process of sentence encoding and later recognition. Subjects listened to a study phase of 40 sentences spoken by 5 male and 5 female talkers and then completed a recognition test phase of 80 sentence which were a mixture of study phase and unstudied talkers and sentences. Results of all four experiments revealed significant effects of voice on sentence judgment accuracy and discrimination scores. These results call into question current theories and models of speech processing which posit preliminary normalization or other variability reduction. While no current theory models speech recognition outside a framework incorporating a formalized, idealized phonemic or segmental stage, it should no longer be taken for granted that speech recognition is a normalizing, abstracting process. Variation in the speech signal may be as important to our understanding and encoding of speech as the regularity of the phonemes and the words they comprise.

Introduction: The Speech Signal

Although usually recognized only as an acoustic phenomenon, the speech signal is actually a dynamic complex of interacting sources of information transmitted through multiple sensory modalities (Summerfield, 1983). Sign language and lip-reading provide examples of alternate modalities that can encode and express language. While these communication modes are usually associated with the deaf who cannot otherwise perceive spoken speech, research has found that lip-reading is also a basic part of normal-hearing listeners' speech perception processes. So basic is lip-reading, in fact, that Sumby and Pollack (1954) found that viewing the face of the talker in a speech-in-noise identification task was equivalent, for increasing accuracy, to a 15dB increase in the signal-to-noise ratio for increasing identification accuracy. McGurk and MacDonald (1976) reported that phonemic perception relied equally on visual and auditory information when subjects were asked to identify phonemes from an audio-video. The sense of touch can even be utilized in speech communication with a glottal pulse train that signals the presence or absence of voicing during articulation (Summerfield, 1987). Deaf listeners who place their hands over the larynx of talkers exploit the same principle. Clearly, then, the speech signal can be and is transmitted through multiple means even for normal-hearing listeners. But exactly what kind of information is being transmitted?

Speech perception is usually considered to be the process by which listeners turn speech into a meaningful string of words and ideas (Goldinger, Pisoni, & Luce, 1996). Yet there is far more to be found in the speech signal than simply a sequence of words or phrases. Information about the age, gender, and the emotional state of the talker as well as information about the talker's background and socioeconomic status is carried in the speech signal. Recent investigations have found that listeners do extract and encode this type of information (Goldinger et al., 1996; Labov, 1963). The traditional view of speech

communication partitions the information carried in the speech signal into two parallel streams: linguistic information such as phonemes, words, and phrases and extra-linguistic or "indexical" information such as physical or emotional state, gender, dialect, etc (Ladefoged & Broadbent, 1957; Pisoni, 1993). This division is based on the meta-theoretical notion that speech perception operates only on linguistic information and hence extra-linguistic information does not contribute to the basic stages of speech processing. Of course, knowledge about the talker's gender and emotional state could contribute to our understanding of speech, but the contribution of this information has traditionally been viewed as part of the context that contributes to our understanding after speech processing has occurred (Ladefoged & Broadbent, 1957; Abercrombie, 1967).

The linguistic versus extra-linguistic division is problematic mainly because the information in the speech signal does not seem to fall naturally into this dichotomy at all. While the emotional state or age of the talker may have little affect on the listeners' phonemic perception, it is obvious from everyday experience that regional dialect and accent can greatly affect the acoustic realization of speech sounds. Moreover, the talker's age and gender also dramatically affect the acoustic realization of phonemes (Peterson & Barney, 1952; Liljencrants & Lindblom, 1972). The effects of such extra-linguistic factors as gender and age have important consequences for the realization of linguistic information in the speech signal and thus may not be so extra-linguistic after all. Thus, it appears that the linguistic vs. extra-linguistic distinction is an expedient assumption of speech processing and language theories rather than reflecting an actual dichotomy in the speech processing system itself.

Processing the Speech Signal

We can understand the motivation behind the linguistic vs. extra-linguistic division of speech signal information by considering the assumptions of traditional models of speech processing. From its beginnings, speech research was heavily influenced by formal linguistics and hence viewed speech perception as the process that recovers phonemes and words from the speech signal (Halle, 1956; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Pisoni & Lively, 1995). The notion of spoken language as simply an acoustic realization of written language had major consequences for speech research. In particular, researchers looked for the counterparts to the serially ordered, formal and discrete letters of written speech (Licklider, 1952; Halle 1956). Even when it became clear that spoken language was anything but a serial and discrete transmission of information, speech researchers maintained the notion that the process of speech perception recovered and identified formal linguistic units such as phonemes or segments (Liberman et al., 1957; Liljencrants & Lindblom, 1972). Thus, the speech signal was conveniently divided up into two sources of information: abstract linguistic information which presumably contributed to the basic process of phoneme identification and word recognition and extra-linguistic information such as the talker's gender or accent which was simply filtered out from the speech signal during initial processing (Halle, 1956; Brown 1990).

The notion of variable, extra-linguistic information as noise which is removed from the speech signal underlies both traditional feature-based phoneme recognition models as well as contemporary kinematic models which view speech perception from an ecological standpoint (Pisoni & Lively, 1995; Klatt 1989; Gaver, 1993; Fowler, 1986). According to both of these approaches, the speech signal undergoes a series of processes in order to extract the symbolic phoneme, segment, or articulation movement (Studdert-Kennedy, 1970). The final result is a translation of the speech signal into an abstract, symbolic idealized string of formal linguistic units which are devoid of such information as talker gender or age (Brown, 1990). Again, these models do not necessarily require that extra-linguistic information be removed altogether from the cognitive system of the listener, but rather, this information is processed by non-linguistic systems.

Encoding the Speech Signal

The result of the traditional view of speech processing is an abstracted, formal representation of speech. It is no surprise, then, that the memory structures these views posit for the encoding and storage of speech lack both variability and extra-linguistic content. From the first conceptualization of the mental lexicon as mental dictionary, it was assumed that the memory representation of speech was formal, abstract and generic in nature (Oldfield, 1966). The traditional models fit well with the then accepted prototype or abstractionist view of memory as a storage of idealized memory components. Hence the mental lexicon and memory for other linguistic units were presumed to be prototypical representations of abstract, formal linguistic units. Again, however, no researcher claimed that gender, age or talker identity had no effect on the listener – obviously we can recognize voices and gender - it was simply ignored in their models of speech perception. Thus, whereas our memories for people, places, and events utilize a rich mental representation, our memories for speech events are a simplified, abstracted idealized representation divorced from the age, gender, or emotional state of the persons who spoke.

After traditional accounts of speech processing and storage are employed, we are left with a lexicon which is little more than Webster's Dictionary encoded neurally – with similarly abstracted and formalized organization (Oldfield, 1966). However, a growing body of new research casts doubt on both the prototypical model of the mental lexicon as well as the formal linguistic notions that have influenced speech-processing research (Pisoni, 1990).

Measuring Speech Perception and Encoding

Challenges to the notion of filtering, abstractionist speech processing were uncovered, but unrecognized, early in the history of speech research. As early as 1955, Peters found that identification of words suffered when multiple talkers spoke the stimulus lists, even though individually the talkers could be easily understood (Peters, 1955). In 1957, Creelman also found that identification of words in noise suffered when stimulus lists were spoken by multiple talkers (Creelman, 1957). These findings were subsequently replicated by Mullennix, Pisoni, and Martin (1989). If the basic process of word recognition begins by filtering the speech signal of extra-linguistic information like talker identity, gender and dialect, we are unable to explain this finding.

Vowel classification experiments found that speakers were able to accurately classify vowels produced by children even when the acoustic characteristics of these vowels were highly dissimilar to adult productions (Gerstman, 1968). Again, unless the extra-linguistic knowledge of a talker's age is somehow utilized during speech perception, it would seem that listeners would be unable to understand the vowels produced by children unless they used qualitatively different criteria for acoustic-phonemic categorization. Though a robust and repeated finding in early speech research (Miller, 1951), the effects of different forms of variability on speech processing were largely precluded from subsequent experiments due to their basic design. The complexity and cost of stimulus creation as well as the general notion that variability was merely a nuisance factor resulted in most experiments utilizing small stimulus sets and even smaller talker pools. Thus, most of speech research until the last two decades was quite simply unable to assess the effects of stimulus variability and this trend had major effects on research and theory during this time.

As the small but important lineage of speech variability research became more widely known, researchers began designing experiments that explicitly addressed these topics. Kuhl and Miller (1982) found that pre-linguistic infants could accurately discriminate the vowels produced by three different talkers. It appears, then, that even infants are able to process and utilize the variability inherent in different talkers' speech. In word recognition paradigms it was found that, like Creelman's (1957) results,

multiple talkers hurt performance (Summerfield & Haggard, 1983). Again this result runs counter to what would be expected from an abstractionist perspective.

The last decade has seen a dramatic increase in speech research specifically designed to measure the effects of stimulus variability. In a serial recall task, Goldinger et al. (1991) found that presentation rate interacted with the number of talkers in the stimulus ensemble to affect subject performance. At slower presentation rates, subjects were better at recalling lists of words spoken by multiple talkers. Multiple talkers hurt performance when presentation rate was faster, however. These results seem to indicate that variable voice information was utilized at slower presentation rates to help recall and thus must have been present in the early neural representation of the words. Using a continuous recall memory task, Palmeri, Goldinger, and Pisoni (1993) found that spoken words were better recognized at test if they were repeated by the same talkers as the original presentation. Moreover, Palmeri et al. also found that performance was worst for words repeated in a new voice from a different gender. Palmeri et al.'s experiment is also noteworthy because subjects were never explicitly instructed to attend to gender or talker identity during the study phase in any of the experiments. These results are simply unexplainable in a framework of abstractionist speech processing which posits that extra-linguistic information such as a talker's identity and gender are stripped from the speech signal before encoding.

In a perceptual identification experiment, Goldinger (1992) found that implicit memory for talker voice attributes was retained in memory long after perceptual analysis was complete. Subjects returned for follow-up tests 5 minutes, one day and one week after initial word identification trails. As long as one week after the original test, subjects showed accuracy improvements for word repetitions in the same voice as they heard during initial testing and accuracy decrements for difference-voice repetitions. Moreover, these effects were not significantly different from the effects observed in the original testing sessions.

Speaking rate has also been found to influence recall in a manner that is similar to talker identity. Nygaard, Sommers, and Pisoni (1992b) found that recall was better for word lists spoken at the same rate at test as at study. In perceptual learning experiments, talker variability has been found to have a beneficial affect on acquiring and retaining phonetic contrast perception for Japanese listeners (Logan, Lively & Pisoni, 1990; Lively, Pisoni, Yamada, Tohkura & Yamada, 1992). Lively et al. (1992), in particular, found that listeners could better retain their perception of the English /l/ and /r/ contrast when they were exposed to a large corpus of stimuli spoken by many talkers.

Even more recently, Remez, Fellowes, and Rubin (1997) found that subject performance in a sinewave sentence identification task indicates that both talker and phoneme identification rely on the same basic processes. In one condition of their study, subjects listened to naturally produced sentences and then had to judge which of two synthetic sinewave sentences had been produced from the original sentence. In another condition, subjects listened to sinewave processed sentences spoken by familiar voices and were asked to identify the talker. The pattern of results across these and another conditions led Remez et al. to conclude that subjects used the same information in processing sinewave sentences to recover both phonemes and talker identity. Thus the distinction between linguistic and extra-linguistic information does not clearly hold for speech recognition processes as basic as phonemic perception.

In summary, we find from a series of studies that variability has important effects for speech perception that cannot be explained from a traditional abstractionist view of speech processing. Remez et al.'s (1997) research is particularly important because it indicates that this variability plays an important role in the earliest stages of speech processing. In addition, talker variability can be encoded in both an incidental and implicit way with neither the awareness nor the intention of the listener (Peters, 1955; Creelman, 1957; Goldinger, Pisoni & Logan, 1991; Palmeri et al., 1993). For such information to be

encoded in memory would require a memory structure for speech events much more rich in information that the abstract, idealized, symbolic representations usually described. These results also require a speech recognition process which, rather than filtering variability from the signal processes, utilizes, and encodes signal variability. The present experiments were designed to further explore the nature of variability in speech processing of sentences.

Sentence Processing

While the role of variability in phoneme and word processing has been examined recently by researchers, little research has been done with spoken sentences. Rather than being the result of a purposeful avoidance of sentence processing experiments, this situation is more likely a result of the assumption that sentence recognition is simply the concatenation of word recognition processes. It is clear from an acoustic analysis of speech, however, that isolated words and words in sentences show significant differences in physical realization (Klatt, 1986; Pisoni & Luce, 1987). In a gating study by Salasso and Pisoni (1985) they found that words removed from their acoustic sentence context were accurately identified only 50% of the time. Clearly, there are important differences between isolated and spoken words that may have implications for the role of variability in sentence processing. If sentences are encoded or processed differently from isolated words, then the effects observed in isolated word research may not be repeated. Due to their length and increased information content and presentation rate, sentences may also require different strategies in processing which may obscure the effects of stimulus variability. These questions concerning the similarities between isolated word and sentence recognition processes can only be addressed by sentence-based research.

When sentences have been used in speech perception experiments in the past, much as the case with single word or phoneme stimuli, few talkers recorded the sentences and the corpus itself was very small. Still, however, several important experiments have utilized sentences as stimuli with the research of Karl (1996) being one of the main motivations behind the present study.

Geiselman and Bellezza (1976) found that talker gender was incidentally encoded in memory for sentences in a recall task. In a similar set of experiments, Fisher and Cuervo (1983) examined memory for extra-linguistic attributes of sentences as they related to sentence comprehension. In their study, subjects were more likely to remember the gender of the speaker or the language of the sentence when this information was important to sentence comprehension.

While the earlier research of Geiselman and Fisher did uncover some effects of stimulus variability, a series of recall experiments conducted by Karl found no effects for talker variability (Karl & Pisoni, 1994; Karl, 1996). Across several variations on a basic free recall task in which subjects heard and then transcribed lists of sentences spoken by one or more talkers, Karl found no significant effects for talker variability on recall performance. In the experiment, sentences were presented in blocks spoken by multiple or single talkers and subjects were asked to recall the sentences by transcribing them in any order when a tone sounded after each block. Karl found that recall was not significantly affected by the number of talkers within each block. Karl concluded that the nature of a free recall task that emphasized rehearsal might not be suitable for uncovering effects of surface features such as talker identity. Karl's suggestion was the elaboration and extraction that his experiment may have encouraged in the subject obscured the surface information present in the stimuli. Thus we find a mixed message from previous sentence processing research examining the role of variability in speech processing. We also find motivation for the present study.

Recognition Memory

Considering Karl's conclusions that a serial recall paradigm is inappropriate for research involving talker-specific variability, we selected a discrete recognition memory paradigm for these experiments. The recognition memory design was also chosen because of the large body of isolated word and speech segment research carried out in this paradigm (Egan, 1948; Snodgrass & McClure, 1975; Squire, Shimamura & Graf, 1985; Henson, Rugg, Shallice, Josephs & Dolan, 1999).

In a discrete recognition memory experiment, the procedure is divided into two parts: a study phase and a test phase. During the study phase, subjects review a collection of stimuli and may additionally perform some specified task after stimulus presentation. Depending on the goal of the experiment, the study phase task may be as simple as controlling the pace of stimulus presentation or the task may require additional judgments or actions with regard to the stimuli. Though the stimuli in these and previous speech experiments are auditory speech samples, the recognition memory design can accommodate principally any set of stimuli such as pictures, objects, or visual speech (Kim, Andreasen, O'Leary, Wiser, Ponto, Boles, Watkins & Hichwa, 1999; Cornell, 1980).

Once the study phase has been completed subjects may perform an interpolated task before the test phase begins. Again, this task can be a distractor task to prevent rehearsal or a task designed to clear short-term memory. No interim task was performed in this series of experiments so subjects moved directly to the test phase.

The test phase, much like the study phase, consists of subjects reviewing a new collection of stimuli. The stimuli used during the test phase consist of some or all of the study phase stimuli mixed with additional new stimuli. The subjects' task in this phase of the experiment is to make a recognition judgment about the stimuli such as whether this stimulus was presented during the study phase (old stimuli) or not (new stimuli). The test phase stimuli are designed to resemble study phase stimuli so that performance in the test phase gives a measure of the discriminability of the new stimuli from the old stimuli as well as a general measure of memory for study phase items. In the present series of experiments signal-detection analyses were performed on response data.

Signal Detection Analysis

The responses to test phase stimuli in these experiments take the form of old (present at study) or new (not preset at study) judgments. Despite the fact that only two responses are possible, there are actually four different response types for these stimuli – two old and two new. Signal detection theory allows us to analyze and make sense of these four response types. The two old response categories are "hits" – correctly labeling an old sentence as old and "false alarms" – incorrectly labeling a new sentence as old. The two new response types correspond to correct new judgments applied to new sentences – "correct rejections" and incorrect new judgments applied to old sentences – misses.

Whereas a standard analysis of performance based on accuracy would lump both old judgments and both new judgments together, a signal detection analysis allows more information to be recovered from the response data. In addition to the scoring for hits, misses, false alarms, and correct rejections, two additional measures can be derived under signal detection analysis. D prime (d') is one of these measures computed by subtracting the false alarm rate from the hit rate. The function of d' is used to assess how well subjects can discriminate truly old items from new items. Without the additional insight d' can provide, a subject could respond old to all sentences and appear to have perfect accuracy for judging old sentences – even when the subject may not have been able to identify old sentences at all. Thus, d' gives

an indication not just of the percent correct, but also of the overall trend of accurate and inaccurate judgments and allows a better assessment of the effects on recognition performance.

A second derived measure is called beta (B) and is computed by normalizing the score for correct rejections. Beta is a measure of the underlying tendency for a given subject to make a specific response. Thus, a subject who responds old to all sentences would have a high bias for old responses or, conversely, would be very bad at making new judgments. These two independent measures provide important information about the underlying sensory and decision processes used by subjects during the experiment. For the current experiments the "signal" to be detected is the oldness or newness of a sentence. As the following discussion will show, signal detection analysis allows us to extract important information from an otherwise simple pattern of old and new responses in test phase data.

Experiment 1: Transcription at Study

Experiment 1 represents the prototype for the four experiments described in this paper. Both the corpus of stimuli and the materials described in depth here are identical across all four experiments.

Method

Subjects

Subjects were 33 undergraduate psychology students at Indiana University who were given class credit in an introductory course for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability. These requirements were not formally enforced - though each subject was informally assessed by the experimenter before participation. It should be noted, however, that subjects did complete individual data forms which corroborate the assumption that subjects followed the sign-up sheet guidelines.

Materials

All stimuli used in this experiment were taken from the Indiana Multi-Talker Speech Database (IMTSD). The IMTSD is a corpus of 100 Harvard sentences (Egan, 1948) recorded by 10 male and 10 female talkers for a total of 2,000 sentence tokens. Each sentence type consists of 5 content words and a variable number of function words. All sentences are meaningful, declarative or imperative statements such as, "These days a leg of chicken is a rare dish" or "Throw the box beside the parked truck." The Harvard sentences have a long history in speech research and so were used here despite their somewhat dated constructions. Though at times quaint in their grammar, the 100 sentences are easily understood and no subject reported a problem in comprehending a sentence. All of these stimuli were stored in the form of digital computer sound files sampled at 20Hz with 16-bit resolution. For a more complete discussion of the collection and creation of the IMTSD see Karl (1996) and Karl and Pisoni (1994).

By using a collection of descriptive statistics compiled for the IMTSD by Bradlow, Torretta, and Pisoni (1995), the original 100 types were sorted into 80 sentences representing the top 80 most intelligible sentences averaged across all 20 talkers. These statistics were used in the selection process because individual talkers differed in their average intelligibility and thus the same sentences did not fill the top 80% for each talker. The twenty talkers varied in their intelligibility from 81.1% to 93.4% – based on average words correct per sentence across 200 listeners (see Bradlow, Torretta & Pisoni, 1995). While actual population estimates for the range of talker variability have never been formally assessed and hence we cannot relate the magnitude of this difference to that outside the laboratory, it is clear that a considerable amount of variation is present across these 20 talkers in the database.

A computer program was then created to select the individual stimulus files from this collection of 1,600 sentence tokens. The program generated lists of stimulus file names that consisted of a sentence label, a talker label and a gender label. The program worked first by randomly selecting 5 male and 5 female talkers from the list of 20 talkers. The ten selected talkers were the alpha talkers for each stimulus file and the 10 unselected talkers were the beta talkers. From each of the alpha talkers, 4 sentences were then randomly selected from the 80 sentence types. The program additionally monitored sentence selection within each set so that 40 different sentence types were chosen for each file. These sets of 40 sentences were used for the study phase stimuli. Stimulus files were randomly sampled after generation to check that the program had selected a correct set of sentences.

The test phase stimuli were created by processing the study phase files arrays. First, the second and fourth sentences of each alpha talker were duplicated and their talker label changed to a beta talker. Each of these 20 study phase sentences was then replaced by a randomly selected new sentence type spoken by a beta talker. The first and third sentence of each alpha talker were likewise duplicated but with their sentence label changed to a randomly selected new sentence. The result of these operations was a set of test stimuli consisting of 80 sentences matched to the study stimuli.

The following notation will be used to identify the sentences in the test phase sets. A sentence spoken by an alpha talker that appears in both the study and test phase files will be an O_sO_t (old sentence/old talker) sentence. New sentence types spoken by alpha talkers at test will be N_sO_t (new sentence/old talker) sentences. A sentence type spoken by a beta talker that appears in both the study and test sets is an O_sN_t (old sentence/old talker) sentence. Finally, N_sN_t refers to a new sentence type spoken by a beta talker (new sentence/new talker). Each test phase file consists of 20 sentences each of OO, NO, ON, NN.

These elaborate stimulus selection strategies were undertaken to avoid any confounds from the inherent differences in sentence and talker intelligibility. It is important that the differences between items inherent in any corpus of speech samples be minimized through randomization. While an item-by-item analysis of each stimulus set file was not undertaken, the multi-step randomization procedure could reliably be expected to evenly distribute the effects of both talker and sentence differences.

Stimulus presentation and data collection were controlled by two computer programs which ran the study and test phases. Sentences in both phases were presented to subjects binaurally over matched Beyer Dynamic DT-100 headphones at a comfortable listening level (75 dB SPL). The programs ran on multiple personal computers and data were written to formatted text files. The keyboard and electronic button-boxes were used to collect responses.

Procedure

Upon completion of consent and credit forms as well as subject information sheets, each subject took a seat in a computer kiosk. Subjects were instructed to find a comfortable position from which to view the CRT monitor and reach the computer keyboard. Subjects were handed headphones and then given instruction as to the task they were to perform. Subjects were told they would be hearing spoken sentences over their headphones. After each sentence had ended, subjects were asked to type in what they heard at an arrow input prompt that appeared on the monitor. Subjects were told the sentences would be simple, meaningful statements in English. Finally, subjects were told that phonetic spellings were acceptable and that blanks could be left for words they could not understand. Most importantly, subjects were not told anything about the ensuing memory test nor were they instructed to attend to any particular

property or set of attributes of the sentences. Subjects were allowed to ask questions to clarify any unclear instructions and then each subject began the experiment.

Subjects hit enter on the keyboard to begin the study phase of the experiment. Upon hitting enter, the program started by reading in a specified study stimulus set and randomizing the files before presentation. A one second delay preceded the beginning of each sentence. Once the sentence had played in full, an arrow prompt -> appeared which was the signal for subjects to type in the sentence they heard. After typing the sentences and making any corrections, subjects hit enter to record their response and advance to the next sentence. Upon completion of the last transcription, the program ended and subjects waited for the experimenter to initiate the test phase of the experiment. Despite the self-paced nature of stimulus presentation, all subjects usually finished the study phase about the same time within 10 to 15 minutes.

Subjects were again instructed before the beginning of the test phase. Subjects were told that in this phase of the experiment they would be hearing 80 total sentences. Subjects were then shown the button boxes used for their response input. The button box consisted of two buttons each with a corresponding LED indicator. Subjects were instructed to judge each sentence as “old”, meaning it was heard during the study phase – or “new”, meaning it was not heard during study. Each button on the button box was labeled as either “old” (1) or “new” (2). Both the number and word label for the button choice appeared on the button box itself as well as the on screen instructions that began each experiment. Subjects were told to guess “old” or “new” in the event that they could not otherwise judge a sentence. Subjects were given no explicit criterion for judging a sentence as “old” or “new” beyond the instructions that the sentence “appeared in the study phase” description. Several subjects did report during debriefing, however, that one of the determining factors of their judgment was a particular speaker’s distinctive way of pronouncing certain sentences that were presented during the study phase.

Subjects began the test phase by hitting enter after which a brief “Ready” message flashed onscreen before the sentence began. After each sentence was presented, the familiar arrow prompt appeared at which point subjects recorded their judgment. Subjects could be certain the computer had received their response by the light above the pushed button flashing briefly and also by the flash of a new “Ready” message onscreen. The minimal response required of the subjects in this phase of the experiment usually resulted in finishing times of less than 10 minutes despite there being twice the number of sentences as in the study phase. Like the study phase, all subjects finished around the same time despite the self-paced stimulus presentation. Upon completion of the test phase, subjects were debriefed and asked for any comments concerning the experiment and then allowed to leave the laboratory.

The data generated for each phase of the experiment were recorded in two corresponding data files. Each file contained a record of the stimulus set files used as well as the subject and session number. Each study phase file also contained the subject-transcribed sentences paired with the actual content of each sentence. The test file sentences contained a numeric code corresponding to the old (1) or new (2) judgment of each sentence paired with the actual label for each sentence.

For ease of scoring, a computer recovery program was written to process each test-phase file into a formatted data file. The program analyzed the subjects’ responses and returned tallies of the number of correct and incorrect judgments for each of the four test sentence categories. Correct judgments for OO and ON sentences were ‘old’ or ‘1’ while ‘new or ‘2’ were the correct judgments for NN and NO sentences. These resulting files were then analyzed using Microsoft Excel.

Results

Due to a computer malfunction that corrupted 8 subject data files as well as additional computer problems that interrupted 3 students during the experiment, 11 subjects' data were discarded before analysis. Finally, an additional subject was eliminated from analysis based on his negative d' score (see below). Thus the data from 21 subjects was used in analysis.

Study phase files were first analyzed to measure transcription accuracy. Following procedure common in transcription task scoring (Karl, 1996), phonetic spellings and obvious typos were counted as correct transcriptions. This was clearly an easy task. All subjects scored above 85% of words accurate. Transcription errors were usually systematic such as 'cow' instead of 'cop' or article substitutions like 'a' for 'the'. From this we can conclude that when subjects did not fully hear parts of a sentence they attempted to compensate using partial stimulus information. The use of partial information, termed "sophisticated guessing," has both a long experimental and theoretical history in psycholinguistic research (Solomon & Postman, 1952; Savin, 1963).

Recognition Accuracy: Figure 1 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

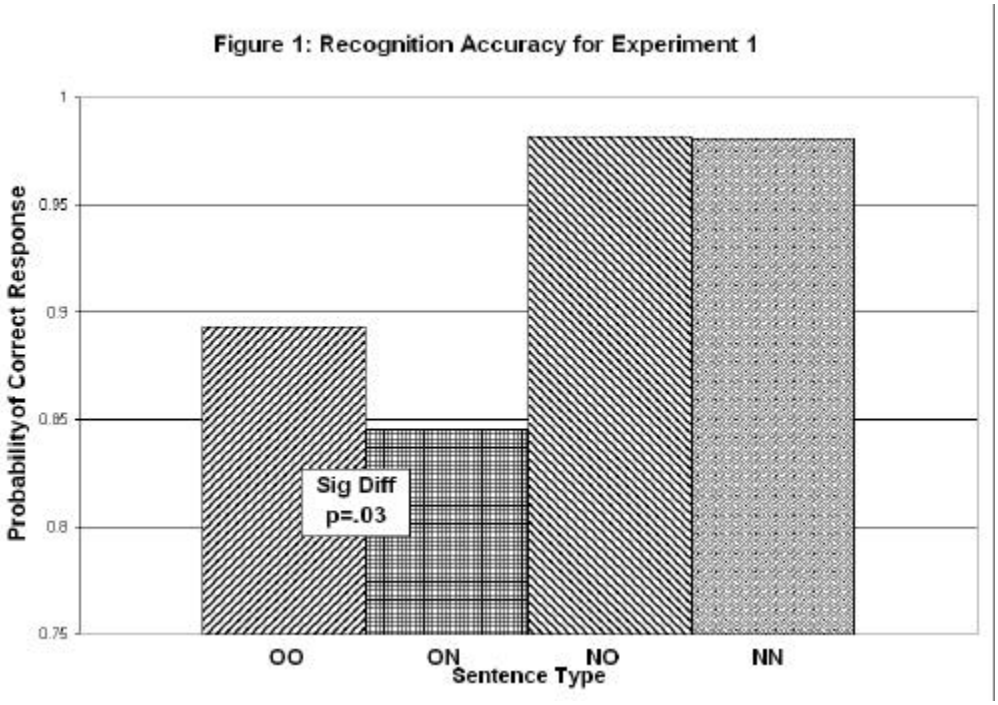


Figure 1: Recognition accuracy for Experiment 1 grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability in paired t-tests.

Test phase responses were first tallied into correct judgment totals for each of the four sentence categories. For analysis, these totals were expressed as probability of correctly recognizing a sentence. Thus, a subject judging OO sentences as old 17 out of 20 times would have 85% accuracy. Subjects generally showed better accuracy in judging OO sentence than ON sentences. Individually, only 3 of the 21 subjects had higher accuracy for ON sentences. Overall, performance accuracy for OO sentences

ranged from 60% to 100%, ON sentence accuracy varied across a smaller range from 70% to 95% correct.

The average percent correct for OO sentences was 89.5% while average correct for ON sentences was 84.5%. This difference was significant using a paired t-test, $t(20)=2.334$, $p=0.027$. No significant differences in accuracy were observed for NN sentences and NO sentences although performance on both these sentence types was near ceiling with accuracy averages of 99% for both NN and NO sentences. The ceiling effect is further revealed by the range of accuracy scores that varied only from 95% to 100%.

Two means were computed for each subject from the category pairs OO/ON and NN/NO. One average thus represents the old sentence average accuracy (OAA) while the second reflects new sentence average accuracy (NAA). Average OAA was 87.5% while average NAA was 98.4%. This difference was significant with a paired t-test, $t(20)=4.96$, $p<0.0005$. This suggests that new sentences are easier to judge than old sentences and that subjects display a response bias to response "new" more often than "old."

A 2x2 ANOVA was then performed on the accuracy results. This analysis yielded a significant main effect for sentence type, $F(3, 80)=13.9$ and $p<0.0005$. Neither the main effect for voice nor the 2-way interactions were significant. The results ANOVA further corroborate findings from the t-tests that old sentences were recognized significantly better then the new sentences. The observed ceiling effect, however, makes these conclusions tentative at best.

d' scores: Two d' measures for each subject were then computed from the recognition data. Hits were scored for every judgment of 'old' given to an OO or ON sentence. Likewise, misses were scored for 'new' judgments of these sentences. A correct rejection was scored for each 'new' judgment given an NN or NO sentence. Finally, false alarms were scored when a subject judged either of these two types as 'old.' A d' measure of sentence discriminability for sentences produced by "old" voices was calculated by subtracting the average false alarms for NO and NN sentences from hits for OO sentences. Subtracting average false alarms for NO and NN sentences from hits for ON sentences generated the d' measure for new-voiced sentences. The false alarm rates for NO and NN were averaged because NN and NO accuracy were not significantly different. This measure provides a better overall measure of subjects' false alarm rate. After the d's had been calculated in this manner for all subjects, one outlier emerged had negative d'. Since the reasons behind this discrepant performance were unknown, the subject was removed from the final analysis and thus all results discussed here exclude this subject. The average d' was 3.56 for old voices while the average d' for new voices was 3.24. The paired t-test for the difference in d's yielded a significant difference, $t(20)=2.522$, $p=0.02$ and thus subjects were able to discriminate between "old" and "new" sentences better when the sentences were in an old voice at test that when the sentences were in a new voice.

Discussion

The results of this experiment suggest two important –and related– conclusions. First, these results provide support for the proposal that the memory representation for spoken sentences is rich enough to preserve voice information. Clearly, the significantly higher accuracy in OO identifications compared to ON identifications suggests that voice information was preserved in memory along with the lexical organization of the sentence itself. Thus, we can conclude that in this task, voice information was responsible for as much as a 5% improvement in recognition accuracy although performance was at or near ceiling. Moreover, under the assumption that subjects were not explicitly attending to voice information, the processing and storage of this information appears to be automatic without explicit conscious attention or awareness. The assumption about subject behavior seems plausible, since voice information, much less a memory task, was not mentioned in subjects' instructions. Still, it is possible

that a confounding number of subjects did specifically attend to voice information during the study phase. A more extensive post-experiment interview than the one conducted for this experiment may be able to resolve this issue.

Secondly, as an obvious corollary to the first conclusion, these results suggest that voice information can also be utilized to aid performance in a recognition memory paradigm. A rich memory representation would be useless without search and retrieval processes that can make use of that additional information. Further, as many subjects did report using voice information in at least some of their judgments, this experiment cannot speak as to the automaticity of voice information use in the recognition process.

The d' results show that subjects were better able to discriminate "new" sentences from "old" sentences when the sentence was spoken by an "old" talker. This provides further evidence for the encoding of voice information since there should be no difference between new voice and old voice sentence discriminability were voice information absent from memory. The fact that the difference in d' occurred without subjects being instructed to attend to voice information further establishes that voice information encoding is an automatic feature of speech event encoding.

Finally, we can see from the data that, particularly in the NN and NO category, performance was near ceiling. This ceiling effect obscures important statistical differences. On the one hand, the accuracy improvement of voice information could be higher than the average 5% here recorded. Additionally, the difference in accuracy between NN and NO sentences may simply have been too small to detect in a collection of scores whose average was only 1% below ceiling. The following experiment was designed to lower subject performance from the ceiling and hence give a better estimate of voice effects in sentence recognition memory.

Experiment 2: No Transcription

In an attempt to lower subject accuracy and move performance down from the ceiling, we modified the study task used in Experiment 1. It is very likely that the depth of processing encouraged by the sentence transcription task resulted in a memory representation that was based on conceptual processing which made the memory particularly accurate for the lexical sentence information and hence caused the ceiling effects. By removing the transcription task from the study phase, we hoped to bring performance down from ceiling and thus yield a more accurate measure of voice effects in sentence recognition.

Method

Subjects

Subjects were 21 undergraduate psychology students at Indiana University who were given class credit in an introductory course for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability. Again, subjects were not formally assessed for meeting these requirements, but subject data forms do corroborate that subjects did follow these guidelines.

Materials

Stimulus materials for this experiment were taken from the same corpus as Experiment 1. The randomly generated stimulus set files used in the previous experiment were also used here. The same

experimental control programs and data-processing routines described for earlier were also used in this experiment.

Procedure

The basic procedure used in this experiment was identical to that described in Experiment 1 except that subjects were not required to transcribe the sentences after each stimulus presentation. Instead, subjects simply hit the enter key on the keyboard to advance to the next stimulus. Subjects were instructed to listen and pay attention to each sentence. As in Experiment 1 no explicit mention was made of voice information in these sentences or the ensuing recognition task.

Results

Recognition Accuracy: Figure 2 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

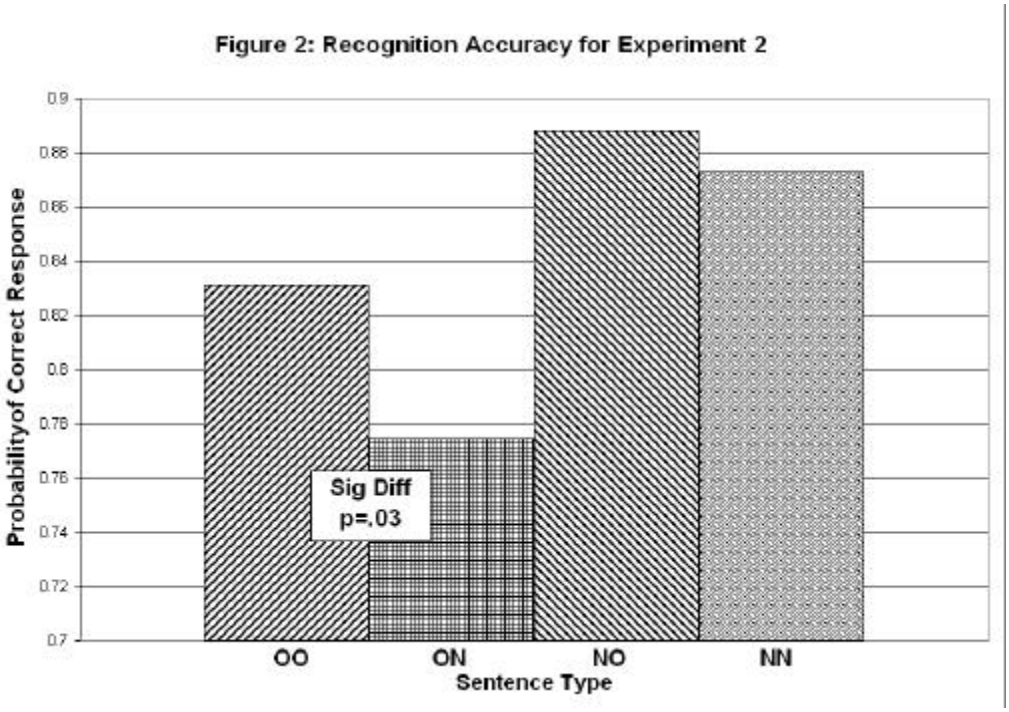


Figure 2: Recognition accuracy for Experiment 2 grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability in paired t-tests.

For OO sentences, subjects accurately judged 83.3% of sentences on average as 'old' in the test phase. ON sentences showed 77.6% average accuracy for "old" sentences with a significant difference of 5.7% in an paired t-test, $t(40)=2.274$, $p<0.05$. While performance in NN and NO sentences was successfully lowered from ceiling, the NN average of 89% and NO average of 87% were still high. Like Experiment 1, there was no significant difference between these two scores.

Old sentence average accuracy (OAA) and new sentence average (NAA) accuracy scores were again computed for each subject. Average OAA was 85.9% average NAA was 87.7%. This difference was not significant in a t-test. The results of experiment 1 which suggested that new sentences were easier

to discriminate than old sentences are thus not replicated in this study – though the direction of the difference in average accuracy is the same.

A 2x2 ANOVA performed on accuracy results yielded a significant main effect for sentence type with $F(3, 80)=7.6$, $p=0.007$. Neither the main effect for voice nor the 2-way interactions were significant.

d' scores: Two d' scores for each subject were calculated. One score compared the two old talker categories while the other score compared performance for the two new talker categories. These d' scores were again calculated using the false-alarm averaging technique described previously. As in Experiment 1, this averaging was justified by a non-significant difference between NN and NO accuracy. The same criterion placed on d' scores in Experiment 1 was also used here, but as no subject had negative d's, all subjects were kept in the analysis. The old-voiced d' average was 2.48 while the new-voiced average was 2.24. This difference was marginally significant, $t(20)=2.01$, $p<.06$. D' scores for old voices ranged from 1.14 to 4.2. New voiced d's ranged from 0.97 to 4.2.

Discussion

The results of this experiment show even stronger voice effects in sentence processing than those of Experiment 1. The 5.7% difference between OO and ON sentences was marginally higher than the 5% difference in Experiment 1. While the non-significant difference between NN and NO sentences was replicated in this experiment, scores were successfully pulled from the ceiling. The difference between the NN and NO scores was also slightly larger than in Experiment 1: 1% in Experiment 2 as compared to no difference Experiment 1. The absence of a transcription task had both the desired effect of lowering performance and increasing the difference between old-voiced and new-voiced category pairs.

The lack of difference between accuracy in new sentence judgments for old and new talkers was replicated in this experiment. Since neither of the two experiments so far discussed have found a difference, it appears that subjects can rely exclusively on lexical information to judge a new sentence accurately regardless of talker. When the sentence is old, however, subjects show significant effects for voice information.

We can explain this pattern of results by considering two situations in which a subject is required to discriminate sentences in this procedure. When presented with a new sentence, subjects easily recognize that the subject matter and phrasing of the sentence are novel. Whether the voice is old or new makes no difference since the words themselves are clearly different. When confronted with an old sentence, however, subjects cannot be certain whether the sentence is the same as one heard before or simply similar to previous sentences in subject matter or phrasing. In such problematic cases subjects may then turn to voice information to judge the sentence. Since old voices were the speakers of old sentences and thus subjects could utilize both their lexical and indexical memories for the sentences, and thus it is more likely that subjects will judge OO sentences correctly than ON sentences. In principle, this effect should occur to some extent in new sentence judgments though the ease of these judgments in most circumstances may mask this affect. By making the study task more difficult, a voice affect for new sentences may emerge.

The d' scores for Experiment 2 were significantly lower than those for Experiment 1. This means that subjects were less sensitive to sentence change after a study task which did not require transcription. This result is not surprising to the extent that transcription requires more attention and conceptual processing than passively listening to a sentence. Thus, subjects would be more sensitive to the semantic differences between sentences after transcription.

It is likely that a passive listening task did not encourage subjects to engage in as intense an attentive process as transcription. Thus, the sentences may not have been as well encoded in memory as suggested by a level of processing perspective (Craik & Lockhart, 1972). This notion is borne-out by the d' prime data which show a decrease in sensitivity to sentence change from experiment 1. General performance accuracy showing a 6% decline also supports this interpretation. But voice information had more of an affect in this experiment than the previous experiment. Apparently, subjects were able to utilize voice information to aide performance in judgment when their conceptual memory for the lexical/semantic characteristics of the sentences was inadequate. Since there was less conceptual information concerning sentences in Experiment 2 by which subjects could make accurate judgments, we can explain both the decrease in performance and discriminability. Moreover, since subjects have made recourse to voice information to compensate for less conceptual information, we can also explain the increase in voice effects between Experiment 1 and Experiment 2.

While the change in study task between Experiments 1 and 2 had the desired effect in pulling performance away from ceiling, it still did not reveal any significant voice effects for new sentences. Unless we posit that new sentences were somehow processed and judged differently from old sentences, there should also be voice effects in old sentence judgment accuracy. Performance was still very high across all conditions and hence the true magnitude of voice effects may still be partially masked. Experiment 3 was designed to lower performance further and hopefully allow more extensive voice effects to be uncovered.

Experiment 3: Word List Transcription

Although the change in study between Experiment 1 and 2 was enough to lower performance from ceiling, subjects still showed extremely accurate judgments for all sentence categories. Thus, in this experiment an additional word list transcription task was added to the study phase to further lower performance and hopefully reveal any voice affects that might be present.

Method

Subjects

Subjects were 21 undergraduate psychology students at Indiana University. Seventeen of the subjects received credit in an introductory psychology. Four other subjects were recruited from an introductory psychology course and paid $5 for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability. Subjects were not formally assessed for meeting these requirements, but subject data forms do corroborate that subjects did follow these guidelines.

Materials

Stimulus materials for this experiment were taken from the same corpus used in experiments 1 and 2. The randomly generated stimulus set files used the earlier experiments were also used in this experiment. An additional list of word pairs randomly selected from the content words of 40 of the 80 sentence types was also used during the study task. For each subject, some of the word pairs would be heard in study phase sentences while others would not. In the test phase, all the word pairs would appear in sentences. The word pair list was designed to reduce the conceptual processing of sentences during study and create interference in recall during test and thus encourage subjects to make more use of voice information in their judgments. The same experimental control programs and data-processing routines described for 1A were also used in this experiment.

Procedure

The procedure in this experiment was the same as in Experiment 2 except that subjects were also required to type in word pairs after each sentence presentation during the study phase. The word pairs were numbered 1 to 40 and presented on a printed page placed beside the keyboard. Subjects were instructed to type in the correspondingly numbered word pair after each sentence was presented. Subjects were further instructed to only read or type word pairs after they had heard the complete sentence. This instruction was given to prevent subjects from reading the word list before sentences were finished and not accurately encoding study sentences. As in the previous experiments, no mention was made of voice information or the recognition memory task.

Results

Recognition Accuracy: Figure 3 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

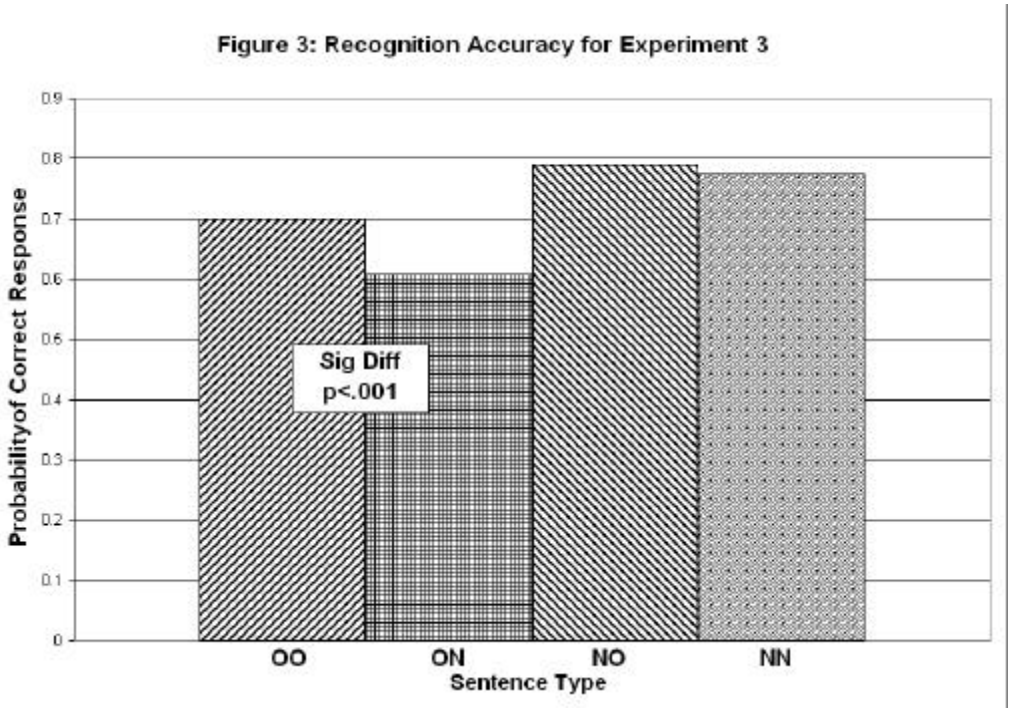


Figure 3: Recognition accuracy for Experiment 3 grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability.

For OO sentences, subjects accurately judged 70% of sentences as 'old' in the test phase. ON sentences showed 60.7% accurate judgments at test for an arithmetic difference of 9.3% which was significant in a paired t-test, $t(20)=4.35$, $p<0.001$. Performance in NN and NO sentences was successfully pulled from ceiling with an NN average of 79% and NO average of 77.5%. However, like experiment 1 & 2, the difference between these two scores was not significant.

Old sentence average accuracy (OAA) and new sentence average (NAA) accuracy scores were again computed for each subject. Average OAA was 70.7% while average NAA was 80.2%. This 9.5% difference neared significance, $t(20)=2.00$, $p=0.06$. This difference is again in the same direction as Experiments 1 and 2 and is nearly significant. Thus we find slight support for the conclusion that new sentences are easier to discriminate than old and additional evidence that subjects has an overall bias to respond "new."

A 2x2 ANOVA performed on accuracy results yielded a significant main effect for sentence type with $F(9, 80)=14.1$, $p<0.005$. Neither the main effect for voice nor the 2-way interactions were significant.

d' scores: Two d' scores for subjects were calculated. One score compared the two old talker categories while the other score compared performance for the two new talker categories. These d' scores were again calculated using the false-alarm averaging technique described previously. As in Experiments 1 and 2, this averaging was justified by a non-significant difference between NN and NO accuracy. The same criterion placed on d' scores in experiment 1a was also used here, but as no subject had negative d's, all subjects were kept in the analysis. The old-voiced d' average was 1.46 while the new-voiced average was 1.18. This difference was significant, $t(20)=4.47$, $p<0.005$. d' scores for old voices ranged from .38 to 3.0. New voiced d's ranged from 0.36 to 2.85.

Discussion

Again, the results from this experiment support a mental representation of speech events that includes extra-linguistic information such as voice information. Also important is a third replication of voice effects for old-voiced despite subjects reporting that they did not attend specifically to voice information during study. Thus, it appears that voice information is encoded automatically without the intention of the listener. Research by Remez et al. (1997) suggesting that voice information is utilized during the phonetic identification offers an explanation for this automaticity: The low-level processes of phoneme identification require voice data and hence lead to voice information encoding during the initial unconscious stages of speech processing.

The difference between accuracy for OO and ON sentences was the largest yet seen in this series of experiments at 9.3%. Thus, the additional voice information provided by OO sentences allows for as much as a 9.3% gain in accuracy. For new sentences, voice information again seemed to make no difference as the 1.5% difference between NN and NO sentences was not significant.

The d' scores for this experiment were the lowest yet seen in this series of experiments. We can interpret these low scores as evidence that the interpolated word-pair transcription task reduced the encoding of sentence information that contributed to higher d' scores in the earlier experiments. We cannot specify, however, whether the word-pair transcription task simply altered the level of processing the sentences were encoded with at study or if the word-pairs interfered with recognition during the study task.

Accuracy in the new sentence categories was pulled well away from ceiling with averages for NN and NO of 79% and 77.5% respectively. Since these scores are even lower than the old sentence accuracy scores from Experiment 1 – which did reveal a significant voice effect – it seems that voice information is simply not utilized to a significant extent when subjects discriminate new sentences. The previous suggestion that voice information may not be utilized because new sentences are easy to discriminate based solely on their novel lexical and semantic content seems well supported by these results.

From Experiments 1 to 3, we find increases in the accuracy difference between old-voiced and new-voiced sentences while overall judgment accuracy and d's significantly decline. Thus, we must consider how voice information provides a greater gain in sentence accuracy judgments when the accuracy of the judgments themselves decreases overall. One possibility is that the full sentence transcription task used during study in Experiment 1 required more attention than the word pair transcription of Experiment 3. The additional requirement was met at the expense of voice information encoding and thus each stimulus could not be processed as deeply as in Experiment 1. Since not as much voice information was encoded at study in Experiment 1, subjects could not benefit as much from this information when making test phase judgments.

A second possibility also results in less voice information being available after sentence transcription, but through a different process. It is possible that the full sentence transcription task itself interfered with voice information encoding to the extent that subjects likely rehearsed the sentences while typing them. This rehearsal may have encouraged conceptual coding of the sentences and obscured surface feature information such as talker voice. Thus, voice information may have initially been encoded to the same extent as in Experiment 3, but the process of rehearsal during transcription typing may have attenuated this voice information and resulted in a loss of surface perceptual features. The result is that less voice information was available at test in Experiment 1 to assist in sentence judgments. The study task in Experiment 3, which did not require sentence rehearsal, would not have encouraged conceptual processing that would likely have obscured surface perceptual information. Thus, when subjects were required to judge sentences as old or new in the test phase, this surface information would have aided in their judgment of sentences.

So far, we have examined test phase voice effects after study tasks that did not direct explicit attention to voice information. The final experiment here discussed was designed to assess voice affects after a task that directly encouraged processing of voice information at the time of study.

Experiment 4: Voice Monitoring

In the first 3 experiments, we measured the indirect effects of voice information when subjects were not instructed or encouraged to pay attention to these properties of the speech signal. The procedure in the present experiment was specifically chosen to encourage subjects to pay explicit attention to voice information without making the real aims of the experiment explicit. This change in task was considered because we had no baseline measurement for the effects of voice information when subjects were explicitly asked to attend to this surface feature. Without information concerning voice effects in an explicit encoding condition, we cannot judge the relative magnitudes of the voice effects from Experiments 1-3. The results of this Experiment 4 can thus be considered a kind of baseline by which to compare the voice effect results from Experiments 1, 2 and 3.

Method

Subjects

Subjects were 21 undergraduate psychology students at Indiana University. Fifteen of the subjects received credit in an introductory psychology. Six other subjects were recruited from an introductory psychology course and paid $5 for their participation. Requirements listed on the sign-up sheet for subject participation specified native English speakers with normal or corrected vision and average typing ability.

Materials

Stimulus materials for this experiment were taken from the same corpus used in the three previous experiments. The randomly generated stimulus set files used in Experiments 1 and 2 were also used in this experiment. This experiment returns to the simpler design of Experiments 1 and 2 and thus did not utilize a word-pair list. The same experimental control programs and data-processing routines described for Experiment 1 were also used in this experiment.

Procedure

The procedure used in this experiment was the same as Experiments 1 and 2 except that now subjects were asked to explicitly identify the gender of the talker during the study phase. After each sentence was played, subjects indicated the gender of the sentence-talker by typing in male or female and then hitting *enter* to record their response. Subjects were allowed to abbreviate their responses to 'm' or 'f'. Subjects were also instructed to guess when they were uncertain as to talker gender. By specifically calling attention to voice information in this deliberate way, it is expected that subjects will show increased effects for voice information in this experiment as compared to Experiments 1-3. These increased effects should emerge in larger differences in accuracy between old and new voiced sentences as well as higher d' scores.

Results

Recognition Accuracy: Figure 4 summarizes recognition accuracy results. The boxes containing probability values span columns whose difference is significant at that probability value.

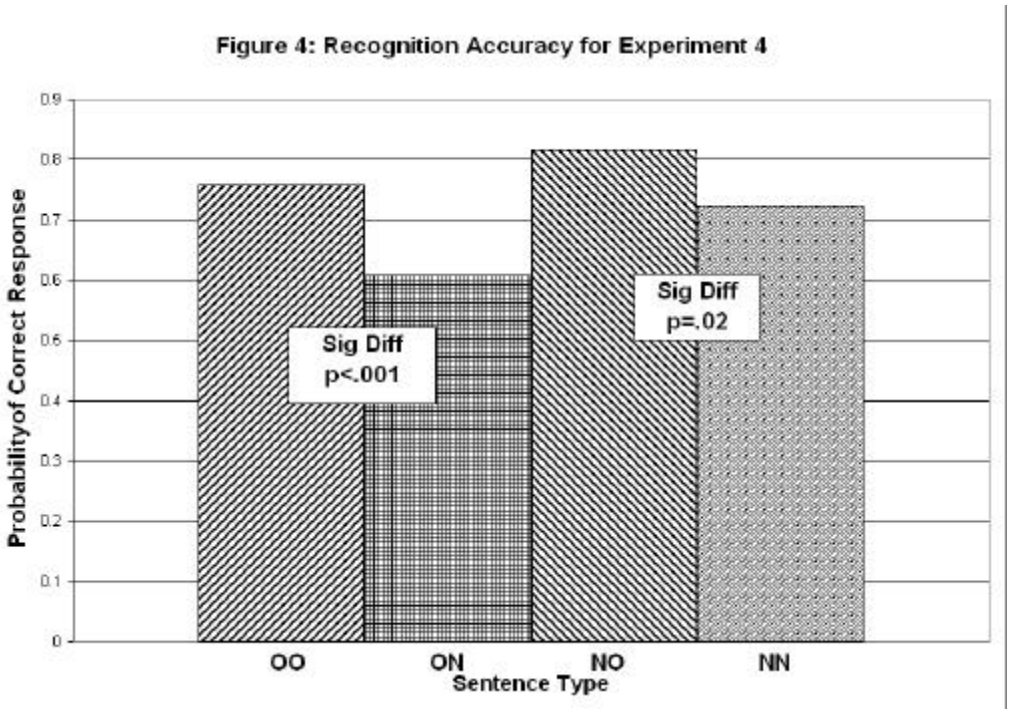


Figure 4: Recognition accuracy for Experiment 4 grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability.

For OO sentences, subjects accurately judged 72.9% of sentences as 'old' in the test phase. ON sentences showed 59% accuracy at test for an arithmetic difference of 13.8% which was significant in a paired t-test, $t(20)=3.54$, $p=0.002$). Performance in NN and NO sentences was well below ceiling with an NN average of 83.4% and NO average of 74%. For the first time in this series of experiments, this new sentence difference across voice was significant, $t(20)=2.58$, $p=0.018$.

Old sentence average accuracy (OAA) and new sentence average (NAA) accuracy scores were again computed for each subject. Average OAA was 66.4% while average NAA was 76.4%. This 10% difference, the second largest encountered so far, was not significant because of the variability in subject performance. This difference is in the same direction as Experiments 1-3. This result provides converging support for the notion that new sentences are easier to judge than old sentences.

A 2x2 ANOVA performed on accuracy results yielded a significant main effect for sentence type with $F(3, 80)=13.9$, $p<0.005$. The main effect for voice was not significant. Likely the result of significant differences between both OO/ON and NN/NO sentences, the 2-way interaction between sentence and voice was significant with $F(3.80)=11.54$, $p=.003$.

d' scores: Two d' scores were also calculated for each subject. One score compared the two old talker categories while the other score compared performance for the two new talker categories to test for differences in discriminability between old voiced and new voiced sentences. Despite the fact that this experiment did find a significant difference between NO and NN accuracy, the same false alarm averaging technique used in experiments 1-3 was utilized. This was done to keep all d's derived in this series of experiments the result of the same calculation procedure and hence to add validity to the overall analyses discussed below. The same criterion placed on d' scores in experiment 1A was also used here, but as no subject had negative d's, all subjects were kept in the analysis. The old-voiced d' average was 1.54 while the new-voiced average was 1.15. This difference was significant, $t(20)=3.50$, $p=0.002$. d' scores for old voices ranged from .91 to 2.57. New voiced d's ranged from 0.45 to 1.94.

Discussion

The recognition results of this fourth experiment are important for several reasons. First, they provide a baseline of voice effects with which to compare the other experiments. Since the focus was on the gender of the speaker of each sentence, subjects were explicitly encouraged to attend to voice information. This manipulation seems to have the desired effect as differences between old and new voiced sentences showed the largest differences in recognition accuracy of all four experiments. Indeed, the difference between OO and ON sentences in this condition is almost 3 times that measured in Experiment 1. The difference between NO and NN sentences is over 7 times larger than the difference in Experiment 1; however, the ceiling effect may explain some of this difference. The additional voice information provided by OO sentences allows for as much as 13.8% gain in accuracy. For the first time in this series of experiments, NN sentences were judged significantly more accurately than NO sentences. The additional "novelty" that a new talker added to NN sentences was worth as much as a 9.3% increase in recognition accuracy. Thus, when subjects were explicitly directed to attend to voice information, they encoded this information better and hence showed greater differences in performance due to the effects of this additional encoding.

Accuracy in the new sentence categories was pulled well away from ceiling with averages for NN and NO of 81.7% and 72.4% respectively. The voice effect found for the new sentences in this experiment is likely a result of the explicit voice monitoring study task that encouraged encoding behavior for sentences that was not present in the earlier experiments. The possibility that this effect was present but obscured by ceiling effects is doubtful because new sentence accuracy scores in this

experiment were not significantly different from new sentence accuracy in Experiment 3 which did not display any voice effects.

The earlier suggestion that voice information is not utilized for new sentence judgment is contradicted by these data. However, new sentences showed only a 9.3% difference in accuracy based on voice compared with the 13.7% difference for old sentences. While voice information was undoubtedly utilized to distinguish NN and NO sentences in this experiment, the effects were still less than those observed for OO and ON sentences. Thus, rather than the previous conclusions of Experiments 1-3 that voice information was not utilized in distinguishing new sentences, we can conclude that voice information was also used in new sentence judgment. Again, this result is not surprising considering the study task of Experiment 4 that encouraged subjects to attend to voice information. Since voice information was likely to be encoded in the sentence memory representations, subjects would more likely have recognized the voice of an old talker speaking a new sentence. This familiarity may have interfered with the subjects' ability to judge the sentence based on the lexical/semantic content alone. New sentences spoken by new talkers would not have this added familiarity and thus subjects could be more accurate in judging these sentences as "new."

These results are also important as a fourth replication of voice effects in recognition memory for sentences. Needless to say, an experimental procedure that calls attention to voice information and then finds effects of that information is not particularly striking. However, these results are still important because subjects again demonstrate incidental encoding of talker identity information in their memories for the sentences. Since no subject was told to pay attention to talker identity apart from gender – and no subject was instructed to remember the sentence or voice of each stimulus – we can reliably conclude that this information is encoded even without the intent of the listener.

As in Experiment 2 and 3, we again find that increases in the accuracy difference between old and new voiced sentences are accompanied by lower overall accuracy and lower d's. Although subjects were less sensitive in this experiment to sentence change when compared to Experiments 1 and 2, the affect of voice information was significantly larger here for both old and new-voiced sentences. Apparently whatever factors caused the decrease in judgment accuracy did not affect the contribution of voice information to sentence judgment. The difference in study tasks is the best explanation for this result. Even though subjects were instructed to listen to all the sentences fully during study, it is possible that subjects attended to stimulus sentences only long enough to determine speaker gender. When confronted with sentences in the test phase, subjects who only processed sentences to uncover talker gender would find it difficult to judge sentences as old or new. Likewise, sensitivity to sentence change would also be less if subjects more or less ignored lexical and semantic information in favor of talker gender information. Thus, preferentially attending to voice quality over lexical information can explain at once the increase in voice effect across both old and new sentences as well as the decrease in overall accuracy and d'. Clearly, then, some way of insuring that subjects listen to sentences fully – without using a transcription-style task – is needed to make confident conclusions from these results.

Overall Results and Discussion

Between Experiments Recognition Accuracy t-Tests: Recognition judgment accuracy was compared between all four experiments first by pair-wise t-tests as data became available over this series of experiments. The pattern which may become apparent in the data in which both d' and recognition accuracy display mostly monotonic behavior from Experiment 1 to Experiment 4 is simply a interesting pattern in the data. Since no particular experimental component was varied systematically across Experiments 1 to 4, the trend in the data is merely coincidence.

Independent t-tests were conducted pair-wise between experiments to test the significance of these changes. Figure 5 summarizes these results.

Figure 5: Mean Accuracy Grouped by Sentence Category

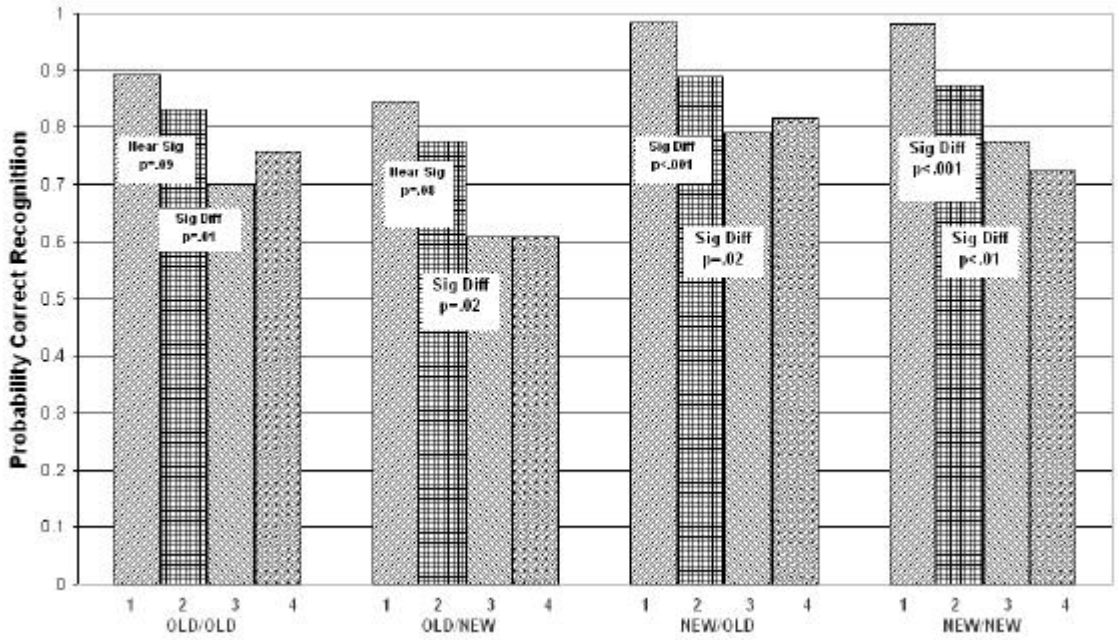


Figure 5: Mean recognition accuracy grouped by sentence category. Probability boxes span the columns whose difference is significant with that probability from unpaired t-tests.

The 6% and 7% decline in recognition accuracy from Experiment 1 to Experiment 2 both neared significance in the unpaired t-tests: $t(40)=1.69$, $p=0.098$ for OO recognition accuracy and $t(40)=1.75$, $p=0.087$ for ON recognition accuracy. OO and ON recognition accuracy in Experiment 3 declined an additional 13% and 17% respectively which were both significant decreases from Experiment 2's results in paired t-tests: $t(20)=2.73$, $p=0.01$ for OO recognition accuracy and $t(20)=3.33$, $p=0.002$ for ON recognition accuracy. Experiment 4's accuracy data is a mix of further declines and some small increases: OO recognition accuracy increased with respect to Experiment 3 while ON recognition remained unchanged. Neither of these values were significantly different from Experiment 3 but were significantly different from Experiment 2 by unpaired t-tests: $t(40)=2.20$, $p=0.03$ for OO recognition accuracy, $t(20)=3.22$, $p=0.002$ for ON recognition accuracy.

ON and NN recognition accuracy show similar trends to "old" sentence recognition accuracy. The average decline of ~ 1% for both "new" sentence categories from Experiment 1 to 2 was significant in paired t-tests: $t(40)=4.46$, $p<0.0005$ for NN recognition accuracy and $t(40)=5.48$, $p<0.0005$ for NO recognition accuracy. Both "new" sentence recognition accuracy scores dropped an additional average of 1.5% from Experiment 2 to Experiment 3. Again, these small declines were significant in unpaired t-tests: $t(40)=2.57$, $p=0.014$ for NO recognition accuracy and $t(40)=2.52$, $p=0.016$ for NN recognition accuracy. The differences in "new" sentence recognition were inconsistent from Experiment 3 to 4, but both were significantly lower than Experiment 2's recognition accuracy results in unpaired t-tests: $t(40)=3.40$, $p<.005$ for NO recognition accuracy and $t(40)=2.27$, $p<.03$ for NN recognition accuracy.

The difference scores between the two "old" sentence categories and the two "new" sentence categories were computed by subtracting each subjects ON recognition accuracy from their OO accuracy. Figure 6 displays the accuracy different scores for each experiment.

Figure 6: Recognition Accuracy Difference Scores

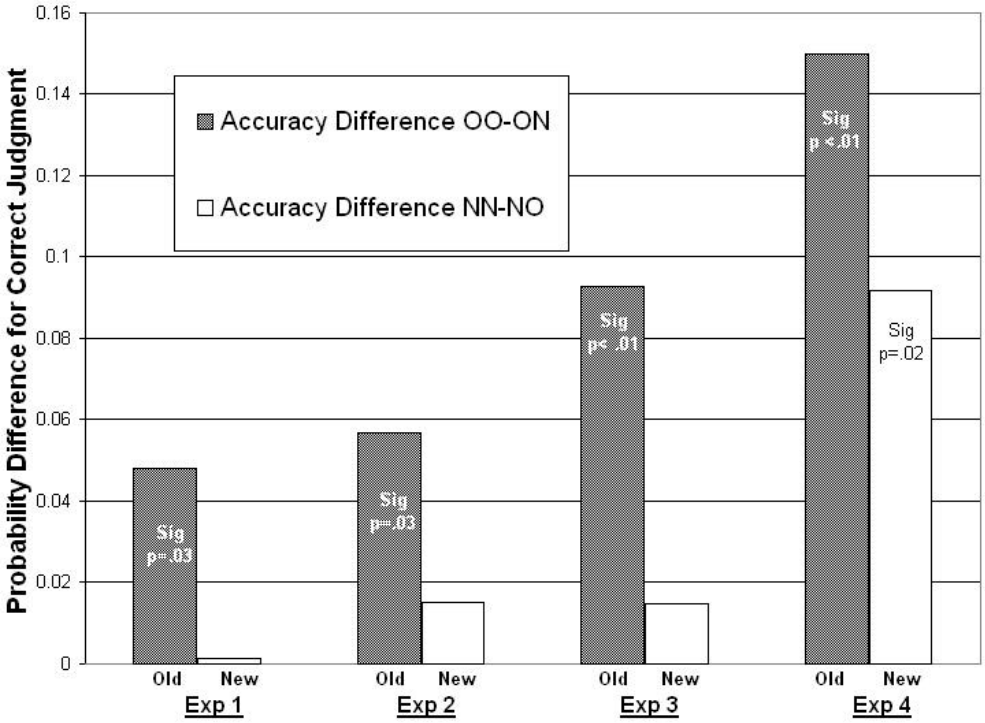


Figure 6: Recognition accuracy difference scores. Probability boxes span the columns whose difference is significant with that probability from paired t-tests.

The same was done for NO and NN accuracy. These new sets of derived measures were then compared in pair-wise t-tests again done as each additional experiment's data became available. Though successfully pulled from ceiling, the accuracy differences observed in all four experiments were not significantly different from one another. The different between OO and ON sentence recognition accuracy in Experiment 1 was 4.8% and not significantly different from the 5.7% in Experiment 2. The "old" sentence accuracy difference of 9.3% was significantly different from Experiment 1 in, $t(40)=4.5$, $p<0.005$ but this score was not significantly different from either Experiment 1's or Experiment 4's "old" accuracy difference. Finally, Experiment 4's accuracy difference for "old" sentences was 15% and significantly different from both Experiment 2 and Experiment 1 in unpaired t-tests: $t(40)=2.31$, $p<.05$ when compared to Experiment 1 and $t(40)=2.04$, $p=0.048$ when compared to Experiment 2.

From the pattern of "old" and "new" sentence recognition accuracy, it appears that recognition was easiest after a sentence transcription task and hardest after an interpolated word list or voice-monitoring task. Moreover, the effect of voice on recognition accuracy became larger relative to overall

recognition accuracy across the four experiments. That is, as average accuracy decreased from Experiments 1 to 4, the voice difference between old sentences increased monotonically. This pattern of results could serve as a guide for future research. Transcription study tasks yield superior recognition accuracy and show comparatively lower voice effects at test than voice monitoring study tasks. In particular, it is clear that a transcription task so thoroughly acquaints subjects with the lexical and semantic content of sentences that "new" sentence judgments are at ceiling. For experiments such as the present series that may look for effects in "new" sentence recognition accuracy, a transcription study task should not be utilized. Likewise, an experimenter looking to minimize voice effects in recognition performance would not want to choose a task such as the interpolated word-list transcription that shows large and significant effects for voice in "old" sentence recognition accuracy.

Old sentence average accuracy scores (OAA) from Experiment 1 were then compared with Experiment 2 OAA in an unpaired t-test that did not reach significance – the difference between averages being only .016%. Old sentence judgments appear to have been equally easy in both experiments. New sentence average accuracy scores (NAA) from Experiments 1 and 2 were significantly different with in an unpaired t-test, $t(40)=3.72$, $p<0$ .001 from an average arithmetic difference of 11%. New sentence judgments were apparently easier for subjects in Experiment 1 than in Experiment 2 and easier than "old" sentence judgments in both experiments.

OAA scores from Experiment 2 were then compared with Experiment 3's OAAs in an unpaired t-test which found a significant difference between the two, $t(40)=2.73$, $p<0.01$ with an arithmetic difference of averages equal to 15%. Old sentence judgments appear to have been easier in experiment 1B – not surprising considering the difference in task requirements for the two experiments. The difference in NAA scores between Experiments 2 and 3 was close to significant in an unpaired t-test, $t(40)=1.84$, $p=0.07$ with an arithmetic difference of averages equal to 7.5%. New sentence judgments seem to have been slightly easier in Experiment 2 than in Experiment 3.

OAA scores from Experiment 2 were then compared with Experiment 4's OAAs in an unpaired t-test which found a significant difference, $t(40)=3.45$, $p=0.001$ with arithmetic difference average of 15%. Old sentence judgments appear to have been easier in experiment 1B – not surprising considering the difference in task requirements for the two experiments. The 11.5% difference in NAA scores between 1B and 1D was significant, $t(40)=2.52$, $p=0.016$. New sentence judgments seem to have been slightly easier in experiment 1B than in 1C. The overall observed trend for subjects to respond "new" also contributed to the consistently higher new sentence accuracy – but these judgments were also quite accurate apart from this bias.

Recognition Accuracy ANOVA: A 4x2x2 analyses of variance were performed on the recognition accuracy data from all four experiments. The judgment accuracy data yielded two significant main effects for experimental condition and sentence with $F(15, 320)=48.0$, $p<0.005$ and 59.6, $p<0.005$ respectively. The main effect for voice neared significance with $F(15, 320)=3.32$, $p=0.07$. Along with the extensive t-tests discussed above, the results of the ANOVA indicate that the different study tasks that were used in each experimental condition did produce changes in recognition accuracy. Additionally, the previous t-tests also preempted the main effect for sentence as the difference in accuracy for old and new sentences was significant in all the previous analyses. The lack of any voice effects obtained for new sentences within 3 of the 4 experiments explains the near-significance of the voice main effect even though old sentences consistently showed an effect of voice. The 3-way interaction was not significant in this analysis.

Between Experiment t-Tests: Figure 7 shows the average d' values for each condition. Figure 8 shows the average d' values compared across experiments.

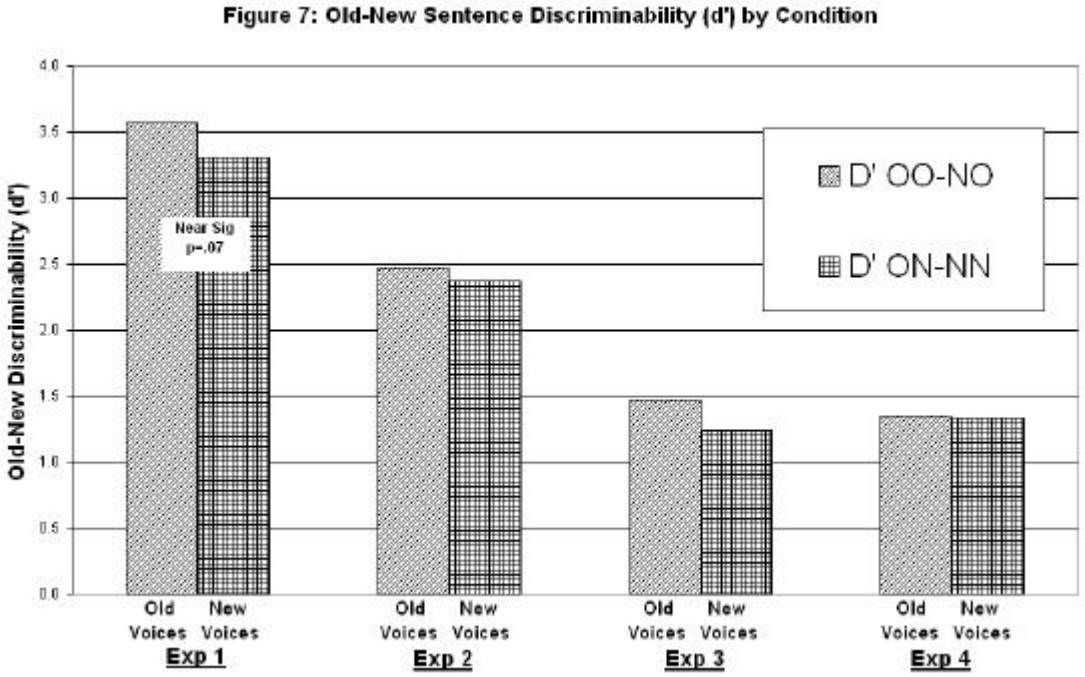


Figure 7: "Old" and "new" sentence discriminability (d') grouped by Experiment. Probability boxes span the columns whose difference is significant with that probability from paired t-tests.

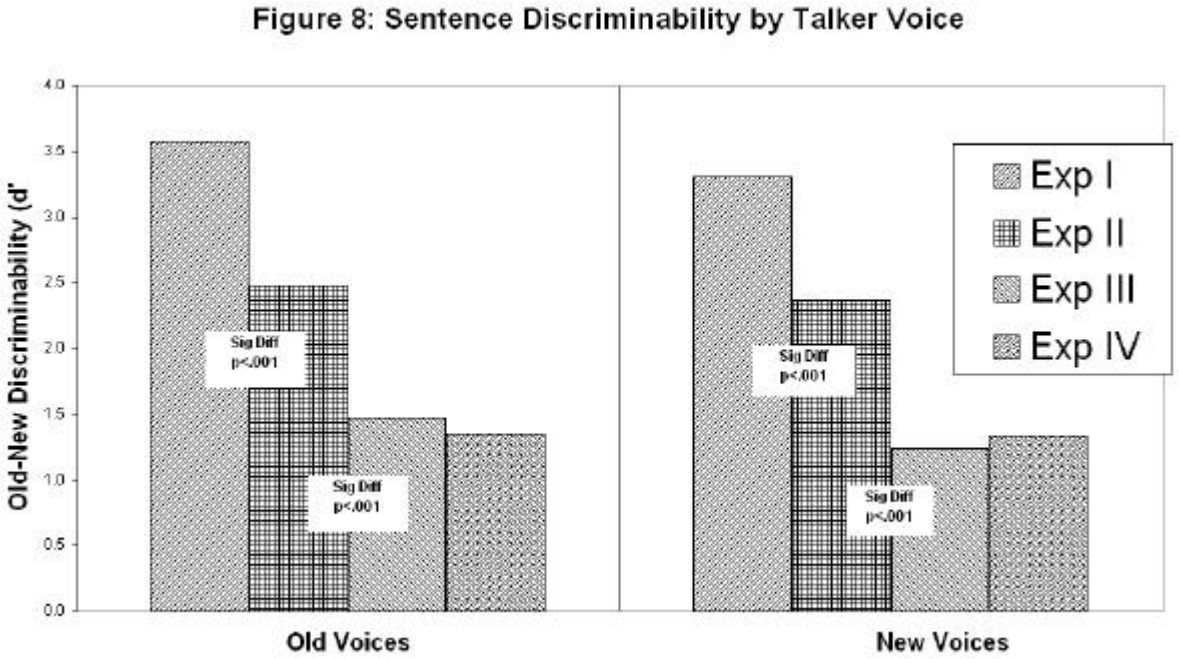


Figure 8: "Old" and "new" sentence discriminability (d') grouped by talker voice. Probability boxes span the columns whose difference is significant with that probability from unpaired t-tests.

In comparison to the average d' scores from Experiment 1, both average d's were significantly smaller for Experiment 2. The average d' scores in Experiment 1 were 1.08 and 1.01 larger than the average d's in Experiment 2 for old voiced and new voiced sentences respectively. An unpaired t-test between old voiced sentence d' scores in Experiment 1 with old-voice d's in Experiment 2 was significant, $t(40)=4.89$, $p<0.001$. The difference between new voiced d's in the two experiments was also significant, $t(40)=3.76$, $p<0.001$. These results suggest that subjects were significantly less sensitive to sentence change across voice condition in Experiment 2 than Experiment 1. This pattern supports the hypothesis that the transcription task encouraged elaboration and a high degree of conceptual processing, which both aided later recognition.

Both of the average d' scores from Experiment 3 were significantly lower than average d's and 2 in unpaired t-tests: $t(40)=3.97$, $p<0.005$ for old voice d' and $t(40)=4.26$, $p<0.005$ for new voice d'. This indicates that the word-pair transcription task had the effect of lowering sensitivity to sentence change. A likely explanation for this phenomena is that subjects were less able to remember whether the content words of the sentence they were judging appeared in a sentences during the study phase (and hence the sentence was old) or whether the content words had appeared in the word pair list (and hence the sentence was not necessarily old).

Like the accuracy scores, the average d' scores for Experiment 3 were not significantly different from Experiment 3's average d' scores. These scores were both significantly lower than Experiment 2's average d' scores in unpaired t-tests: $t(40)=4.1$, $p<.005$ for old voice d' and $t(40)=3.9$, $p<.005$ for new voice d'. Like the word-pair transcription task, the voice-monitoring task also had the effect of lowering sensitivity to sentence change. It is likely that subjects may have paid only enough attention to the study-phase sentences to judge gender. With more of their attention on voice quality and less on the lexical and semantic content of the sentences during study, subjects may have had a harder time at study recognizing both old and new sentences.

d' ANOVA: A 4x2 ANOVA performed on d' scores from all four experiments yielded two significant main effects for experimental condition and sentence type with $F(7, 160)=91.5$, $p<0.005$ and 8.8, $p<0.005$ respectively. As in earlier t-tests, the ANOVA of d's indicates that the study-phase task had a significant effect on how well subjects could distinguish old sentences from new sentences. The second main effect likewise confirms the earlier t-test results that found significant or differences between old and new sentence d's in all four experiments. The 2-way interaction was not significant.

General Discussion

Across four recognition memory experiments using several different study tasks, we observed consistent voice effects for sentences spoken in old voices. Subjects more accurately recognized "old" test phase sentences when these sentences were presented at test in the same voice as their study presentation. This effect was represented in the accuracy data by a difference of 6.2% to 13.7% between OO sentences and ON sentences in the four conditions. These effects are particularly striking because subjects were never told explicitly that a memory test would follow their initial study of stimuli nor were they told to attend to the talker's voice in any conscious, deliberate way - except in Experiment 4 where subjects were required to categorize the talker's gender as male or female. Thus, we can conclude that rather than being filtered from the speech signal during processing, voice information was encoded and retained in the memory for sentences automatically without the conscious intent of the listener.

New sentences failed to display effects for voice in all except the final experiment that explicitly called subjects' attention to a major component of voice information. As mentioned earlier, this pattern of results can be explained by considering the contribution of novel lexical and semantic content of the new

sentences. Few main nouns or verbs appeared more than once in any of the 80 stimulus sentences. Thus, with five content words in each sentence, subjects need only have noted the novelty of one or two of these words in order to recognize that the sentence was novel. Add to this the novelty of the overall meaning of the sentence that was different for each of the 80 stimulus sentences and subjects were unlikely to mistake a new sentence for an old one no matter what the voice. This may explain the overall bias to respond "new" across all four experiments. All conditions showed a greater accuracy for NN sentences than NO sentences even though this difference was not statistically reliable. These findings additionally support the proposal that listeners encode voice information in speech and that this information can be utilized to aid performance in a recognition memory experiment.

Perhaps the most interesting overall result of this series of experiments is the trend for recognition accuracy and discriminability to decrease while the effect of voice increased. This effect was evident in both new and old sentence judgments although old sentence judgments showed the phenomenon more clearly. From Experiment 1 to 4, judgment accuracy decreased from an OAA of 87.5% to an OAA of 66.4%. Average d' likewise decreased from an average old voice d' of 3.56 in Experiment 1 to an average of 1.54 in Experiment 4 – a drop of more than half across these experiments. While both these scores decreased, the difference in accuracy between old-voiced and new-voiced old sentences increased from 6.2% to 13.7% - a two-fold increase. Thus, while the changes in the study-phase task reduced the ability of subjects to explicitly recognize sentences during test, these same procedures served to increase the implicit effects of voice on sentence discriminability.

Experiment 1 showed accuracy near ceiling levels and can thus be considered the upper limit for subject's ability to recognize the study phase sentences. When subjects were no longer required to transcribe the sentences – and hence were not encouraged to focus on the lexical and semantic content of the sentence – their overall recognition accuracy decreased. Since subjects who did not transcribe sentences would be less able to distinguish old from new sentences as subjects who could rely on lexical/semantic content, they may have relied on voice information to support recognition. This account explains the pattern of results obtained in Experiments 1, 2 and 3 well because voice information would likely play a more significant role in recognition when lexical information was reduced or attenuated. Experiment 4 may have encouraged the least amount of lexical information encoding since subjects needed only to recover talker gender from the stimuli.

When explicit attention is directed to talker gender and hence to correlated voice information in the speech signal, subjects may have encoded a comparatively rich representation of the talker's voice information as compared to encoding of lexical information. This trend in the results suggests a flexible cognitive recognition system that can differentially utilize a variety of information when making judgments depending on the memory data available. The flexibility of this system is also characterized by the potential to process and encode a variety of information about speech including lexical/semantic and voice information.

The present discussion can best be concluded with consideration of the shortcomings of the present experiment and some directions that future experiments can take. First, although the stimulus file creation employed several layers of randomization when picking stimulus set files, the relative distinctiveness/similarity among the talkers and sentences was never formally measured. Thus, we cannot be certain that some stimulus set files were not composed of more or less similar talker and sentence combinations than others. Though a lexical analysis of the sentences for this purpose seems rather extraneous, an analysis of the perceptual similarity of talkers would be worthwhile. An important method for assessing this similarity would be a multi-dimensional scaling (MDS) analysis of talkers from this database. A pooling of subject's similarity judgments for speaking sample stimuli in a forced choice ABX task could generate the data necessary for the MDS analysis. If successful, the MDS analysis could

provide a measure of perceptual similarity between talkers and hence allow us to control this factor in stimulus set creation for future experiments.

A second shortcoming of these experiments resides in the ambiguity in the instructions of what characterized an old sentence for the purposes of recognition. Subjects were simply told that an old sentence was one they heard in the study-phase of the experiment. No subject asked for any further clarification, but the fact remains that different subjects may have taken "heard at study" to mean both the voice and the sentence heard at study. We can be reasonably certain that no subject consistently judged sentences based on an old voice and old sentence criterion since OO and ON sentences differed in recognition accuracy by less than 15% in all conditions.

We cannot, however, rule out the possibility that some subjects inconsistently used an old voice/old sentence criterion during their recognition responses. If subjects made some recognition judgments based only on sentence type and some judgments based on both sentence and voice, then we could expect the small but significant differences between recognition accuracy for OO and ON sentences. We neglected to disambiguate the meaning of old sentences precisely because we could not design a way to communicate this distinction to subjects without calling attention to the fact that in all of the experiments the voices changed between study and test. Future experiments could benefit greatly from modifying these instructions that make the meaning of old sentence clear without calling attention to voice information. An additional safeguard against the ambiguity of the "old" criterion could be the inclusion of a post-experiment questionnaire. Such a questionnaire could probe each subjects understanding of "old" and thus their judgment criterion. Subjects who used an incorrect criterion, such as one that required old voice to be considered "old," could be removed from the analysis. Our continued research in this area will incorporate such a post-experiment questionnaire.

With regard to the stimuli themselves, a richer corpus of sentences, particularly one that contained multiple tokens of the same talker speaking the same sentence, could add to the reliability of the pattern of results and conclusions discussed here. Since the OO sentence stimuli were actually the same stimulus files used in the study phase, one could question whether voice information was being used at all. Subjects could simply have based their recognition on some measure of overall acoustic similarity that would have been high for multiple presentations of the same stimulus file. Many subjects did, however, report that they were aware that some sentences were spoken by different speakers in the test phase. Clearly, at least for these subjects, their memory for sentences from the study phase included specific information about the voice of the talker.

The present series of experiments could also have benefited from additional data such as confidence ratings about the sentence recognition judgments. If subjects were told to rate their confidence that a sentence was old or new as well as the recognition judgment itself, additional converging measures could have been derived from the original recognition response. Additional data could also have been gathered after the test-phase such as subjects' estimates of the number of talkers in the stimulus sets. Such measures could shed light on the individual differences in subject performance and could provide additional data about the accuracy of these estimates and how they may be systematically related to recognition performance across conditions.

The initial study-phase instructions could be altered to create an explicit rather than implicit recognition memory experiment. If subjects were told that a memory test would follow, they may adopt quite different encoding strategies during the study phase that could affect voice information utilization at test. If subjects explicitly rehearsed each sentence during the study phase to improve retention, then surface features such as voice information would likely degrade with successive rehearsals. A series of

experiments in which subjects are given differential amounts of time between stimuli to rehearse could uncover whether this strategy might be used.

Finally, the explicit nature of the present memory experiment could be changed to an implicit memory task such as perceptual identification that does not require conscious explicit recollection. Schacter (1987) has suggested that implicit memory is particularly sensitive to stimulus variability and to the perceptual encoding process. If surface perceptual features do exert more of an effect in an implicit memory task, then we would expect even larger effects of voice than those observed in the present set of experiments. The effects of surface information in sentences are even less explored than the small collection of research on explicit sentence memory. Thus, implicit memory experiments represent an important avenue for future research on the encoding of episodic or instance specific properties of sentences and the talkers that produce them.

While the problems mentioned earlier do call into question any specific conclusions about voice effects and recognition accuracy, the general conclusion that spoken language processing retains talker-specific information for encoding stands. This encoding of extra-linguistic information takes place without the conscious intent of the listener. To the extent that talker voice was never mentioned during any of the four experiments yet subjects still showed consistent voice effects, we find evidence in support of incidental encoding of talker-specific, extra-linguistic information. Additionally, since neither voice information nor a test of memory for the sentences was ever mentioned to subjects, it is unlikely this incidental encoding of voice information is in some way an artifact of the experimental procedure. Thus, it is probably the case that listeners automatically and incidentally encode talker-specific information during "everyday" speech processing.

The retention of rich variation in the memory representation of the speech signal means that perceptual processing does not strip away variation in the speech signal before encoding into memory. Thus, our memories of speech events are not idealized sequences of words or phonemes that preserve the gist of what was said but not the specific variation present in the stimulus. Rather, our memories for speech events preserve at least some of the specific, rich variation present at original encoding. If perceptual processing does not filter variation from the signal before encoding into memory, then we can fairly question whether "variation filtering" of the speech stimulus occurs at all. If we abandon the notion that speech recognition requires, at some stage, an idealized, formal representation of the speech stimulus to begin with, then we need not explain how variation can be preserved for encoding on the one hand, but filtered away during the recognition process on the other. That is, the regularities in the multi-modal information for speech may not be interpreted in the form of idealized units extracted from the variable speech signal.

The process of speech recognition could extract higher-level regularities from the speech signal that depend upon analysis of variation in the signal caused by the special circumstances and characteristics of the talker and listener. While no current theory models speech recognition outside a framework incorporating a formalized, idealized phonemic or segmental stage, it should no longer be taken for granted that speech recognition is a normalizing, abstracting process. The variability in the speech signal may be as important to our understanding and encoding of speech as the regularity of the phonemes and words that have traditionally been considered the nucleus of speech recognition.

References

Abercrombie, D. (1967) *Elements of General Phonetics*. Chicago, IL: Aldine.

Bradlow, A.R., Torretta, G. M., & Pisoni, D. B. (1995) Intelligibility of Normal Speech I: Global and Fine-Grained Acoustic-Phonetic Talker Characteristics. In *Research on Spoken Language Processing Progress Report No. 20* (pp. 89-116). Bloomington, IN: Speech Research Laboratory, Indiana University.

Cornell, E. H. (1980) Distributed study facilitates infants' delayed recognition memory. *Memory and Cognition*. Vol 8(6): 539-542.

Craik, F., & Lockhart (1972) Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684

Creelman, C.D. (1957) The case of the unknown talker. *Journal of the Acoustical Society of America*, 29, 655.

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

Fisher, R. P. & Cuervo, A. (1983). Memory for physical features of discourse as a function of their relevance, *Journal of Experimental Psychology: Learning, Memory & Cognition*, 9, 130-138

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3-28.

Gaver, W.W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5(1),* 1-29.

Geiselman, R. E. & Bellezza, F. S. (1976) Incidental retention of speaker's voice. *Memory and Cognition*, 5, 658-665

Geiselman, R. E. & Bellezza, F. S. (1977) Long term memory for speaker's voice and source location. *Memory and Cognition*, 4, 483-489

Gerstman, L.J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16, 1,* 78-80.

Goldinger, S.D., Pisoni, D.B. & Logan, J. S. (1991) On the locus of talker variability effects in the recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17 (1), 152-162.*

Goldinger, S. D. (1992) Words and voices: Implicit and explicit memory for spoken words. *Research on Speech Perception Technical Report No. 7.* Indiana University: Bloomington, IN.

Goldinger, S.D., Pisoni, D.B. & Luce, P.A. (1996). Speech perception and spoken word recognition: Research and theory. In N.J. Lass (Ed.), *Principles of Experimental Phonetics*. Toronto: B.C. Decker, Inc. Pp. 277-327

Halle, M. (1956) For Roman Jakobson: essays on the occasion of his sixtieth birthday, 11, Oct 1956. The Hague: Mouton.

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., and Dolan, R.J. (1999) Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study. *Journal of Neuroscience*,19(10), 3962-3972.

Karl, J.R. & Pisoni, D. B. (1994) Effects of stimulus variability on the recall of spoken sentences: A first report. In *Research on Spoken Language Processing Progress Report No. 19* (pp. 145-194). Bloomington, IN: Speech Research Laboratory, Indiana University.

Karl, J. R. (1996).The role of speakers' voice in listeners' recall of spoken sentences. Unpublished Masters Thesis.

Kim, J. J., Andreasen, N. C., O'Leary, D. S, Wiser, A. K., Ponto, L. L., Boles, Watkins, G. L. & Hichwa, R. D. (1999) Direct comparison of the neural substrates of recognition memory for words and faces. *Brain*. 122, 1069-1083.

Klatt, D. H. (1986) The problem of variability in speech recognition and models of speech perception. In Y. Tohkura, E., Vatikiosis-Bateson & Y. Sagisaka (eds*.) Speech Perception, Production, and Linguistic Structure*. Tokyo: IOS Press, pp. 300-324.

Klatt, D. H. (1989) Review of selected models of speech perception. In W. D. Marslen-Wilson, (Ed.), *Lexical representations and process*. Cambridge, MA: MIT Press (pp. 169-226).

Kuhl, P. K. & Miller, J. D. (1982) Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics*, 31, 279-292.

Labov, W. (1963) The social motivation of sound change. *Word*, 19, 273-309.

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

Liberman, A.M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America*, 29, 117-123.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-461.

Licklider, J. C. R. (1952) On the process of speech perception. *Journal of the Acoustical Society of America*, 24, 590-594.

Lively S.E., Pisoni, D.B., Yamada R.A., Tohkura, Y. & Yamada, T. (1992) Training Japanese listeners to identify English [r] and [l]: III. Long-term retention of the new phonetic categories. *Research on Speech Perception*, Progress Report No. 18. (pp. 185-216) Indiana University, Bloomington, IN.

Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*, 839-862.

Lively, S.E., Pisoni, D.B. & Goldinger, S.D. (1994). Spoken word recognition: Research and Theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, New York: Academic Press. (pp. 265-301).

Logan, J.S. & Pisoni, D.B. (1987) Talker variability and the recall of spoken word lists: A replication and extension. In *Research on Spoken Language Processing Progress Report No. 13* (pp. 307-328). Bloomington, IN: Speech Research Laboratory, Indiana University.

Logan J. S., Lively, S. E. & Pisoni D. B. (1991) Training Japanese listeners to identify the English [r] and [l]: A first report. *Journal of the Acoustical Society of America, 89 (2),* 874-886.

Mattingly, I.G. & Liberman, A.M. (1990). Speech and other auditory modules. In G.M. Edelman, W.E. Gall & W.M. Cowan (Eds.), *Signal and Sense: Local and Global Order in Perceptual Maps*. New York: Wiley. 501-520.

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

Miller, G. A. (1956) The perception of speech. *In Language and Communication*. New York: McGraw-Hill, pp. 47-49.

Mullennix, J. W. & Pisoni, D. B. (1990) Stimulus variability and processing dependencies in speech perception. Perception and Psychophysics, 47, 349-390.

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.

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

Nygaard, L.C., Sommers, M.S. & Pisoni, D.B. (1992b). Effects of speaking rate and talker variability on the recall of spoken words. *Journal of the Acoustical Society of America, 91 (4),* 2340.

Oldfield, R.C. (1966). Things, words and the brain. *Quarterly Journal of Experimental Psychology, 18,* 340-353.

Peters, R. W. (1955b) The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise. *Joint Project Report No. 56 U.S. Navel School of Aviation Medicine,* Pensacola, FL (pp. 1-9).

Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

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.

Pisoni, D.B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. *Proceedings of the 1990 International Conference on Spoken Language Processing,* Kobe, Japan, pp. 1399-1407.

Pisoni, D.B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication, 13*, 109-125.

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

Pisoni, D.B. & Lively, S.E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech Perception and Linguistic Experience*. Pp. 433-459. Baltimore: York Press.

Pisoni, D.B. & Luce, P. A. (1987) Acoustic-Phonetic representations in word recognition. *Cognition*, 25, 21-52.

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

Sachs, J.D.S. (1967) Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics*, 2, 437-442.

Salasso, A., & Pisoni, D. B. (1985) Interaction of knowledge sources in spoken word recognition. *Journal of Memory and Language*, 24, 210-231.

Schacter, D. L. (1987) Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13, 501-518.

Schacter, Daniel, L. (1999) The seven sins of memory. *American Psychologist*, 54 182-203.

Snodgrass, J. G. & McClure, P. (1975) Storage and retrieval properties of dual codes for pictures and words in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*.1(5), 521-529.

Squire, L. R., Shimamura, A.P., and Graf, P. (1985) Independence of recognition memory and priming effects: A neuropsychological analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 11(1), 37-44.

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. (1983) Audio-visual speech perception, lipreading, and artificial stimulation. *Hearing and Science Disorders*, London: Academic Press.

Summerfield, Q. (1987) Some Preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.) Hearing by Eye: The Psychology of Lip-reading, Hillsdale, NJ: LEA.

Summerfield, Q. & Haggard, M. P. Vocal tract normalization as demonstrated by reaction times. *Report of Speech Research in Progress*, 2(2) Queens University of Belfast (pp. 12-23).

Studdert-Kennedy, M. & Shankweiler, D. P. (1970) Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America, 48*, 579-594.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Audio-Visual Perception of Sinewave Speech in an Adult Cochlear Implant User: A Case Study[1]

Winston D. Goh,[2] David B. Pisoni,[3] Karen I. Kirk,[3] and Robert E. Remez[4]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This research was supported by NIH-NIDCD Research Grant DC00111 to Indiana University, NIDCD K23 Research Grant DC00126 to Indiana University School of Medicine, and NIDCD Research Grant DC00308 to Barnard College. We would like to thank Stacey Yount for gathering information on the CI patients' test scores, and Lorin Lachs for helpful comments during the preparation of this article.
[2] Also, Department of Social Work & Psychology, National University of Singapore.
[3] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.
[4] Department of Psychology, Barnard College, New York, NY.

Audio-Visual Perception of Sinewave Speech in an Adult Cochlear Implant User: A Case Study

Abstract. We investigated a post-lingually defeaned cochlear implant user's ability to perceive sinewave replicas of spoken sentences. The patient, Mr. S, transcribed sinewave sentences under audio-only (AO), visual-only (VO), and audio-visual (A+V) conditions. His performance was compared to the data collected from a group of normal-hearing participants in an earlier study by R.E. Remez, J.M. Fellowes, D.B. Pisoni, W.D. Goh, and P.E. Rubin (1998). Results showed that Mr. S derived a larger gain from additional visual information provided by the talker's face than the normal-hearing controls. The increase in performance under A+V presentation reflected the superior lip-reading skills that this patient displayed and his ability to use this skill to integrate the information provided by the talking face and the sinewave speech to perceive the underlying sentence. Implications of these findings for multimodal phonetic coherence in speech perception are discussed.

It has long been known that combining audio and visual information facilitates the perception of speech. In their pioneering study, Sumby and Pollack (1954) demonstrated that the intelligibility of spoken words can be enhanced by as much as +15 dB in noisy environments if listeners are able to see the talker's face. This is a substantial gain in performance that surpasses even the best hearing aid devices. Using visual information from a dynamically articulating face for phonetic and lexical identification is a skill which almost everyone will benefit from when listening in noisy environments, especially when people get older and their hearing deteriorates (Summerfield, 1987). For the hearing-impaired population, the visual route may play an especially major role in speech perception. Some of the speech cues for consonants that are difficult to hear are easy to see and vice-versa (Walden, Prosek, Montgomery, Scherr, & Jones, 1975). For example, /f/ and /θ/ are auditorily confusable, but they are very distinct visually when the talker's articulatory movements can be seen. The enormous gain from seeing the talker's face is eloquently captured by a question frequently asked of hearing-aid practitioners – "Doctor, why can I understand you so much more clearly when I wear my glasses?" (Summerfield, 1987). Research into the nature of audio-visual integration in speech perception can therefore provide substantial insights and applications for rehabilitative procedures, techniques, and training methods to assist the hearing impaired. The study of multimodal speech perception also raises many important theoretical issues about the scope and domain of current models of speech perception and spoken language processing (see Berstein, Demorest, & Tucker, in press; Massaro, 1998).

The absolute gain in performance observed from the visual aspects of speech is highest in a noisy environment or in other conditions that make auditory perception difficult (Sumby & Pollack, 1954). Therefore, the best way to observe the influence of visual information is to look at identification performance with impoverished auditory stimuli. The traditional way of studying this problem in the past was to manipulate the signal-to-noise ratio (SNR) for the environment in which the speech stimuli is presented. Another way involved reducing the amount of information that is normally available in the auditory speech waveform. One such technique is to use sinewave speech instead of natural speech (Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez, Rubin, Pisoni, & Carrell, 1981).

In sinewave speech, time-varying sinusoidal waveforms are generated by a digital synthesizer to match the LPC-derived center frequencies and amplitudes of the formants in the natural utterance. The synthetic sinewave pattern preserves the dynamics of frequency and amplitude variations observed in natural speech over time, but differs from natural speech in several important ways. There are no

harmonics, broadband formant structures, formant frequency transitions, steady-state formants, or changes in fundamental frequency. In short, sinewave speech patterns contain none of the "traditional" speech cues that are assumed to form the basis of speech perception – e.g., formant frequency transitions that cue manner and place of articulation (see Remez et al. 1981).

Despite the unnatural characteristics of sinewave speech, these sound patterns are still intelligible (Remez et al., 1981; 1994). The absence of traditional acoustic cues for phonetic perception implies that sinusoidal replicas of speech should be perceived as independently changing tones and not as an integrated, linguistic percept. However, listeners are still able to extract the phonetic and lexical properties of the utterance from the highly impoverished, skeletal representation of the natural token that is preserved in the sinewave replica. This result suggests that sufficient phonetic information is still encoded in the relational and time-varying structure that is represented in the sinewave pattern, even though the synthetic waveform is obviously not producible by a vocal tract. Sinewave speech perception also shows the multimodal facilitation observed for natural speech (Remez, Fellowes, Pisoni, Goh, & Rubin, 1998). A considerable increase in identification performance was found when the sinewave patterns are presented in an audio-visual context compared to an audio-only context.

Previous studies on sinewave speech perception have so far used only participants who have normal hearing at the time of testing. Since audio-visual speech perception may be even more critical for people who have hearing impairment, it is important to begin investigations into how members of this clinical population perceive sinewave speech. In particular, how would hearing-impaired individuals fitted with a cochlear implant (CI) fare in listening to sinewave speech under different presentation conditions? We are especially interested in patients who perform very well with their CI and who demonstrate the ability to use visual information to mitigate their hearing impairment. Would such users be able to integrate visual information with very unnatural auditory patterns? In this paper, we report the performance of one patient, Mr. S, in transcribing sinewave sentences under audio-only (AO), visual-only (VO), and audio-visual (A+V) conditions and then compare his performance to a group of normal-hearing participants whose data was collected by Remez et al. (1998).

Patient Background

Our patient, Mr. S, is a 35-year-old Caucasian male with a graduate degree. He has a profound hearing loss due to cryoglobial anemia and autoimmune syndrome. Onset of his deafness occurred in 1993 when he was 29. His hearing impairment was diagnosed as a profound loss a year later and he was implanted with a Clarion 8-channel CI in 1995. He has been using the CI for the past 4 years and is considered to be an exceptionally good user by the clinical staff. We will now describe Mr. S's performance on a battery of standard clinical tests that were collected in 1998. All tests were conducted while he was using the CI.

The Iowa Consonant Test (Tyler, Preece, & Tye-Murray, 1983) is a closed-set test of consonant recognition in which the listener is familiarized with 16 different consonants in the /aCa/ format. The listener is then asked to identify the consonant he hears out of a choice of 16 alternatives. Chance performance for consonant identification is approximately 6%. This test can also be analyzed in terms of the listener's ability to identify phonetic features. Chance performance for consonant voicing, manner, and place of articulation identification is 50%, 33%, and 20% respectively. On the Iowa Consonant Test, Mr. S achieved a total score of 79% correct, 96% on voicing, 94% on manner, and 85% on place.

The CUNY Sentences Test (Boothroyd, Hannin, & Hnath, 1985) is an open-set sentence recognition task in which the listener is presented with sentences in three listening conditions: AO, VO, and A+V. The test is scored in terms of the total number of words correctly identified. On this test, Mr. S

obtained a perfect score of 100% in the A+V condition, 92% in the AO condition, and 63% in the VO condition.

Table 1 compares Mr. S's scores on these tests and the average scores of 28 other CI patients. Table 1 also lists the performance of Mr. S and the other CI patients in a recent study (Kaiser, Kirk, Pisoni, & Lachs, 2000) that tested the participants' ability to identify isolated consonant-vowel-consonant (CVC) words from the Hoosier audiovisual multi-talker database (Lachs & Hernandez, 1998; Sheffert, Lachs, & Hernandez, 1997) under AO, VO, and A+V presentations, using both single-talker and multiple-talker presentation conditions. Generally, Mr. S's performance on these speech perception tests indicate that he is able to perceive speech without any content cues in a controlled test environment. It is clear that Mr. S is an exceptionally good implant user relative to the other CI patients. His lip-reading performance is consistently at least two standard deviations higher than the average CI patient, as shown in his scores for the various tests in the VO conditions in Table 1.

	Mr. S	Other CI Patients (N = 28)	
		M	*SD*
Iowa Consonant Test*			
Total	**79**	**45.0**	17.3
Voicing	**96**	**88.7**	12.9
Manner	**94**	**67.1**	15.7
Place	**85**	**52.3**	16.3
CUNY Sentences Test**			
Audio-only (AO)	**92**	**55.0**	29.6
Visual-only (VO)	**63**	**24.3**	15.6
Audio-visual (A+V)	**100**	**91.2**	9.1
Indiana Multi-talker Isolated CVC Words***			
Single talker condition			
Audio-only (AO)	**67.0**	**30.1**	19.1
Visual-only (VO)	**30.5**	**15.7**	5.1
Audio-visual (A+V)	**88.5**	**69.0**	13.8
Multiple talker condition			
Audio-only (AO)	**55.5**	**30.00**	16.57
Visual-only (VO)	**30.5**	**14.79**	7.07
Audio-visual (A+V)	**86.0**	**60.03**	14.60

Table 1. Comparison of Mr. S's percent correct scores and the average percent correct scores of other CI patients' on several speech perception tests.

* Tyler, Preece, and Tye-Murray (1983).

** Boothroyd, Hannin, and Hnath (1985).

*** From Kaiser, Kirk, Pisoni, and Lachs (2000).

Mr. S also achieved an auditory digit-span score of 10 on the WISC-forward and 8 on the WISC-backward collected under AO conditions. On the word familiarity test (FAM; Lewellen, Goldinger, Pisoni, & Greene, 1993), which indexes subjective familiarity with words of varying frequencies, he had a mean FAM score of 3.57 on low frequency words, 6.39 on mid-frequency words, and 6.85 on high

frequency words. These scores are comparable to the average scores obtained for normal-hearing, high-vocabulary participants as described in Lewellen et al. (1993). His performance on the WISC digit-span and FAM tests indicate that his short-term memory capacity and word familiarity are comparable to normal-hearing subjects.

Method

Participants

The normal-hearing participants whose data we used as a comparison group consisted of 25 young adults from the Indiana University community. These participants were a subset of the sample that participated in the study described in Remez et al. (1998). All participants were native speakers of English and reported normal hearing and normal vision or corrected-to-normal vision at the time of testing. None of the participants had any previous exposure or familiarity with sinewave analogs of speech signals. All participants were students enrolled in Introductory Psychology classes. They received either course credit for participation or they were paid as a volunteer. Our patient, Mr. S, was paid as a volunteer. He also had no prior experience with sinewave speech before the present tests and he had corrected-to-normal vision. He was tested on the sinewave speech in August 1999.

Apparatus and Materials

The 18 sentences used in the present study were obtained from the database developed by Remez et al. (1998) and are listed in the Appendix. The original sentences were recorded and digitized and then an expert phonetician analyzed the sampled data to estimate the formant center frequencies and amplitudes. Formant center frequencies were obtained by comparing discrete Fourier spectra and linear prediction estimates. The synthesis parameters were created by tracing the formant patterns over time. The fundamental frequency of phonation was estimated from a narrow-band Fourier representation of the natural spectra. A software synthesizer was then used to convert the frequency and amplitude values taken at 10 msec intervals for F0, F1, F2, F3 and fricative formants to time-varying sinusoids (Rubin, 1980). The sinewave replicas of each sentence were composed of tone analogs of the three oral formants. A fourth tone was used to reproduce fricative formants when these were present and discontinuous with the oral formants.

The first 8 sentences were spoken by one of the authors (RER), and were used for the familiarization phase and AO condition. The other 10 sentences were spoken by an adult female speaker whose natural speech intelligibility had been verified by other normal-hearing volunteer participants as acoustically intelligible (see Bradlow, Torretta, & Pisoni, 1996). The female speaker's sentences were used in the VO and A+V conditions. The sinewave patterns for her sentences were combined and synchronized with the video clips using Abobe Premiere 4.2. All stimulus materials were presented to participants via a Macintosh Quadra 950 machine with a Targa 2000 video card. This system presented the 14-bit color video samples at 30 frames per second in full-screen mode at 640x480 resolution on a 17-inch monitor.

Design and Procedure

The 18-sentence presentation sequence was fixed for all participants. All normal-hearing participants listened to the audio track via a pair of Beyer Dynamic DT100 headphones. The audio stimuli were presented at approximately 75 dB SPL. Our patient, Mr. S, listened to the audio track via a set of Labtec LS-1020 desktop computer speakers with his CI turned on. Prior to the start of the experiment, we

calibrated the output amplitude of the speakers to a signal level where he could correctly identify five auditorily presented, naturally spoken CVC words in a row.

The sequence of testing for Mr. S was as follows. The first three sentences were used as a familiarization sequence to acclimatize him to the unnatural timbre of the sinewave sentences. These materials were presented audio-only and the sentences were already transcribed for him on the answer sheet. Each sentence was repeated five times with 10 seconds between repetitions and 20 seconds between sentence blocks. A warning tone occurred before the start of a new sentence block. The next five sentences followed the same procedure and comprised the AO condition. Our patient was asked to transcribe these sentences while he listened. For the familiarization phase and the AO condition, the video monitor remained blank while the audio signals were played out via the speakers. The control participants in Remez et al. (1998) followed precisely this same procedure except that the signals were presented over individual headphones and the participants were run in groups of five or smaller.

In Remez et al. (1998), the VO and A+V conditions were run between-subjects using all ten of the female talker's sentences. It was obviously not possible to follow this same procedure with our patient so several changes were made in the presentation format. For our CI patient, after the AO condition, the first five sentences of the female talker were presented in the VO condition, followed by the last five sentences of that same talker in the A+V condition. The CI patient and the normal-hearing control participants were instructed to look at the video monitor and write down their responses only during the intervals between repetitions and sentence blocks.

Scoring

The number of syllables correctly transcribed was used as the dependent measure for both the CI patient and the normal-hearing control group. Because there were some small procedural differences between the Remez et al. (1998) data collection and the session with our CI patient, we had to ensure that the control data obtained from the earlier study was a valid comparison to make. For the control participants who were assigned to the VO condition in Remez et al., we only scored their responses for the *first* five sentences, since these were exactly the same sentences presented to the CI patient in the VO condition. Conversely, for the control participants assigned to the A+V condition in Remez et al., we only scored their responses for the *last* five sentences, since these were the same sentences used in the CI patient's A+V condition. All five sentences were scored in the AO transcription of RER's sentences because these sentences were identical for both the CI patient and the participants in Remez et al. These scoring procedures ensured that appropriate comparisons could be made between our patient's responses and those obtained from the normal-hearing participants.

Results

The transcription performance of our CI patient, Mr. S, and the relevant data from the normal-hearing participants from Remez et al. (1998) are summarized in Table 2. The results show that the CI patient's AO performance (53%) was not as good as the average of the normal-hearing controls (65%). It should be emphasized here that the CI patient's performance is, however, within one standard deviation of the mean of the normal-hearing controls' performance. It is very likely that the inherent difficulty in perceiving sinewave speech is further compounded by reliance on a CI device for speech perception. Normal-hearing participants would not have as much difficulty because of their intact hearing abilities. The average performance of the normal-hearing controls displayed here is very similar to the previous results reported for AO listening conditions (Remez et al., 1981).

	Audio-only (AO)	Visual-only (VO)	Audio-visual (A+V)
Mr. S	**52.5**	**43.2**	**89.7**
Normal-hearing*			
M	**64.7**	**18.0**	**85.6**
SD	24.4	10.6	13.5
N	25	14	11

Table 2. Average percentage of correct syllables for Mr. S and normal-hearing listeners for the three presentation conditions.
* from Remez, Fellowes, Pisoni, Goh, and Rubin (1998).

For the VO condition, it is clear that Mr. S showed superior performance (43%) relative to the normal-hearing controls (18%). His performance is more than two standard deviations from the mean of the normal-hearing listeners. This is probably due to the enhanced lip-reading abilities that are typical of the hearing-impaired population (Summerfield, 1987). Of greater interest to us, however, is his performance in the A+V condition. With the addition of visual information, our CI patient's performance (90%) is slightly above the average performance of the normal-hearing controls (86%). This patient is clearly able to use and integrate information from the dynamically changing articulators in the video display with the auditory information, despite the unspeechlike qualities of the auditory sinewave speech patterns. The additional visual information drives his performance up to levels that are comparable to the average performance of a group of normal-hearing controls.

One of the more interesting questions about Mr. S's performance deals with assessing the contribution of the visual information relative to the possible information available in the absence of visual stimulation (see for example Sumby & Pollack, 1954). The actual contribution from the visual modality is the difference in scores between the A+V and AO conditions. The possible available information is the difference between the total information possible and the performance in the AO conditions, i.e. 100% minus the AO performance. The ratio of the actual contribution to the possible available information is the amount of gain obtained from seeing the talker's face, and can be considered a measure of visual enhancement. This ratio normalizes for absolute differences in AO performance and is formalized below:

$$R = (\text{A+V} - \text{AO}) / (100 - \text{AO})$$

Using this metric, our patient displays a 78.3% gain from seeing the talker's face, compared to a 59.2% gain for the mean of the normal-hearing controls. This difference in visual enhancement suggests that Mr. S is deriving a much larger benefit from the visual information than the normal-hearing controls to achieve his exceptionally high level of performance in the A+V condition. This level of performance is clearly not due to his superior lip-reading ability alone, since he was only performing at 43% accuracy in the VO condition, although this ability is probably an influential factor. The improvement observed under A+V presentation is due to his ability to use and integrate the information provided through the visual modality with the cues provided by the time-varying dynamics of the tone analogs to perceive the underlying sentences.

Discussion

If sinewave replicas can be perceived as speech even though a natural source cannot be attributed to the signal, the results imply that the perceptual organization of speech depends on the establishment of coherence among dissimilar sensory elements (Remez et al., 1994) – in the case of sinewave speech, it is

the coherence of independently changing tones. In multimodal contexts, the visual perception of a dynamically articulating face combined with the auditory perception of the very unspeechlike quality of a sinewave signal makes the event especially incoherent. In other words, no natural talker can be made to appear to an observer as the original source of the sinewave speech. Thus, multimodal presentation provides a strong test of phonetic coherence despite the incoherence of the visual and auditory information as an integrated perceptual event. In Remez et al. (1998), the combination of the video signal plus an auditory track containing a single tone analog of the F2 formant was the only single-tone condition that showed a facilitation over the VO condition. The interpretation of this result by Remez et al. was that both the F2 sinewave analog and the visual information provided congruent cues regarding the underlying gestures for place of articulation, thus producing multimodal phonetic coherence of the perceptual event from two separate and independent sensory inputs.

In this study, multimodal perception of sinewave speech was investigated for the first time in a CI user. Although Mr. S did not perform as well as the normal-hearing controls in the AO condition, his performance was comparable or slightly better than the normal-hearing controls in the A+V condition. Clearly, the additional complementary visual information in the latter condition allowed both Mr. S and normal-hearing participants to better perceive the sinewave sentences. More importantly, however, Mr. S obtained a much larger benefit from seeing the talker's face than the normal-hearing controls (78% gain versus 59% gain, respectively). This pattern is consistent with previous findings showing that the visual modality plays a large role in the speech perception of the hearing-impaired population (Summerfield, 1987). It is also important to emphasize here that the task used in the present study was to transcribe highly impoverished sinewave speech patterns, not normal or natural speech samples. Mr. S's ability to simultaneously integrate visual information with the sinewave tracks of the auditory signal is impressive and suggests that multimodal phonetic coherence can occur even when a profoundly hearing-impaired listener perceives highly impoverished sound patterns through a CI device. These findings from Mr. S, a patient with a CI, provide additional support for the original proposal of Remez et al. (1981) – speech perception can occur without traditional speech cues. Elements of speech perception from multiple sensory modalities depend on the establishment and preservation of perceptual coherence among individual elements and attributes at a more abstract level that reflects the underlying common source of the speech signal, as represented in the talker's articulatory gestures (Remez et al., 1994; 1998).

We are confident that this exploratory case study will spur future research on sinewave speech perception in the hearing-impaired and other clinical populations such as the elderly and language-delayed children. The use of sinewave speech patterns with these populations will provide useful converging evidence about the similarities and differences in the nature of speech perception processes in different populations who have known and well-documented sensory, perceptual and cognitive impairments in the ability to encode and perceive sensory input. Data from these listeners should therefore provide valuable new insights into the multimodal organization of speech and spoken language processing.

References

Bernstein, L.E., Demorest, M.E., & Tucker, P.E. (in press). Speech perception without hearing. *Perception & Psychophysics*.

Boothroyd, A., Hannin, L., & Hnath, T. (1985). A sentence test of speech perception: Reliability set equivalence and short term learning (internal report RCI 10). New York: City University of New York.

Bradlow, A.B., 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.

Kaiser, A., Kirk, K.I., Pisoni, D.B., & Lachs, L. (2000). Audiovisual speech integration in adults with cochlear implants or normal hearing: Lexical and talker effects. Paper to be presented at the ARO midwinter meeting, St. Petersburg Beach FL, February 20-24 2000.

Lachs, L., & Hernàndez, 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.

Lewellen, M.J., Goldinger, S.D., Pisoni, D.B., & Greene, B.G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General, 122,* 316-330.

Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.

Remez, R.E., Fellowes, J.M., Pisoni, D.B., Goh, W.D., & Rubin, P.E. (1998). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication, 26,* 65-73.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review, 101,* 129-156.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947-950.

Rubin, P.E. (1980). Sinewave synthesis. Internal memorandum. New Haven: Haskins Laboratories.

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.

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: Erlbaum.

Tyler, R.S., Preece, J., & Tye-Murray, N. (1983). *The Iowa cochlear implant tests*. Iowa City: Department of Otolaryngology - Head and Neck Surgery, University of Iowa.

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.

Appendix

Sentence Materials (adapted from Remez et al., 1998)

Familiarization

The bill was paid every third week.
The soft cushion broke the man's fall.
Two blue fish swam in the tank.

Audio-only (AO)

A small creek cut across the field.
The fruit peel was cut in six slices.
Her purse was filled with useless trash.
The stray cat gave birth to kittens.
Where were you a year ago?

Visual-only (VO)

Always close the barn door tight.
This is a grand season for hikes on the road.
He ran halfway to the hardware store.
Kick the ball straight and follow through.
The term ended in late June that year.

Audio-visual (A+V)

Use a pencil to write the first draft.
Cut the pie into large parts.
The boy was there when the sun rose.
A cup of sugar makes sweet fudge.
What joy there is in living.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Sublexical Influences on Lexical Development in Children[1]

Holly L. Storkel[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

[1] This work was supported by NIH-NIDCD DC00012 and DC01694 to Indiana University, Bloomington. Thanks to Judith Gierut for comments regarding the design of this study and to Michael Vitevitch for comments on an earlier version of this manuscript.
[2] Also, Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN 47405.

Sublexical Influences on Lexical Development in Children

Abstract. Previous experimental research has found that adults and infants are sensitive to the likelihood of occurrence of sequences of segments, or probabilistic phonotactics, in the ambient language. One hypothesis that emerges from this finding is that probabilistic phonotactics, a sublexical factor, may influence rate of lexical acquisition. Preliminary results are reported from a study involving 21 typically developing preschool children. Children participated in a multi-trial word learning task involving eight nonwords of varying phonotactic probability. Each nonword was paired with a picture of an unusual object having no apparent corresponding label in English. Referent and item identification tasks were used to monitor lexical acquisition during learning and retention trials. Results indicated that high probability nonwords were learned with fewer exposures than low probability nonwords across both test measures. This finding suggests that sublexical representations influence lexical development in children.

Introduction

Children rapidly acquire a large number of words without the benefit of direct instruction (Dollaghan, 1985; Heibeck & Markman, 1987; Jusczyk & Aslin, 1995; Rice & Woodsmall, 1988). This ability has been termed *fast mapping* (Carey & Bartlett, 1978) or *quick incidental learning* (QUIL; Rice, 1990). Researchers have proposed a number of underlying mechanisms or language learning devices to account for this robust ability to learn novel words. Past proposals have focused on pragmatic or syntactic cues that children may use to identify the meaning of a novel word (Akhtar & Tomasello, 1996; Baldwin, 1991, 1993a, 1993b; Brown, 1957; Soja, 1992; Tomasello & Barton, 1994; Waxman & Kosowski, 1990). Alternatively, constraints or biases have been proposed to narrow possible semantic interpretations of newly encountered words. Proposed constraints include whole-object, taxonomic, and mutual exclusivity (Golinkoff, Mervis, & Hirsh-Pasek, 1994; Markman, 1989, 1994; Markman & Hutchinson, 1984; Markman & Wachtel, 1988; Merriman & Bowman, 1989; Waxman & Kosowski, 1990). These accounts all focus on how children determine the meaning or referent of a novel word. Less attention has been devoted to identifying the processes that influence acquisition of the form of a novel word. It is possible that sublexical, or phonological, factors may affect lexical acquisition. In particular, the probability of the phonological sequence may facilitate acquisition of the novel word. We first consider the role of phonotactic probability in language processing generally, and then explore the evidence supporting the influence of sublexical factors on lexical acquisition.

Phonotactic Probability and Language Processing

One observation that has emerged from the study of language structure is that certain sequences of segments are more likely to occur. Experimental psycholinguistic research has found that adult speakers are sensitive to this likelihood of occurrence or *phonotactic probability* (Vitevitch & Luce, 1998, 1999; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997). Adults recognize high probability nonwords faster than low probability nonwords (Vitevitch & Luce, 1998, 1999). This finding suggests that adults learn the likelihood of occurrence of sounds in the ambient language and that high phonotactic probability may play a role in the recognition of spoken words. The facilitory effect of high phonotactic probability appears to be dependent on the lexicality of the stimuli and the task. For example, an inhibitory effect of high phonotactic probability has been documented in recognition tasks involving real words and in lexical decision tasks using both real words and nonwords (Vitevitch & Luce, 1998, 1999). To account for the stimuli and task dependent nature of this effect, two types of representations have been proposed: sublexical and lexical (Vitevitch & Luce, 1998, 1999). Sublexical representations presumably contain information related to

phonotactic structure. For this type of representation, it is assumed that high phonotactic probability facilitates processing because common sound patterns are more readily activated. In contrast, lexical representations are seemingly organized into neighborhoods based on similarity of form (Luce & Pisoni, 1998). For this type of representation, it is hypothesized that high phonotactic probability inhibits processing because of the correlation between phonotactic probability and neighborhood density (Vitevitch, Luce, Pisoni, & Auer, 1999). Specifically, words composed of common sound patterns will tend to be similar to many other lexical entries. In this way, words composed of high probability phonological sequences will tend to reside in *high density neighborhoods*. This high density may create competition among lexical items. In any given task, it is thought that one type of representation will dominate processing leading to the inhibitory or facilitory patterns observed. For instance, nonwords are assumed to lack lexical representations leading to sublexical dominance in processing. Under these conditions, facilitory effects of high phonotactic probability have been reported (Vitevitch & Luce, 1998, 1999). The evidence from experimental psycholinguistic research seems to indicate that adult speakers represent phonotactic probability. Furthermore, the phonotactic probability of a word or nonword may influence performance when sublexical representations dominate processing.

Infants, like adults, appear sensitive to the phonotactic probabilities (density) in the ambient language. Jusczyk, Luce, and Charles-Luce (1994) demonstrated that 9-month-old infants listen longer to high probability (density) than low probability (density) sound sequences. This preference indicates that infants learn the distribution of sounds in the ambient language and may have form based representations of words without semantic knowledge (Jusczyk et al., 1994). The implication of this preference for language processing and learning is unclear. One possibility is that phonotactic probability (density) aids the infant in determining which sound sequences are likely to form words in their native language (Jusczyk et al., 1994; Luce & Pisoni, 1998). This hypothesis suggests the high probability (density) preference is indicative of sublexical facilitation of language processing and learning. A second possibility is that infants listen longer to high probability (density) sound sequences because these sequences are more difficult to discriminate from other sound sequences. This second hypothesis assumes that the high probability (density) preference results from lexical inhibition of language processing and learning. At issue here is the dominant type of representation in this listening task. Taken together, it appears that 9-month-old infants represent phonotactic probability (density), but the locus of the effect of phonotactic probability (density) and the influence on language learning are unclear in this group. Given the difficulties in investigating learning in young infants, evidence from older infants and children may provide insights into the representational structure of young language learners.

Research with older children suggests the presence of both sublexical and lexical representations (Gathercole, Frankish, Pickering, & Peaker, 1999; Kirk, Pisoni, & Osberger, 1995; Messer, 1967; Pertz & Bever, 1975). The presence of lexical representations can be inferred from the finding that hearing-impaired children are less accurate recognizing high probability (density) than low probability (density) words (Kirk et al., 1995). This finding indicates an inhibitory effect of high probability (density) typically associated with lexical dominance in processing. The presence of sublexical representations is supported by the findings from metalinguistic and memory research. Children and adolescents show metalinguistic awareness of sublexical structure (Messer, 1967; Pertz & Bever, 1975). In terms of memory, children recall more high probability (density) than low probability (density) nonwords (Gathercole et al., 1999). The finding of a facilitory effect of high probability (density) for nonwords supports the conclusion that sublexical representations presumably dominate processing in this memory task. These results appear to indicate that young children may have two types of representations and that different representations may dominate processing depending on the task and stimuli used. However, the dominant representation in lexical acquisition remains unclear.

Sublexical Influences on Lexical Acquisition

Given that a novel word has not been encountered previously, it can be considered similar to a nonword. Based on this assumption, novel words presumably do not have a lexical representation and sublexical processing may dominate tasks involving novel words. As a result, one would expect sublexical representations to influence word learning leading to a facilitory effect of high probability (density). Past research supports the hypothesis that sublexical factors may influence lexical acquisition (e.g., Leonard, Schwartz, Morris, & Chapman, 1981; Schwarz & Leonard, 1982; Storkel & Rogers, in press).

Research related to sublexical influences on lexical acquisition typically has examined the effect of phonological development on lexical acquisition. The question addressed is whether or not the child's *phonological inventory*, the sounds the child produces, influences lexical acquisition. Correlational studies indicate that the phonological characteristics of infant babbling are highly similar to the phonological characteristics of the child's first spoken words (Oller, Wieman, Doyle, & Ross, 1975; Vihman, Ferguson, & Elbert, 1987). This same association between the lexicon and the phonological inventory has been documented in children with advanced language development and children with delayed language development (Paul & Jennings, 1992; Stoel-Gammon & Dale, 1988; Thal, Oroz, & McCaw, 1995; Whitehurst, Smith, Fischel, Arnold, & Lonigan, 1991). In addition, experimental studies indicate that children more readily learn to *produce* novel words composed of sounds in their phonological inventory (known sounds) than those composed of sounds out of their phonological inventory (unknown sounds). In contrast, these same studies show that children learn to *comprehend* novel words composed of known sounds or unknown sounds at equivalent rates (Leonard et al., 1981; Schwarz & Leonard, 1982 but see also Bird & Chapman, 1998). Taken together, the findings seem to support the presence of phonological selectivity in lexical acquisition. Children appear to learn words composed of known sounds more readily than words composed of unknown sounds, although this influence is asymmetrical affecting primarily production. This suggests that sublexical representations play a role in lexical acquisition. The research documenting phonological selectivity has focused almost exclusively on children who produce fewer than 50 words. It is proposed that a developmental change occurs in lexical acquisition at this 50-word threshold resulting in a rapid increase in rate of lexical acquisition (Behrend, 1990; Bloom, 1973; Dore, 1978; Gopnik & Meltzoff, 1986; Mervis & Bertrand, 1994). For this reason, it is unknown if phonological selectivity continues to govern word learning in children beyond the 50 word stage.

There is evidence to suggest that sublexical representations influence lexical acquisition in children with productive lexicons greater than 50 words. Storkel and Rogers (in press) examined the effect of phonotactic probability (density) on lexical acquisition in typically developing 7-, 10-, and 13-year-old children. Results showed that 10- and 13-year-old children learned more high probability (density) nonwords than low probability (density) nonwords. This finding supports the claim that sublexical representations influence lexical acquisition. In contrast, 7-year-old children showed no consistent effect of probabilistic phonotactics on lexical acquisition. It was unclear if the cause of this null result related to methodological considerations or to differences in representations and processing in the youngest group of children. Also of note, the difference between high probability and low probability nonwords was small, although statistically reliable. The methods used may have attenuated the effect of phonotactic probability. Word learning was only examined in a comprehension task at one point in time. One possibility is that phonotactic probability may have a greater effect on learning the form of the novel word rather than the referent. Additionally, phonotactic probability may have a greater influence initially following limited exposure, and the effect of phonotactic probability may be attenuated as a lexical representation is formed.

The initial hypothesis that high phonotactic probability may facilitate lexical acquisition in children is based on the underlying assumptions that children are sensitive to the likelihood of occurrence of sound

sequences and that sublexical representations may influence lexical acquisition. Past research provides evidence that children do learn the distributional regularities of the ambient language. In addition, sublexical representations do appear to influence lexical acquisition at least during acquisition of the first 50 words (i.e., 12-18 months) and in older children (i.e., 10- and 13-year-old children). What remains less clear is the influence of sublexical representations, specifically phonotactic probability, on word learning in preschool children. In addition, it is unknown how phonotactic probability influences the learning of forms versus referents and how this influence changes over time. The purpose of the current study was to extend previous findings by examining the influence of phonotactic probability, a sublexical factor, on lexical acquisition in preschool children using multiple measures of learning over the course of learning. It was predicted that preschool children would learn high probability nonwords with fewer exposures than low probability nonwords, due to the hypothesized dominance of sublexical representations. Furthermore, it was predicted that phonotactic probability would affect learning of both forms and referents, but the effect would be greater for form learning. Finally, it was expected that the difference between high and low probability words would decrease with greater exposure as learning approached ceiling.

Method

Participants

Twenty-one typically developing monolingual preschool children (*M* = 4; 3, range = 3; 2 - 5; 10) were recruited by public announcements to participate in a multi-trial novel word learning task. Speech, language, hearing, and cognitive development were screened using a parent questionnaire related to medical history; the *Goldman-Fristoe Test of Articulation* (GFTA; Goldman & Fristoe, 1986); the *Peabody Picture Vocabulary Test - Revised* (PPVT-R; Dunn & Dunn, 1981); and a hearing screening (ASHA; 1985). Eligible children were required to score at the 32nd percentile or above on the GFTA (*M* =71; range 36 - 99) and the PPVT-R (*M* = 73, range 39-99).

Stimuli

The left-hand columns of Table 1 display the stimuli used in the multi-trial nonword learning task. Eight consonant-vowel-consonant (CVC) nonwords of varying phonotactic probability were chosen for the nonword learning task. Phonotactic probability can be decomposed into 2 measures: positional segment frequency and biphone frequency. *Positional segment frequency* is the likelihood of occurrence of a given sound in a given word or syllable position. *Biphone frequency* is the likelihood of occurrence of a given sound preceding or following another sound. These frequencies were computed using a 20,000 word on-line dictionary and were weighted for word frequency. High phonotactic probability was operationally defined using a median split of all legal CVC nonwords (median positional segment frequency = 0.1152; median biphone frequency = 0.0030). The four high probability nonwords had a mean positional segment frequency of 0.1639 (range 0.1157-0.2123) and a biphone frequency of 0.0055 (range 0.0036-0.0066). The four low probability nonwords had a mean positional segment frequency of 0.0849 (range 0.0595-0.1072) and a biphone frequency of 0.0010 (range 0.0004-0.0018). The four high probability nonwords were also high density (*M* = 13; range 12-18). In complement, the four low probability nonwords were also low density (*M* = 5; range 2-6). Phonemes were not repeated in the same word position across the eight nonwords to decrease the confusability among items. All nonwords were composed of early acquired sounds having a 75% level of acquisition of 3; 6 or younger (Table 5; Smit, Hand, Freilinger, Bernthal, & Bird, 1990). Mean age of phoneme acquisition using this 75% criterion was 3-years for both high and low probability nonwords. In addition, data from the GFTA was used to determine that the participating children accurately produced the phonemes used in the nonwords.

Form Characteristics		Referent Characteristics		
High PP	Low PP	Category	Item 1	Item 2
w æ t	n aʊ b	Toys	punch toy (Geisel & Geisel, 1958; p. 53)	cork gun (Geisel & Geisel, 1958; p. 45)
h ʌ p	g i m	Horns	orange trumpet downward orientation (Geisel & Geisel, 1954; p. 50)	yellow hand-held tuba (Geisel & Geisel, 1954; p. 50)
p i n	m ɔɪ d	Candy Machines	red candy + 1 shoot (invented)	blue candy + 2 shoots (invented)
k oʊ f	j e p	Pets	green gerbil with antenna (DeBrunhoff, 1981; p. 132)	purple mouse-bat (Mayer, 1992, p. 43)

Table 1. The phonetic transcription of the high and low probability nonword stimuli and their corresponding referents. Referents were invented or adapted from published children's stories.

In an attempt to equate semantic and conceptual factors, two nonsense object referents were selected from the same semantic category (Storkel & Rogers, in press). A nonsense object was defined as an object that had no corresponding single word label for adults. Objects were adapted from children's stories or were invented. The right-hand columns of Table 1 contain a description of the chosen objects. Each referent was arbitrarily paired with a high or low probability nonword. This pairing of referents and nonwords was counterbalanced across participants.

Scenes from multiple children's stories by Mercer Mayer (1993) were combined and adapted to incorporate the nonsense objects. The resulting story contained three story episodes with each episode containing the eight nonsense objects. Each story episode featured a common routine such as selecting objects to take on an outing, using objects in a competition, or hiding objects. The story pictures were 8 x 11 color drawings, mounted on a solid background, and placed in a storybook. The narrative was created so that exposure sentences were identical for each nonword within a semantic category. See the appendix for an example of the story narrative. The first story episode provided one exposure to the eight nonwords. The second and third story episodes each provided three massed exposures to the eight nonwords (refer to the appendix). A female speaker recorded two versions of the story, corresponding to the different pairings of forms and referents for counterbalancing.

Procedures

Children participated in three sessions approximately one-week apart. In the first session, entry testing consisting of the parent questionnaire, GFTA, PPVT-R, and the hearing screening was completed. In addition, children were required to accurately repeat the nonwords. The purpose of the nonword

repetition task was to ensure that all children could accurately produce the nonwords. In the second session, children participated in the multi-trial nonword learning task. Children listened to the first story episode via desktop speakers while viewing the accompanying picture book. The investigator pointed to the appropriate main character to help the child follow the dialogue. Recall that story episode 1 provided one exposure to each of the eight nonwords. Following this first exposure, nonword learning was assessed using two tasks: referent identification and form identification. In the *referent identification task*, the child heard a word presented over the speakers and was asked to select the correct picture referent from a field of three choices. The choices included the target, the foil from the same semantic category, and an unrelated foil presented in the story. In the *form identification task*, the child saw one picture and heard three choices for the name of the object. The examiner pointed to a colored square as each form choice was played over the speakers. The child then pointed to the square corresponding to his or her response. Foils were selected in the same manner as in the referent identification task. Prior to testing with the nonword targets, each task was explained and each child was required to provide accurate responses to three known words to demonstrate understanding of the task. This series of story listening and testing was repeated for story episodes 2 and 3. Recall that story episodes 2 and 3 provided three exposures to each nonword. In this way, nonword learning was tested following 1, 4 and 7 cumulative exposures in both referent and form identification tasks. Approximately 1-week after this initial learning phase, children returned for a third session to examine retention of the nonwords using both referent and form identification tasks.

Results

The current study was designed to examine the effect of phonotactic probability across different levels of exposure as revealed by two measures of learning (referent identification vs. form identification). Data collection for this project is ongoing and the results reported here are preliminary. At the outset of this project, it was predicted that children would learn high probability nonwords more rapidly than low probability nonwords. This finding would support the hypothesis that children encode phonotactic probability and that sublexical representations influence lexical acquisition. Further, it was hypothesized that stronger evidence for the effect of phonotactic probability would be found in the form identification task than in the referent identification task. This finding would seem to suggest that sublexical representations are particularly important in learning the form of a novel word. Finally, it was predicted that the difference between high and low probability nonwords would diminish with increasing exposure as learning approached ceiling or maximum performance. To provide evidence related to these predictions, the effect of phonotactic probability and exposure will be considered for each measure of learning, in turn.

Analysis

The proportion of correct responses for all four nonwords in a given phonotactic probability condition (high, low) was computed for each participant at each test point. These aggregated accuracy scores for each task (referent identification, form identification) were then submitted to a repeated measures analysis of variance (ANOVA) with the factors Exposure (1, 4, 7, Retention) and Phonotactic Probability (high, low). The purpose of this analysis was to examine the effect of phonotactic probability on nonword learning across three levels of exposure. In addition, qualitative comparisons were made between the results for referent identification and those for form identification.

Referent Identification

All 21 children were able to successfully complete the training component of the referent identification task. The results of the ANOVA showed a significant main effect of Exposure (F (3, 60) = 4.933; $p < 0.01$) and Phonotactic Probability (F (1, 20) = 4.401; $p < 0.05$). The interaction of Exposure x Phonotactic Probability failed to reach significance at this time ($F < 1$). Figure 1 displays mean proportion

of correct responses across children for high and low probability nonwords at each exposure. Proportions between 0.29 - 0.38 are not significantly different from chance (exact binomial, $p \leq 0.05$). As seen in Figure 1, high probability nonwords were learned more rapidly than low probability nonwords as predicted. Responses to high probability nonwords were significantly above chance following 4 and 7 exposures. In contrast, low probability nonwords were not responded to with greater than chance accuracy until 7 exposures. As predicted, the difference between high and low probability nonwords gradually decreased across exposures. Following 4 exposures, the greatest difference between high and low probability nonwords was observed (51% vs. 38%). Following 7 exposures, the difference in response accuracy between high and low probability nonwords was minimal (56% vs. 52%).

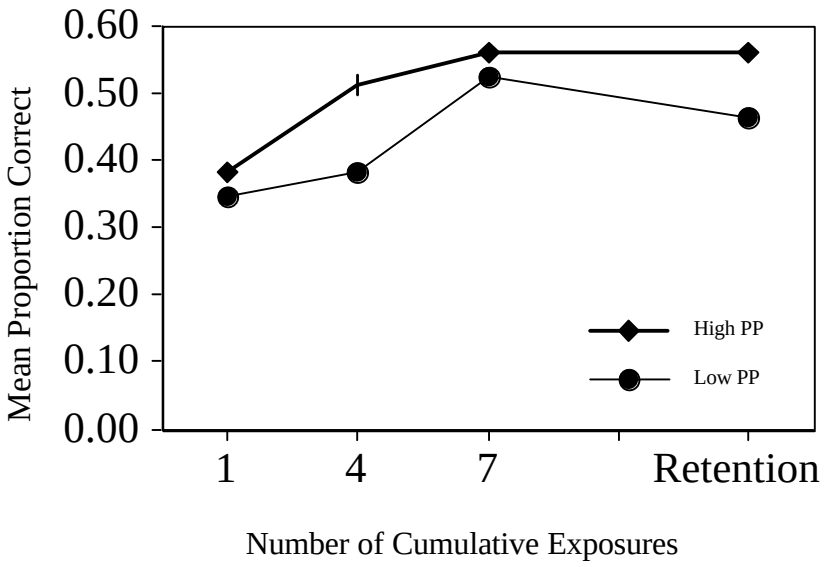


Figure 1. Mean proportion of correct responses in the referent identification task for high probability and low probability nonwords following 1, 4, and 7 exposures and retention. Note that proportions greater than 0.38 of less than 0.29 differ significantly from chance (exact binomial, $p < 0.05$)

Form Identification

Two of the 21 children were unable to complete the training component of the form identification task. The results for form identification are based on the remaining 19 children. The ANOVA analysis showed a significant main effect of Exposure (F (3, 54) = 3.362; $p < 0.05$) and Phonotactic Probability (F (1, 18) = 6.840; $p < 0.05$). The interaction of Exposure x Phonotactic Probability failed to reach significance at this time ($F < 1$). Figure 2 displays the mean proportion of correct responses for high and low probability nonwords for each cumulative exposure. Proportions greater than 0.38 or less than 0.28 differ significantly from chance (exact binomial, $p \leq 0.05$). Results from the form identification task complement the results from the referent identification task. Inspection of Figure 2 shows that high probability nonwords were acquired more readily than low probability nonwords. Responses to high probability nonwords were significantly above chance after 4 and 7 cumulative exposures. In contrast,

responses to low probability nonwords were not above chance until retention testing. In addition, the difference between high and low probability nonwords varied over time. The difference between high and low probability nonwords initially increased with 4 exposures (difference of 0.12) and 7 exposures (difference of 0.21) and then decreased at retention testing (difference of 0.11).

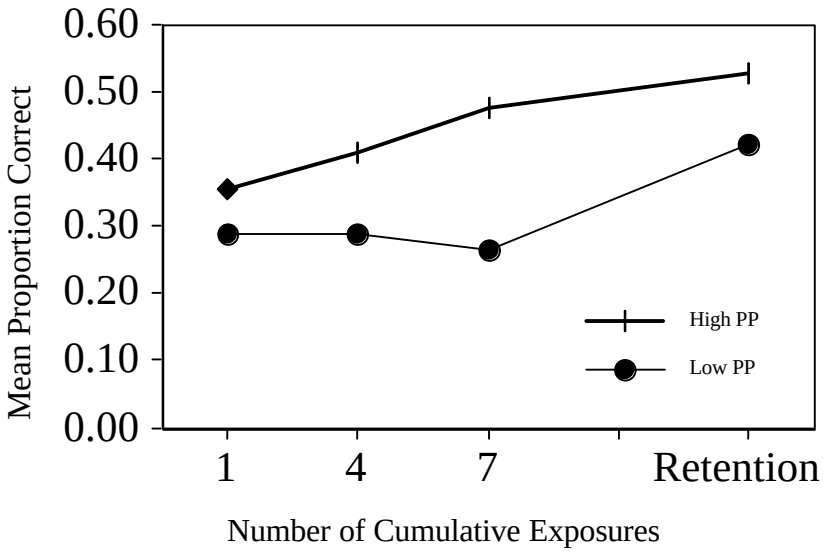


Figure 2. Mean proportion of correct responses in the form identification task for high probability and low probability nonwords following 1, 4, and 7 exposures and retention. Note that proportions greater than 0.38 of less than 0.28 differ significantly from chance (exact binomial, $p < 0.05$)

Discussion

The current study examined the effect of phonotactic probability on lexical acquisition over time as evidenced by two measures, referent and form identification. Preliminary results indicated that high probability nonwords were learned with fewer exposures than low probability nonwords. This effect was observed in both referent and form identification tasks across all test points. Although observed at all test points, the difference between high and low probability nonwords was more pronounced at specific test points and varied across referent and form identification tasks. In the referent identification task, a substantive difference between high and low probability nonwords was observed after 4 exposures only. For the form identification task, differences between high and low probability nonwords were observed after 4 and 7 exposures and at retention testing. In both tasks, the differences between high and low probability nonwords ultimately decreased across learning and retention phases. The implications of these findings for theories of lexical acquisition will be considered, in turn.

Sublexical Influences on Lexical Acquisition

The current findings provide further evidence that children learn the likelihood of occurrence of sound sequences in the ambient language and that this representation of phonotactic probability may influence language learning. The observed facilitory effect of high phonotactic probability supports the initial hypothesis that sublexical representations may play a role in lexical acquisition. It is possible that

other phonological variables, such as prosody, word length, and syllable structure, may influence lexical acquisition. For example, words with more likely stress patterns, as in the dominant trochaic pattern of English, may be easier to learn than words with less likely stress patterns. Furthermore, the influence of sublexical representations in lexical acquisition may have more global consequences for language acquisition. The learning advantage of high probability words over low probability words facilitates the creation of dense lexical neighborhoods. These dense lexical neighborhoods may highlight the relevant contrasts in the language supporting development in other areas, such as word recognition and productive phonological knowledge (Charles-Luce & Luce, 1990; 1995).

Form vs. Referent Learning

The current study provides evidence that sublexical representations influence both referent and form learning. Recall that the previous studies of phonological selectivity showed only an effect on the child's ability to learn to produce a novel word. Measures of comprehension failed to show evidence of phonological selectivity. The current results suggest that phonotactic probability influences both form and referent learning. It is possible that sublexical representations play a role in both lexical and conceptual learning. The effect of phonotactic probability on form learning suggests that sublexical representations may support the formation of lexical representations or the connection between sublexical and lexical representations. In addition, the effect of phonotactic probability on referent learning replicates the findings of Storkel and Rogers (in press) using a different method and a different age group. The implication of these referent identification results is that sublexical representations may have consequences for acquisition of the link between conceptual knowledge and word forms.

Effect of Phonotactic Probability over Time

We initially hypothesized that a nonword would eventually form a lexical representation and achieve 'word' status during learning. When a lexical representation is firmly established for a nonword, sublexical processing presumably becomes less influential than lexical processing. In addition, the sensitivity of the learning task to further changes in representations and processing diminishes as the unknown word becomes "known." Given this hypothesis, we predicted that phonotactic probability should be highly influential initially after minimal exposure. The current results support this prediction. The advantage for high probability words was observed after minimal exposure and began to dissipate with further exposure. Additional changes in representations and processing may be observed if other more sensitive psycholinguistic tasks were employed in conjunction with the word learning paradigm. That is, once the word becomes "known" other changes in representations and processing for the newly learned word might be revealed in a word recognition or production task.

Conclusions

The current study provides evidence that children learn the distributional regularities of the ambient language. Moreover, sublexical representations appear to support lexical acquisition leading to an advantage of high probability words over low probability words. Sublexical representations seem to have a robust effect on word learning influencing the acquisition of both the form and the referent.

References

Akhtar, N., & Tomasello, M. (1996). Two-year-olds learn words for absent objects and actions. *British Journal of Developmental Psychology, 14*, 79-93.

ASHA Committee on Audiologic Evaluation. (1985). Guidelines for identification audiometry. *ASHA, 27*, 49-52.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development, 62*, 875-890

Baldwin, D. A. (1993a). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*, 832-843.

Baldwin, D. A. (1993b). Infants' ability to consult the speaker for clues to word meaning. *Journal of Child Language, 20*, 395-418.

Behrend, D. A. (1990). Constraints and development: A reply to Nelson, 1988. *Cognitive Development, 61*, 681-696.

Bird, E. K. R., & Chapman, R. S. (1998). Partial representations and phonological selectivity in the comprehension of 13- to 16-month-olds. *First Language, 18*, 105-127.

Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax.* The Hague: Mouton.

Brown, R. W. (1957). Linguistic determinism and the parts of speech. *Journal of Abnormal and Social Psychology, 55*, 1-5.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17-29.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicon. *Journal of Child Language, 17*, 205-215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language, 22*, 727-735.

DeBrunhoff, L. (1981). *Babar's anniversary album*. New York, NY: Random House.

Dollaghan, C. A. (1985). Child meets word: 'fast mapping' in preschool children. *Journal of Speech and Hearing Research, 28*, 449-454.

Dore, J. (1978). Conditions for the acquisition of speech acts. In I. Markova (Ed.), *The social context of language* (pp. 87-111). New York: Wiley.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test Revised*. Circle Pines, MN: American Guidance Service.

Gathercole, S.E., Frankish, C.R., Pickering, S.J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25,* 84-95.

Geisel, T. S., & Geisel, A. S. (1954). *Horton hears a who!* New York, NY: Random House.

Geisel, T. S., & Geisel, A. S. (1958). *Cat in the hat comes back.* New York, NY: Random House.

Goldman, R., & Fristoe, M. (1986). *Goldman-Fristoe Test of Articulation.* Circle Pines, MN: American Guidance Service.

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language, 21*, 125-155.

Gopnik, A., & Meltzoff, A. N. (1986). Words, plans, things and locations: Interactions between semantic and cognitive development in the one-word stage. In S. Kuczaj & M. Barrettt (Eds.), *The development of word meaning* (pp. 199-223). New York: Springer-Verlag.

Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: an examination of fast mapping. *Child Development, 58*, 1021-1034.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1-23.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33,* 630-645.

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

Leonard, L. B., Schwartz, R. G., Morris, B., & Chapman, K. (1981). Factors influencing early lexical acquisition: Lexical orientation and phonological composition. *Child Development, 52*, 882-887.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing, 19*, 1-36.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua, 92*, 199-227.

Markman, E. M., & Hutchinson, J. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology, 16*, 1-27.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121-157.

Mayer, M. (1992). *Professor wormbog in search for the zipperump-a-zoo.* Italy: Rainbird Press.

Mayer, M. (1993). *Little critter's read-it-yourself storybook: Six funny easy-to-read stories*. New York, NY: Golden Book.

Merriman, W. E., & Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development, 54* (Serial No. 220).

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name nameless category (N3C) principle. *Child Development, 65*, 1646-1663.

Messer, S. (1967). Implicit phonology in children. *Journal of Verbal Learning and Verbal Behavior, 6*, 609-613.

Oller, D. K., Wieman, L, Doyle, W., & Ross, C. (1975). Infant babbling and speech. *Journal of Child Language, 3*, 1-11.

Paul, R., & Jennings, P. (1992). Phonological behavior in toddlers with specific expressive language delay. *Journal of Speech and Hearing Research, 35*, 99-107.

Pertz, D. L. & Bever, T. G. (1975). Sensitivity to phonological universals in children and adolescents. *Language, 51*, 149-162.

Rice, M. L. (1990). Preschoolers QUIL: quick incidental learning of novel words. In G. Conti-Ramsden & C. E. Snow (eds), *Children's Language, vol. 7* (pp. 171-195) Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Rice, M. L., & Woodsmall, L. (1988). Lessons from television: children's word learning when viewing. *Child Development, 59*, 420-429.

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language, 9*, 319-336.

Soja, N. N. (1992). Inferences about the meanings of nouns: The relationship between perception and syntax. *Cognitive Development, 7*, 29-45.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, J. A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*, 779-798.

Stoel-Gammon, C., & Dale, P. (1988, May). *Aspects of phonological development of linguistically precocious children*. Paper presented at Child Phonology Conference, University of Illinois, Champaign-Urbana.

Storkel, H. L., & Rogers, M. A. (in press). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics and Phonetics.*

Thal, D., Oroz, M., & McCaw, V. (1995). Phonological and lexical development in normal and late-talking toddlers. *Applied Psycholinguistics, 16*, 407-424.

Tomasello, M., & Barton, M. (1994). Learning words in nonostensive contexts. *Developmental Psychology, 30*, 639-650.

Vihman, M., Ferguson, C. A., & Elbert, M. (1987). Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics, 7*, 3-40.

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., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech, 40*, 47-62.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain & Language, 68*, 306-311.

Waxman, S. R. & Kosowski, T. D. (1990). Nouns mark category relations: Toddlers and preschoolers word learning biases. *Child Development, 61*, 1461-1473.

Whitehurst, G., Smith, M., Fischel, J., Arnold, D., & Lonigan, C. (1991). The continuity of babble and speech in children with expressive language delay. *Journal of Speech and Hearing Research, 34*, 1121-1129.

Appendix: Sample Narrative

Version 1

Story episode 1 "What about the pets?" asked Little Sister. "We'll take them with us" said Big Brother. "I'll get [jep]. Little Sister said, "I'll get [koʊf].

Story episode 2 "I can make our pets do more tricks than you," said Little Sister. "Uh-uh," said Big Brother. Big Brother made [jep] do tricks. He made [jep] roll-over. He made [jep] jump up and down. Next, it was Little Sister's turn. Little Sister made [koʊf] do tricks. She made [koʊf] roll-over. She made [koʊf] jump up and down.

Story episode 3 "Let's hide our pets," said Big Brother. "I'll hide [jep]. Don't make any noise [jep]." Little Sister looked and looked for [jep]. "Here he is!" Little Sister said, "I'll hide [koʊf]. Don't make any noise [koʊf]." Big Brother looked and looked for [koʊf]. "I found him."

Version 2

Story episode 1 "What about the pets?" asked Little Sister. "We'll take them with us" said Big Brother. "I'll get [koʊf]. Little Sister said, "I'll get [jep].

Story episode 2 "I can make our pets do more tricks than you," said Little Sister. "Uh-uh," said Big Brother. Big Brother made [koʊf] do tricks. He made [koʊf] roll-over. He made [koʊf] jump up and down. Next, it was Little Sister's turn. Little Sister made [jep] do tricks. She made [jep] roll-over. She made [jep] jump up and down.

Story episode 3 "Let's hide our pets," said Big Brother. "I'll hide [koʊf]. Don't make any noise [koʊf]." Little Sister looked and looked for [koʊf]. "Here he is!" Little Sister said, "I'll hide [jep]. Don't make any noise [jep]." Big Brother looked and looked for [jep]. "I found him."

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

The Effect of Linguistic Experience on Perceptual Similarity Among Nasal Consonants: A Multidimensional Scaling Analysis[1]

James D. Harnsberger

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH-NIDCD Training Grant DC00012 to Indiana University.

The Effect of Linguistic Experience on Perceptual Similarity Among Nasal Consonants: A Multidimensional Scaling Analysis

Abstract. In cross-language speech perception studies, the perceptual categories of a listener group are often assumed to be adequately represented by single, abstract labels, such as the position-dependent allophonic variant of a phoneme. However, listeners of different languages may also vary in their perceptual weighting of acoustic cues that signal a given phoneme, or its allophonic realization, in their language. These two competing units of analysis were evaluated in a cross-language perceptual similarity test employing a broad range of non-native stimuli and listener groups. In this experiment, an AXB classification test using nasal consonants from Malayalam (bilabial, interdental, alveolar, retroflex, palatal, velar) was administered to three sets of listener groups with common coronal nasal inventories: dental-retroflex (Marathi, Punjabi), alveolar-retroflex (Tamil, Oriya), and alveolar (Bengali, American English). A two-dimensional multidimensional scaling analysis of the similarity scores revealed language-specific differences that were not predictable from the test groups' nasal inventories, as represented by position-dependent allophones. The dental-retroflex and alveolar groups showed intra-group differences in their clustering of stimuli and their weighting of both perceptual dimensions, leading to language-specific perceptual spaces. Only the alveolar-retroflex group spaces were similarly organized. The results demonstrate that descriptions of the native perceptual categories of listeners must be made at the level of the individual acoustic cues that are used to match acoustic input to particular perceptual categories, rather than abstract labels.

Introduction

Cross-language speech perception research has demonstrated that specific linguistic experience can limit listener sensitivity to some non-native phonemic distinctions (Abramson and Lisker, 1970; Miyawaki, Strange, Verbrugge, Liberman, Jenkins, and Fujimura, 1975; Werker, Gilbert, Humphrey, and Tees, 1981). Non-native consonant stimuli that correspond to, or are most similar to, a single phoneme in a listener's native language have often proven to be difficult for listeners to accurately discriminate and identify, such as /l-ɹ/ for Japanese listeners (MacKain, Best, and Strange, 1981), Hindi dental-retroflex stops for English listeners (Werker et al., 1981), and Salish /k'-q'/ for English and Farsi listeners (Werker and Tees, 1984a, b). However, non-native consonant contrasts have been shown to vary in their discriminability, from chance to near native level performance (Polka, 1991; Pruitt, 1995). Moreover, listener groups varying in native language, but sharing a similar phonemic inventory, can differ in the extent to which they find a given non-native consonant contrast difficult to discriminate. For instance, Japanese and Cantonese, and Japanese and Korean, listeners differ in their ability to identify and/or discriminate natural American English (AE) or Australian /l-ɹ/, despite the fact that all three languages have only one native liquid phoneme (Henly and Sheldon, 1986; Ingram and Park, 1998). Japanese and AE listeners' perception of the four dental-retroflex stop contrasts from Hindi differed significantly, with Japanese listeners on the whole finding the contrasts easier to discriminate (Pruitt, 1995). The variability in sensitivity of listener groups to non-native, non-phonemic contrasts has led some researchers to focus on alternate descriptions of perceptual categories and models of perceptual similarity. Candidates for perceptual category descriptors, or units of analysis, include context-dependent allophones (Strange, 1995), phonetic features or cues (Bennett, 1968; Gottfried and Beddor, 1988), and distributions of individual exemplars (Pisoni and Lively, 1995; Yoneyama and Johnson, 1998). Perceptual similarity is frequently

assumed to be a transparent, linear mapping between a stimulus and a category that share a common descriptor, usually a phoneme or allophone.

This study was designed to address the unit of analysis and perceptual similarity issues in cross-language speech perception. The study addressed the following, related questions: Is the position-dependent allophone as the unit of analysis sufficient to describe listeners' perceptual categories in cross-language speech perception? Do listener groups with common phonetic inventories for a given non-native sound show any differences in their perception of such sounds? These questions were assessed by examining the perceived similarity among a set of non-native sounds by a large set of listener groups varying in their native inventories, which were described in terms of position-dependent allophones.

Six types of nasal consonants varying in place of articulation served as stimulus materials in this study. They included bilabial ([m]), interdental ([n̪]), alveolar ([n]), retroflex ([ɳ]), palatal ([ɲ]), and velar ([ŋ]) nasals. Nasal consonants were chosen as stimuli because it was hoped that they would prove to be a perceptually challenging set for some or all of the non-native listener groups. In prior work, nasal consonants varying in place of articulation have been shown to be confusable relative to other contrasts (Mohr and Wang, 1968; Hura, Lindblom, and Diehl, 1992). Non-native sounds that are highly confusable may elicit significant cross-language differences in perceived similarity that would not be predictable from abstract descriptions of a listener's language, based on units such as the phoneme or allophone.

The listener groups tested with this stimulus set were speakers of Malayalam, Marathi, Punjabi, Tamil, Oriya, Bengali, and AE. Malayalam listeners served as a control group. The six non-native listener groups were chosen to represent three different types of coronal nasal consonant inventories: dental-retroflex (Marathi, Punjabi), alveolar-retroflex (Tamil, Oriya) and alveolar (Bengali, AE). Listener groups were selected on the basis of their *coronal* nasal consonant inventory because prior work on the perception of non-native place distinctions (Polka, 1991), as well as pilot testing (see Harnsberger, 1998), had indicated that dental, alveolar, and retroflex nasals may be more confusable as a group relative to other pairings of nasals (i.e. bilabial and alveolar, palatal and velar), and thus could elicit significant cross-language differences in perceived similarity.

The listener groups were presented with triads of nasal consonants varying in place of articulation in an AXB format, and were asked to choose which nasal consonant stimulus, A or B, was more similar to X. The frequency with which two stimulus types, such as dental and alveolar, were grouped together across all test trials served as a raw similarity score. A full set of these scores, for all possible pairings of the six types of nasal consonants, were then submitted to an ALSCAL multidimensional scaling analysis (MDS) for the purpose of mapping each listener group's perceptual "space" for this stimulus set, that is, the arrangement of the six stimulus types in an n-dimensional space illustrating their degree of similarity. The perceptual spaces of each listener group were compared to see if any cross-language differences emerged that could not be predicted from their nasal consonant inventories, described in terms of allophones. Specifically, the six non-native listener groups, paired off by their common coronal nasals into three test groups (dental-retroflex, alveolar-retroflex, and alveolar), were chosen to compare two hypotheses, the *allophonic category center* and *cue-weighting* hypotheses. The allophonic category center hypothesis states that listeners' native categories are abstract and sufficiently described by the context-dependent allophonic variants of phonemes. The hypothesis represents an assumption made by several cross-language speech perception models (Best, 1995; Flege, 1995). In contrast, the cue-weighting hypothesis maintains that listener groups (varying in native language) may systematically differ in their weighting of the acoustic cues for allophonically the "same" nasal consonants. For example, Bengali and AE listeners, two listener groups which both have native alveolar nasals, may differ in the degree to which they attend to particular

acoustic cues to alveolar nasals. Cue-weighting as a source of cross-linguistic variability has been suggested by the results of a number of studies, including work by Bennett (1968), Gottfried and Beddor (1988), and Rochet (1991). If the allophonic category center hypothesis is correct, we would expect to see internal consistency in the perceptual spaces of all three groups, and systematic differences between each group. For example, we would expect to see no differences in the perceptual spaces of Marathi and Punjabi listeners (both dental-retroflex groups), Tamil and Oriya listeners (both alveolar-retroflex) or Bengali and AE listeners (both alveolar). If instead, the cue-weighting hypothesis is correct, then we would expect to observe perceptual spaces that vary on a listener group by listener group basis. Such individual listener group variability would suggest language-specific weighting of the perceptual dimensions of a space, each of which corresponds to an acoustic cue or a complex of cues,. That is, we might observe significant differences between the Marathi and Punjabi spaces, between the Tamil and Oriya spaces, and/or the Bengali and English spaces.

Methods

Stimulus Materials

The stimulus materials were restricted to two exemplars each of six types of nasal consonants varying in place of articulation ([m], [n̪], [n], [ɳ], [ɲ], [ŋ]) from a single speaker of Malayalam, a Dravidian language spoken in southern India. The stimuli appeared as medial geminates in an [i*N*i] context. Every exemplar was not matched with every other in generating trials. Instead, arbitrarily, all trials consisted of only tokens from the first set of exemplars or the second set (e.g., [m_1]-[n_1]-[$ŋ_1$], [m_2]-[n_2]-[$ŋ_2$], but never [m_1]-[n_2]-[$ŋ_1$]). Each set of exemplars was combined, resulting in 20 different kinds of triads. Both sets of 20 triads appeared in six orders (ABC, ACB, BAC, BCA, CAB, and CBA) for a total of 240 trials. The interstimulus, intertrial, and interblock intervals for the similarity test were 1 s, 5 s, and 6 s, respectively, with 20 trials per block (total test time: 33.5 minutes). The use of just two exemplar sets for a total of 240 trials, from a single talker, was necessitated by the length of two other tests administered along with the AXB classification test, given the amount of time available at testing facilities in India, and the issue of subject fatigue.[2] The stimuli from this single talker were consistently correctly judged as representative of Malayalam nasals in pilot tests, and elicited similar results as those of a second Malayalam talker in additional identification and discrimination tests (Harnsberger, 1998).

Participants

Speakers of Malayalam (N=18), Marathi (N=18), Punjabi (N=14), Tamil (N=14), Oriya (N=16), Bengali (N=17), and AE (N=18) participated in this study. All but the AE listeners were tested in India, in order to recruit subjects who varied little in terms of age, dialect, and overall linguistic experience. The Malayalam listener group was recruited to serve as a control group. The six non-native listener groups served in three test groups representing different types of native nasal consonant inventories, defined specifically in terms of coronal nasal consonants: a test group with a native dental-retroflex nasal

[2] For example, if instead of two exemplar sets (A_1-X_1-B_1 and A_2-X_2-B_2), all of the twelve stimuli had been combined in all possible orders, the number of test trials would have been 960 (8 exemplar sets * 6 order * 20 types of triads), which would have required approximately 2.5 hours of testing. Unfortunately, testing in India did not allow for the use of testing facilities for the number of sessions necessary to run such a lengthy test, in combination with several other tests that were administered (Harnsberger, 1998).

consonant contrast (Marathi, Punjabi[3]), an alveolar-retroflex test group (Tamil, Oriya[4]), and an alveolar test group (Bengali, AE)

Procedure

A forced-choice AXB classification test was administered, in which participants decided whether A or B was more similar to X. Unlike AXB discrimination, A, B, and X in a classification test are tokens from three rather than two categories.[5] Subjects were instructed to decide which nasal consonant was more similar to that of middle word, the nasal consonant of the first or the third word. Subjects were told that, while all three nasal consonants may sound quite different from one another, they were to judge which, generally speaking, was more similar to X, A or B. Prior to the test, the subjects listened to 10 randomly chosen trials to familiarize them with the test format and the kinds of stimuli being compared, with no feedback provided by the investigator.

All listener groups, except for Punjabi and AE listeners, were tested in sound-attenuated chambers affiliated with private studios in New Delhi (Malayalam, Marathi, Tamil, Oriya) and Calcutta (Bengali) in India. Native Punjabi speakers were tested in a quiet room in Amritsar, India. AE subjects were tested in a sound-attenuated chamber in the Phonetics Lab at the University of Michigan. Within a single session in India, up to eight subjects were tested, with instructions provided in English by a native speaker of Indian English. At the University of Michigan, up to four subjects were tested at one time, with instructions provided in American English.

Predictions

Two general predictions were generated by the hypotheses tested in this study, allophonic category center and cue-weighting. According to the allophonic category center hypothesis, listeners' perceptual categories possess category centers that correspond to a phoneme's rule-governed phonetic manifestation in a given context, with context defined in terms of position within a syllable or word, any proximate conditioning vowels or consonants, or position in prosodic structure. Listeners with common allophonic category centers would map stimuli to those categories in a common manner, which in turn would affect their similarity judgments of stimuli that fall into these categories, as well as their overall similarity space for a class of speech sounds. In contrast, the cue-weighting hypothesis predicts that languages with common phoneme inventories, and even common allophonic distributions of those phonemes, can nevertheless differ in their perceptual weighting of critical cues to those phonemes or allophones, which in turn would determine their overall perceptual space for a class of speech sounds. The report of the results will focus in particular on the perceived similarity for the coronal nasals, as the dental-retroflex and alveolar-retroflex groups show possible intragroup differences in their noncoronal nasal consonant inventory (see footnotes 3 and 4). However, an examination of the complete perceptual spaces of the six

[3] Marathi and Punjabi only differ in terms of the places of articulation exploited in their nasal series in one case: Marathi is sometimes described as having a velar nasal which contrasts with /m̩/, /n̪/ and /ɳ/ in final position. This velar nasal is a product of the reduction of an /ŋg/ to [ŋ] in casual speech. See Harnsberger (1998) for a summary of the literature on the phonetics of Marathi nasals. This difference between Marathi and Punjabi could have manifested itself in the position of the non-native [iŋˡi] stimuli in perceptual space (see Results and Discussion).

[4] Tamil and Oriya only differ in terms of the places of articulation exploited in their nasal series in one case: Tamil, unlike Oriya, has a contrastive palatal nasal in all but final position. See Harnsberger (1998) for a summary of the literature on the phonetics of Tamil nasals. This difference between Tamil and Oriya could have manifested itself in the position of the non-native [iɲˡi] stimuli in perceptual space (see Results and Discussion).

[5] See Goldinger (1998) for another example of AXB classification.

listener groups individually provides important descriptive information concerning the relationship between the nasal consonant inventory of a listener group and its perception of non-native nasal contrasts.

Results and Discussion

Cross-Language Differences

The results of the AXB classification test were analyzed by listener group. The raw similarity scores for all possible stimulus pairs (e.g. bilabial and alveolar, retroflex and palatal) were calculated by determining the frequency with which a stimulus type that appeared as X was judged as similar to another stimulus type, across all triads in which they appeared together. For example, the frequency with which [n] was judged as similar to [m] was calculated over all triads in which [m] was X and [n] was either A or B. These similarity scores were calculated for each individual subject, and the scores were averaged together to calculate the similarity scores for an entire listener group. In addition, the scores for similar stimulus pairs were geometrically averaged for submission to an ALSCAL MDS analysis. For example, the score for [m]-[n] when [m] was X was geometrically averaged with the score for [n]-[m] when [n] was X. Thus, a single similarity score for two stimulus types, such as [m] and [n], was used in the MDS analysis, as the analysis requires.

Each listener group's mean similarity scores were then submitted separately to both a one- and two-dimensional (2D) ALSCAL MDS. In all cases, the 2D analysis provided a substantially better fit to the similarity scores than the one-dimensional analysis. Table 1 lists the stress and proportion of variation values for each listener group for the 2D analyses. The perceptual spaces derived from the analyses for the control and test groups appear in Figures 1-4, with each axis representing similarity on some undefined, perceptually relevant dimension. Figure 1 provides the perceptual space of the control group for this experiment, the Malayalam listeners. For this task, the Malayalam listeners were expected to show a relative lack of clustering among all six stimulus types, given that all six nasals are used contrastively in the language. This expectation was upheld in all interstimulus distances, with the exception of the interdental and alveolar nasals, which were judged as significantly more similar to one another than either was to any other nasal in the perceptual space in a factorial ANOVA ($df = 14$, $F = 13.9$, $p \leq 0.0001$). Such close similarity relationships are indicated in Figure 1, and in Figures 2-4 as well, with a solid-line circle encompassing the similar stimulus types. While the close similarity between the interdental and alveolar nasals was somewhat surprising, these distances are measures of relative similarity, which do not necessarily entail that the stimuli were similar enough to be confusable for Malayalam listeners. In fact, these stimuli were successfully discriminated by the same Malayalam listeners in a prior experiment, a categorial AXB discrimination test in 96% of all test trials (Harnsberger, 1998). Overall, the Malayalam listeners produced a perceptual space in which most nasals occupied their own region of perceptual space, consistent with their phonemic status in the language.

Figures 2-4 show the perceptual spaces of the alveolar, dental-retroflex, and alveolar-retroflex listener groups. As predicted by the cue-weighting hypothesis, listener groups with similar nasal consonant inventories at the allophonic level differed from one another in several cases. For instance, the alveolar test group, composed of Bengali and AE listeners with identical nasal consonant inventories, differed from one

another in their apparent attention to the two perceptual dimensions and in the distances between different stimulus types. On one hand, the Bengali group's space showed an attention to both perceptual dimensions only in the differentiation of the bilabial stimuli from the other five stimulus types. The remaining stimuli varied on roughly a single dimension. Moreover, Bengali listeners also judged the coronal stimuli to be highly similar to one another relative to the AE listeners. All of the English coronal interstimulus distances (interdental-alveolar, alveolar-retroflex, dental-retroflex) were significantly greater ($p \leq 0.05$) than the corresponding Bengali distances in post-hoc t-tests in a repeated measures mixed model ANOVA, in which Native Language (F (1,33) = 5.44, $p \leq 0.05$), Stimulus Pair (F (14,462) = 41.89, $p \leq 0.0001$), and their interaction (F (14,462) = 3.34, $p \leq 0.0001$), were all significant. In addition, the coronal interstimulus distances in the Bengali space were significantly shorter than all other distances between coronal and noncoronal stimulus types in the Bengali space. In other words, Bengali listeners judged the non-native coronal stimuli to be highly similar to one another and relatively dissimilar to [m], [ɲ], and [ŋ]. Overall, the Bengali space can be apportioned into three sets of sounds: the coronal stimulus types, the palatal and velar stimulus types, and the bilabial stimulus type. These clusters of similar sounds are indicated in Figure 2 by circles around stimulus types that were judged to be similar to one another relative to their similarity to stimulus types outside of a given circle.

Group	Stress[6]	R^2
Malayalam	0.07129	0.96907
Marathi	0.03193	0.99476
Punjabi	0.04794	0.98985
Tamil	0.00486	0.9999
Oriya	0.01615	0.99882
AE	0.05194	0.9822
Bengali	0.03993	0.99133

Table 1. Fits of the 2D MDS analyses to the similarity scores

In contrast to the Bengali listeners, the AE listeners weighted the two dimensions differently to produce an alternate set of similarity relations and relative lack of clustering, despite the fact that AE and Bengali share a common nasal consonant inventory at the allophonic level, for these stimuli. AE listeners used both dimensions to not only differentiate the bilabial but also the retroflex stimuli from all other

[6] "Stress" in an MDS analysis is a measure of the fit of the derived distances to the raw similarity scores submitted to the analysis.

stimuli. The resulting distances between coronal stimulus types did not differ significantly from the distances of individual coronal stimulus types and near-neighbors such as [ɲ] for [n] or [m] for [ɳ], as indicated by dotted-circle lines in Figure 2. Thus, the AE space could not be nicely apportioned in the same manner as the Bengali space, a result that could not have been predicted by the allophonic category center hypothesis and one that was congruent with the cue-weighting hypothesis.

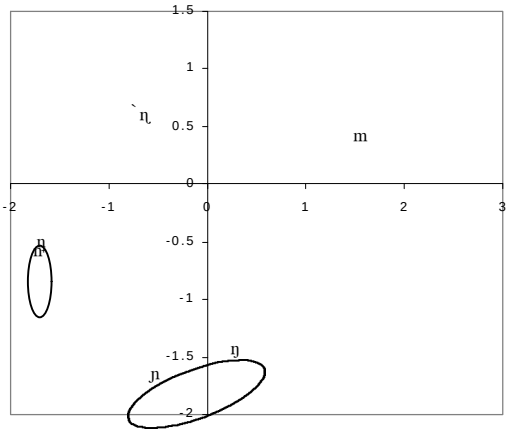


Figure 1. 2D MDS analysis of the Malayalam similarity scores. Solid-line circles encompass nasals whose distances were not significantly different from one another *and* whose distances were significantly different from the distance between them and all other nasals outside of the circle.

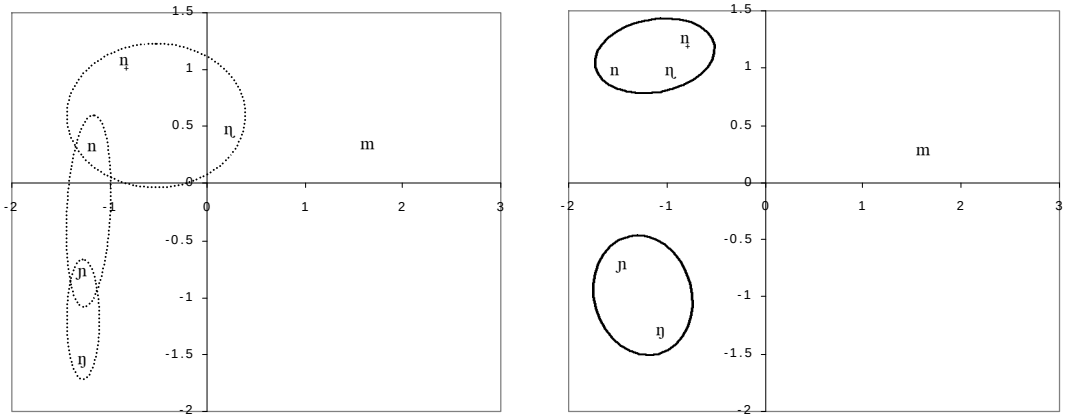


Figure 2. 2D MDS analysis of the alveolar test group' similarity scores (Left: AE, Right: Bengali). Solid-line circles encompass nasals whose distances were not significantly different from one another *and* whose distances were significantly different from the distance between them and all other nasals outside of the circle. Dotted-line circles indicate less exclusive, partially overlapping clusters of nasals.

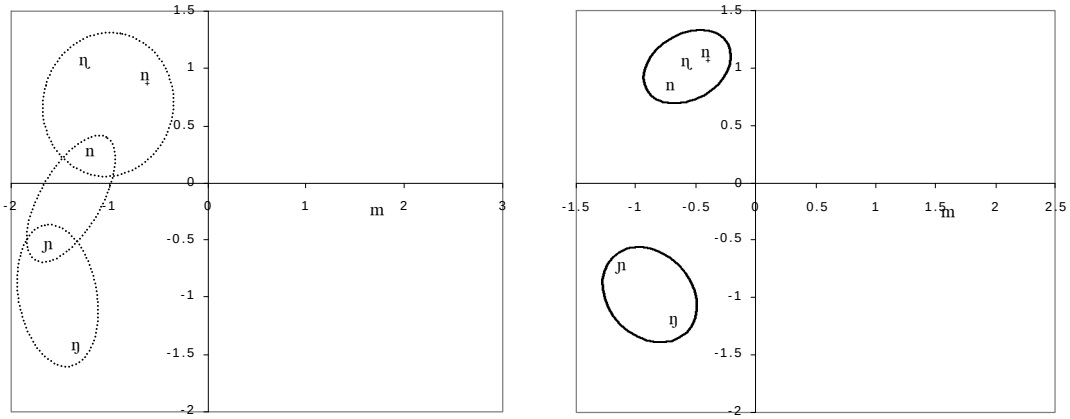


Figure 3. 2D MDS analysis of the dental-retroflex test groups' similarity scores (Left: Marathi, Right: Punjabi). Solid-line circles encompass nasals whose distances were not significantly different from one another *and* whose distances were significantly different from the distance between them and all other nasals outside of the circle. Dotted-line circles indicate less exclusive, partially overlapping clusters of nasals.

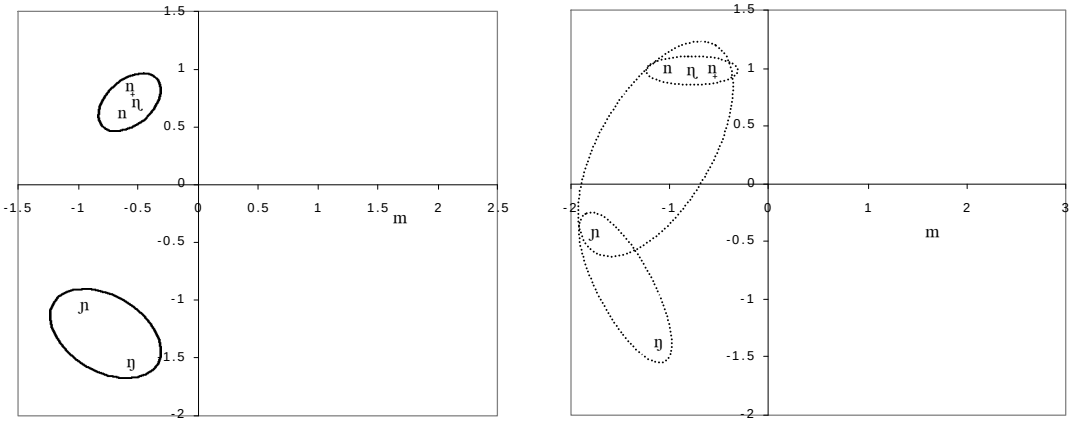


Figure 4. 2D MDS analysis of the alveolar-retroflex test groups' similarity scores (Left: Tamil, Right: Oriya). Solid-line circles encompass nasals whose distances were not significantly different from one another *and* whose distances were significantly different from the distance between them and all other nasals outside of the circle. Dotted-line circles indicate less exclusive, partially overlapping clusters of nasals.

The dental-retroflex test group, whose spaces appear in Figure 3, also showed intra-group differences not predicted by the allophonic category center hypothesis. The Marathi listeners, like the Bengali, appeared to use a single perceptual dimension in judging the [n], [ɳ], [ɲ], [ŋ], and, to a lesser extent, [n̪] stimuli. And like AE, Marathi listeners demonstrated relatively less clustering among the coronal nasals. In contrast, Punjabi listeners resembled Bengali, and not Marathi listeners, in tightly clustering the [n̪], [n], and [ɳ], reflecting perhaps less sensitivity to phonetic cues that are not contrastive in Punjabi. These general observations were confirmed in a repeated measures mixed model ANOVA, with interstimulus distances compared in post-hoc t-tests. While the Native Language of the listeners was not a significant factor (F (1,30) = 0.79, n.s.), Native Language did interact significantly with Stimulus Pair (F (14,240) = 1.77, $p \leq 0.05$). Stimulus Pair alone was also significant (F (14,240) = 91.02, $p \leq 0.0001$). In post-hoc t-tests, the two groups differed in their judged distances between alveolar and palatal nasals, and between interdental and retroflex nasals. Both of these differences were a product of the relative clustering of the coronal nasals by the two groups. With tight clustering in the Punjabi space, the coronal nasals were judged as highly similar to one another, and highly dissimilar to the palatal nasal. In post-hoc t-tests, the [n̪]-[n], [n̪]-[ɳ], and [n]-[ɳ] distances were not significantly different from one another, and all were significantly different from all coronal-noncoronal distances. In contrast, the relative spread of the coronal nasals in the Marathi space resulted in relatively a more dissimilar interdental-retroflex pair, and in a relatively similar alveolar-palatal pair. Post-hoc t-tests confirm these aspects of the Marathi space: the alveolar-palatal distance was not significantly different ($p \leq 0.05$) from the distance between the alveolar and retroflex nasal. This similarity relationship is indicated in Figure 3 by a dotted circle encompassing the alveolar and palatal nasal symbols. These results, like those of the alveolar group, cannot be accounted for by the allophonic category center hypothesis, and illustrate the need for a description of the cues that differentiate nasal consonants in these languages.

Of the three test groups, the alveolar-retroflex group showed the greatest similarities, and thus their data strongly supported the allophonic category center hypothesis. Both Tamil and Oriya listeners tightly clustered the [n̪]-[n]-[ɳ] series, and both appeared to judge the non-bilabial stimuli along a single dimension. The similarities between these two spaces were supported by comparing the interstimulus distances of both groups in a repeated measures mixed model ANOVA. In this analysis, Native Language was not significant (F (1,28) = 2.72, n.s.), nor was its interaction with Stimulus Pair (F (14,392) = 1.39, n.s). Stimulus Pair was itself significant (F (14,392) = 88.42, n.s.), as expected. In both groups' spaces, the coronal nasals clustered together, and had interstimulus distances that were significantly shorter than distances involving noncoronal stimuli ($p \leq 0.05$). In the Tamil space, the palatal and velar nasals also clustered together, as the solid circle in Figure 4 indicates. However, in the Oriya space, the distance between the palatal and velar nasal was large enough to insure that the palatal-alveolar and palatal-velar distances were not significantly different. This small difference was the only one observed between the two spaces. Interestingly, the distance between the palatal nasal and the coronal nasals was not significantly greater in the Tamil space than the Oriya space, despite the fact that Tamil possesses a palatal nasal phoneme along with their alveolar and retroflex nasals. Overall, the alveolar-retroflex spaces supported the allophonic category center hypothesis.

Language-General Patterns

Finally, in addition to cross-language differences, several general patterns emerged across most of the listener groups that were not anticipated prior to the experiment. First, the stimulus types formed three groups in terms of similarity, {[m]} {[n̪]-[n]-[ɳ]} and {[ɲ]-[ŋ]}. The bilabial stimuli were typically judged as highly dissimilar from the other stimuli in the experiment, and occupied an extreme corner of perceptual

space. In contrast, the interdental, alveolar, and retroflex stimuli clustered together for five out of seven listener groups, including the dental-retroflex and alveolar-retroflex test groups. In addition, all listener groups placed the palatal and velar stimuli in their own region of perceptual space, with these stimuli showing less clustering relative to the [n̪]-[n]-[ŋ] series. These general groupings, made by a diverse set of listener groups spanning a range of nasal consonant inventories, may reflect general psychoacoustic differences between the stimuli that are robust enough to be perceived despite the "filtering" effect of the listeners' native nasal consonant inventory. Such language-general similarity patterns could serve as benchmarks for models of perceptual similarity that attempt to predict the identification of a non-native sound based on the native phonetic inventory of a listener's language, assuming that these patterns hold in similar experiments with a greater number of stimuli from multiple talkers.

Conclusions

The results of the 2D MDS analysis of the similarity scores from the classification test revealed substantial effects of language experience in the organization of perceptual spaces that cannot be accounted for by reference to abstract category centers such as allophonic variants. Listener groups, instead of being easily classified by their nasal consonant inventory, were instead defined by their weighting of the cues or dimensions that primarily signal the contrast between the non-native stimuli. Cue-weighting itself, of course, involves abstraction, reducing the rich, highly redundant signal into a few key components. Moreover, a two-dimensional cue- or dimension- weighting model of the perceptual category can be seen as a conclusion driven by this particular form of analysis. Undoubtedly, listeners use more than just two cues, however defined, in perceiving speech under a variety of adverse environments for communication, including visual as well as auditory information (Summerfield, 1987).

What appears to be cue weighting could also be the product of a language-specific distribution of individual exemplars of a nasal consonant, each of which corresponds to an instance or episode stored in long-term memory (Pisoni and Lively, 1995). That is, language-specific differences might fall directly out of the sum total of experiences a given listener group (or a given listener) has with a given stimulus type, such as nasal consonants. An episodic-based account might take the following form: Bengali and AE listeners show, for instance, a difference in their relative clustering of the [n̪]-[n]-[ŋ] series. Bengali listeners cluster these more, and their clustering may reflect their greater experience hearing dental and retroflex nasals. Bengali listeners encounter dental and retroflex nasals more frequently than AE listeners as allophonic variants of their /n/, due to place assimilation before dental and retroflex oral stops, which are much more frequent in Bengali than AE. Thus, Bengali listeners may have more experience than AE listeners in classifying dental and retroflex nasals as alveolar, accounting for their tight clustering of non-native [n̪], [n], and [ŋ].

The results of this experiment provide no support for a model of language-specific cue weighting, over an episodic-based model of speech perception. The differences between listener groups that share the same nasal consonant inventory only highlight the need to go beyond the allophone in our search for the proper unit of analysis in cross-language speech perception. This is not a new idea, but it is one that has not been taken to heart in much cross-language research. One issue this study raises is the need for more detailed descriptions of the phonetics of languages than exist for most of the world's languages. To test a cue-weighting model, we will need to describe how the phonetic cues for a given phoneme or allophone are weighted in a linguistic community. Testing an exemplar-based model would require even more descriptive work on a language, entailing the development of large phonetic databases from which distributions of speech sounds along given dimensions could be extracted. However, without such descriptive work, we

may never be able to develop models of cross-language speech perception that are capable of making quantitative predictions of the magnitude of cross-language differences.

In summary, this study evaluated two competing units of analysis in cross-language speech perception, the position-dependent allophone and the phonetic cue, in a perceptual similarity test employing a broad range of nasal consonant stimuli and listener groups varying in their nasal consonant inventory. A two-dimensional multidimensional scaling analysis of the similarity scores revealed language-specific differences that were not predictable from the test groups' nasal inventories, as represented by position-dependent allophones. The two sets of listener groups, representing dental-retroflex and alveolar nasal inventories, showed intra-group differences in their clustering of stimuli and their weighting of perceptual dimensions, leading to language-specific perceptual spaces. Only the spaces of the two alveolar-retroflex were organized in a similar manner. This study, even with its inherent limitations in terms of stimuli (only two exemplars of each speech sound from one talker), indicated that abstract phonological or phonetic models of the perceptual category do not capture important and significant variability in the cross-language data, and that new techniques and new theoretical assumptions are warranted in our study of cross-language speech perception.

References

Abramson, A. S., and Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In Hála, B., Romportl, M. and Janota, P. (ed.), *Proceedings of the Sixth International Congress of Phonetic Sciences*, pp. 569-573. Prague: Academia.

Bennett, D. C. (1968). Spectral form and duration as cues in the recognition of English and German vowels. *Language and Speech, 11*, 65-85.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in Cross-language speech perception*, pp. 171-204. Baltimore: York Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in Cross-language speech perception*, pp. 233-277. Baltimore: York Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Gottfried, T. L. and Beddor, P. S. (1988). Perception of temporal and spectral information in French vowels. *Language and Speech, 31*, 57-75.

Harnsberger, J. D. (1998). *The perception of non-native nasal contrasts: a cross-linguistic perspective*. Unpublished doctoral dissertation, University of Michigan.

Henly, E. and Sheldon, A. (1986). Duration and context effects on the perception of English /r/ and /l/: A comparison Cantonese and Japanese speakers. *Language Learning, 36*, 505-521.

Hura, S.L., Lindblom, B., and Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech, 35*, 59-72.

Ingram, J. and Park, S-G. (1998). Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America, 103,* 1161-74.

MacKain, K. S., Best, C. T., and Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2,* 369-90.

Miyawaki, K., Strange W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception and Psychophysics, 18,* 331-65.

Mohr, B. and Wang, W. S.-Y. (1968). Perceptual distance and the specification of phonological features. *Phonetica, 18,* 31-45.

Pisoni, D. and Lively, S. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research,* pp. 433-458. Baltimore: York Press.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *Journal of the Acoustical Society of America, 89,* 2961-777.

Pruitt, J. S. (1995). *The perception of Hindi dental and retroflex stop consonants by native speakers of Japanese and AE*. Unpublished doctoral dissertation, University of South Florida, Tampa.

Rochet, B. L. (1991). Perception of the high vowel continuum: A cross-language study. *Proceedings of the International Congress of Phonetic Sciences, 4,* 94-97.

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research,* pp. 3-45. Baltimore: York Press.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R. (ed.), *Hearing by Eye: The Psychology of Lip-Reading,* pp. 3-51. Hillsdale, NJ: LEA.

Werker, J. F., Gilbert, J. H. V., Humphrey, K. and Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development, 52,* 349-53.

Werker, J. F. and Tees, R. C. (1984a). Cross-language speech perception: Evidence for perceptual reorganization in the first year of life. *Infant Behavior and Development, 7,* 49-63.

Werker, J. F. and Tees, R. C. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75,* 1866-78.

Yoneyama, K. and Johnson, K. (1998). An instance-based model of Japanese speech recognition by native and non-native listeners. *Journal of the Acoustical Society of America, 103,* 3090.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

A Voice is a Face is a Voice: Cross-Modal Source Identification of Indexical Information in Speech[1]

Lorin Lachs

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by a grant from the NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington. Special thanks go to David Pisoni and Luis Hernàndez for excellent suggestions and brainstorming sessions. Thanks also to Tyler Emley, Patrick Kelley and Jaime Brumfield for all their help in collecting and processing this data.

A Voice is a Face is a Voice: Cross-Modal Source Identification of Indexical Information in Speech

Abstract. Recent evidence from experiments using sinewave speech shows that the linguistic content of a message, as well as the indexical characteristics of the talker can be perceived from the isolated kinematic form of speech utterances. Similarly, isolated visual kinematic information in the form of point-light displays has been shown to behave in much the same way that full visual displays of a talker articulating do (e.g., by enhancing intelligibility in noise). If the isolated kinematic visual form of speech is informative in speech perception, and the isolated kinematic acoustic form of speech can carry indexical information, then visual information should also be able to carry information regarding the indexical properties of the talker. If this is true, then perceivers should be able to use the information about an utterance obtained through one sensory modality (e.g., vision) and use it to identify the same utterance in the other sensory modality (e.g., audition). The present study examined the ability of participants to perceive and use either auditory or visual information about articulation across sensory modalities in identifying source characteristics of a talker's voice.

Optical information about articulation has been shown to have substantial effects on speech perception (Massaro & Cohen, 1995). In the absence of auditory stimulation, visual information is sufficient for accurate speech perception (Bernstein, Demorest, & Tucker, in press). In conjunction with auditory information, visual information can enhance speech intelligibility in noise by +15 dB (Sumby & Pollack, 1954). Alternatively, incongruent information in the auditory and visual aspects of multimodal stimuli can interact to form illusory percepts (the “McGurk” effect McGurk & MacDonald, 1976).

Because visual information about articulation can have such profound effects on speech perception, some theorists have proposed that the perceptually useful information in speech signals must be transmittable via acoustic as well as optic media. Indeed, some researchers have gone so far as to propose that the information is amodal; that is, the information for speech is not constrained to any particular sensory modality. In fact, the McGurk effect has been replicated by using auditory and tactile information about speech (Fowler & Dekle, 1991) demonstrating that some degree of useful information about speech can be obtained through sensory modalities other than audition.

In another experiment designed to demonstrate that speech information can be carried in multiple sensory modalities, Green and Kuhl (1989) showed that the perceived VOT boundary for a synthetic /bi/-/pi/ continuum was shifted toward the VOT boundary for a /di/-/ti/ continuum when the stimuli were paired with the visual specification of a talker uttering the syllable /gi/. That is, the VOT boundary shifted in an appropriate manner for the illusory percept invoked by the McGurk illusion. In an earlier study, Green and Miller (1985) showed that the speaking rate information in optical displays influenced the identification of voiced or voiceless segments on an acoustic continuum that remained constant. These studies demonstrate that the dynamic aspects of visual information play a role in the perception of speech, and that the same auditory information can be perceived differently depending on the kinds of visual information available during perception.

The exact form of such information remains the subject of some debate, but a growing body of research points to the possibility that the acoustic or optical forms of speech signals carry kinematic or dynamic information about the articulation of the vocal tract, and that such information drives the

perception of linguistically relevant utterances (Fowler, 1986; Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985; Rosenblum & Saldaña, 1996; Summerfield, 1987). The use of dynamic information has been demonstrated across multiple contexts. For example, Green and Gerdeman (1995) showed that cross-modal discrepancies in the *vowel* portion of McGurk stimuli influenced the degree to which the *consonant* portion was susceptible to the McGurk effect. The findings suggest that the perceptual system must be sensitive to non-segmental, coarticulatory information when it attempts to make sense of multimodal inputs.

Another method used to study the problem of audiovisual integration in speech perception is the point-light technique (Johansson, 1973). By placing small reflective patches at key positions on a talker's face and darkening everything else in the display, one can isolate the kinematic aspects of visual displays of talkers articulating speech (Rosenblum, Johnson, & Saldaña, 1996; Rosenblum & Saldaña, 1996). Such "kinematic primitives" have been shown to behave much like unmodified, full visual displays of speech (Rosenblum & Saldaña, 1996). For example, the McGurk illusion can be induced by dubbing visual point-light displays onto phonetically discrepant auditory syllables (Rosenblum & Saldaña, 1996). In addition, an extension of Sumby and Pollack's (1954) findings has demonstrated that providing point-light information about articulation in conjunction with auditory speech embedded in noise can result in increased intelligibility (Rosenblum et al., 1996).

All of the studies reviewed above, and indeed, most of the previous investigations of the effects of audiovisual information on speech perception have focussed on what are commonly referred to as the *linguistic* aspects of the signal: phoneme or syllable identification and spoken word recognition. However, a growing body of literature has shown that speech signals also carry information about the *indexical* properties of the talker, and that this information is perceived, stored in memory and used during speech perception and spoken word recognition (see Goldinger, 1998; Pisoni, 1997, for a review).

Numerous recent studies have shown that the indexical properties of a talker's voice are stored in long-term memory (Bradlow, Nygaard, & Pisoni, 1999; Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989). For example, using a continuous recognition task, Palmeri, Goldinger, and Pisoni (1993) showed that repeating a word in the same voice that produced it during study facilitated later recognition of that word. Furthermore, the size of this effect did not change depending on the number of talkers uttering test items. This suggested that the encoding of voice attributes in memory is automatic and not controlled by strategic processes.

The link between memorial encoding of fine-grained details of spoken words and perceptual processes has also been established (Nygaard, Sommers, & Pisoni, 1994; Nygaard, Sommers, & Pisoni, 1995). In one experiment, Nygaard and Pisoni (1998) trained participants to identify a set of novel talkers from their voices alone. Once the participants had learned the voices using a set of training stimuli, Nygaard and Pisoni found that the knowledge of talker characteristics obtained also generalized to new stimuli. Furthermore, the perceptual learning of the trained voices transferred to a novel task: words spoken by familiar voices were recognized more accurately in noise than words spoken by unfamiliar voices.

But what kind of information about a talker is contained in speech, and how does that information contribute to speech perception? In an examination of the acoustic correlates of talker intelligibility, Bradlow, Torretta and Pisoni (1996) showed that while global characteristics such as fundamental frequency and speaking rate had little effect on intelligibility, acoustic-phonetic properties of voice, such as vowel space reduction and "articulatory precision", were strong indicators of overall intelligibility. These findings suggest that indexical properties of a talker may be completely intermixed with the

phonetic realization of the utterance, with no real dissociation between the two sources of information in the speech signal.

More direct evidence for this hypothesis comes from recent studies using sinewave replicas of speech. Sinewave speech is made by generating sinusoidal tones that trace the center frequencies of the three lowest formants produced during a natural utterance. The resulting tone complex sounds completely unnatural, but can be perceived by listeners as speech (Remez, Rubin, Pisoni, & Carrell, 1981). Indeed, not only is the linguistic content of the utterance perceptible, but specific aspects of a talker's unique identity are also preserved in sinewave replicas. Remez, Fellowes, and Rubin (1997) showed that participants could identify specific familiar talkers from sinewave replicas of their utterances. This finding is remarkable because sinewave speech patterns preserve none of the traditional stimulus aspects that cue vocal identity, such as fundamental frequency, or the average long-term spectrum. Furthermore, Fellowes, Remez, and Rubin (1997) also showed that while the gender of talkers could be perceived from sinewave replicas, the correct perception of gender was not a precondition for the identification of a specific talker. In the words of Fellowes et al., (1997), "Personal information is available in an aspect of the signal that does not arise through anatomical variation alone" (p. 848).

These findings raise some important questions about the domain of speech perception. Sinewave speech strips an utterance of all information except the time-varying properties of the resonances generated by articulatory motion. In this way, sinewave speech is much like a point-light display; it isolates the kinematic information in an acoustic display that relates to the underlying articulation. If the common metric for auditory and visual information about speech is kinematic, and kinematic auditory information has been shown to provide information for the identity of a talker, then visual kinematics should also provide the same kind of information. However, evidence for the visual perception and use of nonlinguistic information from articulatory activity has been reported only recently. In one study, Gagné, Masterson, Munhall, Bilida, and Querengesser (1994) showed that the visual intelligibility of tokens spoken in the "clear speech" speaking style was higher than the visual intelligibility of tokens spoken in a conversational style. In another study, Rosenblum, Yakel, Baseer, and Panchal (1999) showed that visual point light displays of a talker articulating could be matched accurately to unaltered visual displays of the same talker articulating.

If the important information for talker identity and source recognition in speech is contained in kinematic and dynamic aspects of articulating vocal tracts, then perceivers should be able to match the identity of a speaking face across sensory modalities, because the criterial information is not necessarily carried *solely* by the acoustic specification of speech. In theory, kinematic and dynamic information is modality-neutral, and can be conveyed by both optical and acoustic media. The question, then, is whether perceivers can actually use this information to make judgments of source identity across sensory modalities.

In order to answer this question, an experiment was designed that used a 2-alternative forced choice paradigm. Perceivers were required to match a talker presented in one modality with the same talker presented in the other modality. The two sensory modalities used for this experiment were visual and auditory. Several factors might play a role in the perceiver's ability to perform such a task. First, hit rates would be expected to be very high if the two alternatives were of different genders. Accordingly, all the talkers identified by a particular participant were of the same gender. Different groups of participants identified male or female talkers, in order to test for differences in performance based on the gender of the talkers. A second factor that might also play a role in this matching task was the order (or direction) in which the judgment was made. For example, it might be the case that seeing a face and then judging which of two voices matched it is easier than the converse situation: hearing a voice and judging which of two faces matched it. Both conditions were tested as repeated measures to find any difference in

performance dependent on the order in which the modalities were presented. In addition, it is possible that fine-grained details of the stimulus will be lost if the stimulus is unintelligible in one or the other modality. In order to test this possibility, stimulus items were balanced for their intelligibility. Because the stimulus items used in this study were all highly intelligible in audio-only identification tests, stimulus items were split into low and high groups according to their visual intelligibility based on visual-only identification tests (Lachs & Hernández, 1998). The visual-only (VO) intelligibility of the stimulus items was manipulated as a repeated-measures variable. Finally, confidence ratings were collected in order to determine whether participants were aware of the particular trials on which they performed well. If confidence ratings were higher on correct trials, and lower on incorrect trials, then participants must have a good estimate of their ability to perform this unusual task.

Method

Experimental Design

A two-alternative forced choice procedure was used in a 2 x 2 x 2 repeated measures factorial design. The two levels for the between-subjects Gender factor were "male" and "female". The two levels for the within-subjects Direction factor were "A-V" (where participants identified the correct visual stimulus after viewing the test auditory stimulus) and "V-A" (where participants identified the correct auditory stimulus after viewing the test visual stimulus). The levels of this factor were blocked and counterbalanced across participants for the order in which they were presented. The two levels of the within-subjects visual intelligibility factor were "low" and "high". Stimuli in the "low" group were words whose average VO intelligibility was in the bottom 1% of the distribution of VO intelligibilities for the HAVMD (Lachs & Hernández, 1998). Stimuli in the "high" group were taken from the top 5% of the same distribution. The percentages are different because of the extreme leftward skew of the VO intelligibility distribution (i.e., relatively few words had better than average accuracy scores). The levels of this factor were randomly distributed in each block for the Direction factor.

Participants

Participants were 40 undergraduate students enrolled in an introductory psychology course who received partial credit for participation. All of the participants were native speakers of English. None of the participants reported any hearing or speech disorders at the time of testing. In addition, all participants reported having normal or corrected-to-normal vision.

Stimulus Materials

Two Apple Macintosh computers, each equipped with a 17" Sony Trinitron Monitor (0.26 dot pitch) and a TARGA 2000 video processing board were used to present the visual stimuli to subjects. The video processing boards were each capable of handling clips digitized at 30fps with a size of 640 x 480 and 24-bit resolution. Auditory stimuli were presented over Beyer Dynamic DT100 headphones.

The stimuli were a subset of tokens selected from the Hoosier Audiovisual Multitalker Database ((HAVMD, Lachs & Hernández, 1998; Sheffert, Lachs, & Hernández, 1996). The video portion of HAVMD tokens was digitized at 30 fps with 24-bit resolution with 640 x 480 pixel size. The audio portion of HAVMD tokens was digitized at 22 kHz with 16-bit resolution. Movie clips from eight talkers were used in this study (F1, F2, F3, F4, M1, M2, M3, and M4).

Procedures

Participants were told that they would be seeing two blocks of stimulus trials. In one block (the "V-A" level of the Direction factor), they would see a video clip of a talker uttering an isolated, English word, but they would not be able to hear it. Shortly after seeing this video display, they would be presented with two audio clips. One of the clips would be the same talker they had seen in the video, while the other clip would be a different talker. Participants were instructed to choose which audio clip matched the talker they had seen. The same instructions were provided for the trials in which the participant heard the audio clip first, and had to make their decision based on two video displays (the "A-V" level of the Direction factor).

On each trial, the test stimulus was either the video or audio portion of one movie token based on an isolated word spoken by one talker. The correct target choice was the other (cross-modal) portion of the same movie. The distractor choice was the same word spoken by one of the other three talkers, presented in the same modality as the target alternative. The order in which the target and distractor choices were presented was randomly determined on each trial. All responses were made with the mouse and recorded in a log file for further analysis.

After a response was made using dialog boxes and mouse inputs, participants were asked to record a confidence judgment for their response. The ratings were made on a scale of 1 to 7, with 1 marked as "not confident at all" and 7 marked as "very confident". At the completion of the session, participants were asked to briefly describe any strategies they used in order to accomplish the task.

Results

Determining Chance Performance

The data from each participant was analyzed first to determine if his/her performance in either the "A-V" or "V-A" conditions differed significantly from chance. A binomial distribution was used to calculate the probability that the number of successful trials in each block was due to chance performance (p(correct) = 0.5). Chance performance was rejected if $p < 0.05$.

Because the Direction of Judgment factor was manipulated within-subjects, the data from 40 participants was available for the V-A and A-V conditions. The accuracy of 11 participants in the "V-A" condition did not significantly differ from chance. However, of those 11 participants, the performance of 8 participants was only marginally attributable to chance ($p < 0.1$). Similarly, the accuracy of 15 participants in the "A-V" conditions were not significantly different from chance, although 7 participants were within marginal range ($p < 0.1$). These data indicate that a little less than 75% of the participants were able to accomplish this task, with slightly fewer being able to perform as well in the "A-V" conditions.

Of the 40 participants, 33 were able to respond above chance in at least one of the "V-A" and "A-V" conditions. This indicates that most participants were able to do the matching task in one form or the other. The 7 participants who did not perform above chance in either condition were eliminated from the final data analysis.

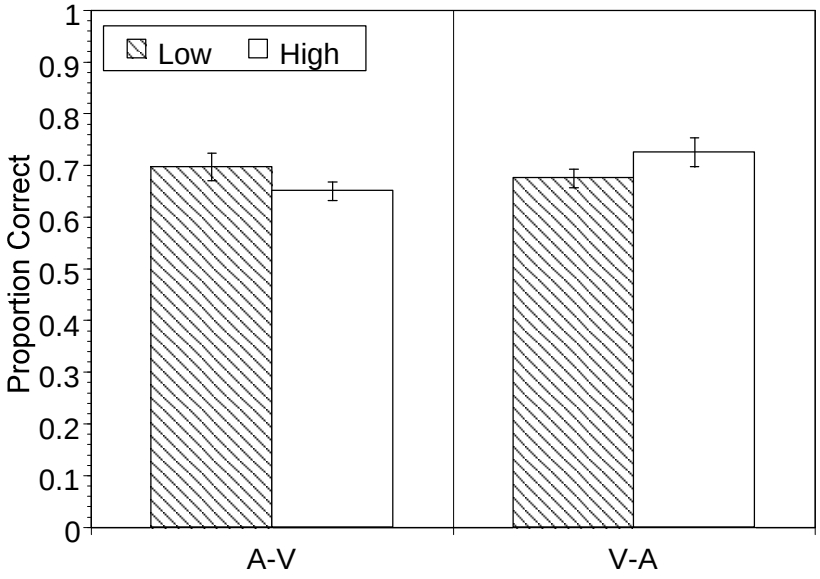


Figure 1. The interaction between visual intelligibility and the direction of the judgment. The striped bar represents performance when intelligibility is low, and the open bar represents performance when intelligibility is high.

Accuracy Analyses

The data obtained from 33 participants were then submitted to a 2 (Direction) x 2 (Visual Intelligibility) x 2 (Gender) x 2 (Order) repeated-measures ANOVA using percent correct as the dependent variable. Differences were evaluated with an α of 0.05. No main effects for any of these variables were observed, but the ANOVA did reveal a significant interaction between Direction and Intelligibility, F (1, 29) = 5.743, MSE = 0.075, p = 0.023, η^2 = 0.165. This interaction is illustrated in Figure 1. The set of bars on the left represent performance in the A-V trials, while the set of bars on the right represent performance on the V-A trials. Within each set of bars, the shaded bar represents performance when stimuli were of low visual intelligibility and the open bar represents performance when stimuli were of high visual intelligibility. Examination of this figure shows that participants performed better on V-A trials when the words were of high visual intelligibility. In contrast, there was a small difference in the degree to which low visual intelligibility words influenced participants' ability in the A-V direction. Post-hoc pairwise comparisons revealed that the source of this interaction could be localized in a significant difference between performance on low and high intelligibility words in the V-A direction.

The analysis also revealed a significant interaction between Direction and Order, F (1, 29) = 9.438, MSE = 0.152, p = 0.005, η^2 = 0.246. Differences in performance in either direction depended on whether they received that direction in the first or second block of the experiment. Figure 2 illustrates this interaction. The left panel shows the average performance of participants in the A-V block, and the right panel shows average performance in the V-A block. Within each panel, the shaded bar represents performance when the participant received the A-V block first, and the open bar represents performance when the participant received the V-A block first. As shown in the figure, participants performed better in the A-V block when they had received the V-A block first. In contrast, participants performed better in

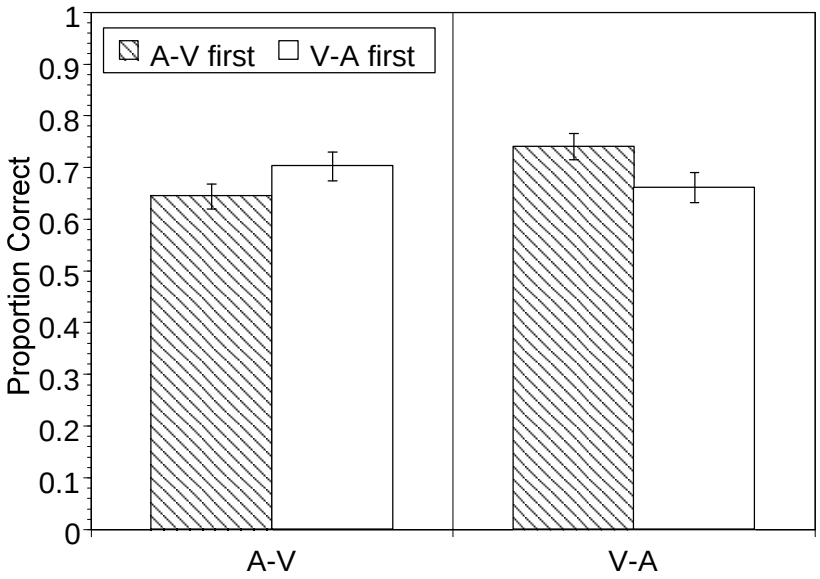


Figure 2. The interaction between the direction of judgment and the counterbalancing order. The striped bar shows performance when the A-V condition was presented first, and the open bar shows performance when the V-A condition was presented first.

the V-A block when they had received the A-V block first. That is, participants tended to be more accurate in the second Direction block they received. Post-hoc pairwise comparisons showed that significant improvement in the second block was only seen when participants received the V-A block first.

Because the counterbalancing order affected performance, the Direction factor was replaced with another within-subjects factor ("Block)", which consisted of two levels: "first" and "second". This analysis was carried out in order to examine the effects of block order regardless of the specific conditions implemented in the block. The reapportioned data were submitted to a 2 (Block) x 2 (Intelligibility) x 2 (Gender) x 2 (Order) repeated-measures ANOVA. As reported earlier, there was a main effect of Block, F (1, 29) = 9.438, MSE = 0.152, p = 0.005, η^2 = 0.246. Performance in the second block (M = 0.721, $S.E.$ = 0.019) was always better than performance in the first block (M = 0.652, $S.E.$ = 0.019), regardless of the direction in which the judgments were made.

In addition, the analysis revealed a significant interaction between Intelligibility and Block, F (1, 29) = 7.636, MSE = 0.1, p = 0.01, η^2 = 0.208. The degree to which the visual intelligibility of the stimuli affected performance depended on whether the trial was in the first or second block. Figure 3 shows the average performance of participants on low and high visual intelligibility stimuli in the first and second blocks. There was clearly an improvement in performance across the blocks in the ability of participants to perform this task when given low visual intelligibility stimuli. However, no improvement is observed for the high visual intelligibility stimuli. Post-hoc pairwise comparisons confirmed that this interaction was due to a difference in performance across blocks for low intelligibility words, but not for high ones. Interestingly, this improvement occurred *regardless of the direction in which the matching judgments were made.*

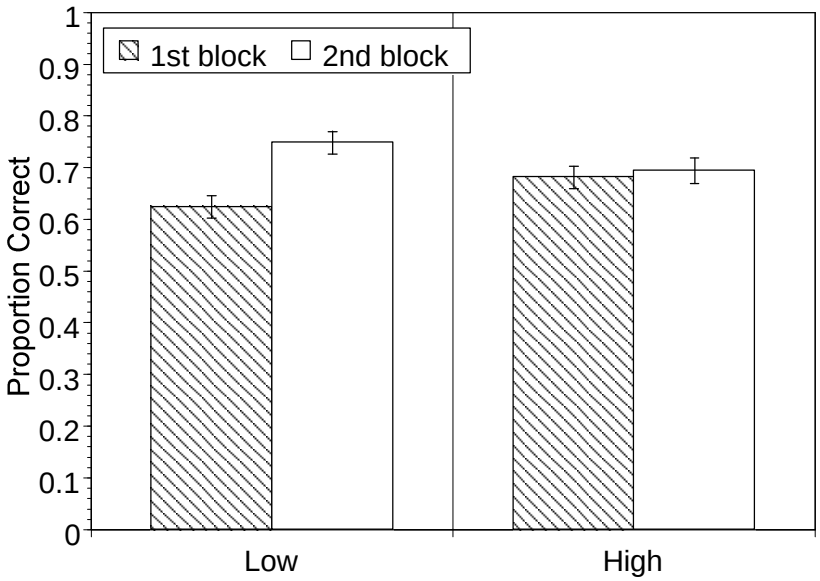


Figure 3. The interaction between Block order and visual Intelligibility. Striped bars represent performance on the first block of trials, while the open bar represents performance on the second block of trials.

Finally, this analysis revealed a significant three-way interaction between Block, Intelligibility and Order, F (1, 29) = 5.743, MSE = 0.075, p = 0.023, η^2 = 0.165. This is shown in Figure 4. The improvement in performance from the first to second block for low intelligibility words was affected by which direction was presented first. Simple effects analyses were conducted to examine this interaction at each level of Counterbalance Order. These analyses showed a significant interaction between Block and Intelligibility when the V-A block was presented first (right panel of Figure 4; F (1, 12) = 6.957, MSE = 0.152, p = 0.022, η^2 = 0.367). Post-hoc pairwise comparisons between the conditions at this level of Order revealed that the interaction was due to the difference in performance between blocks for stimuli with low visual intelligibility. The difference between blocks for high intelligibility words was not significant. No interaction between Block and Intelligibility when the A-V block was presented first (left panel of Figure 4; $F < 1$, n.s.).

Summary of accuracy scores. In summary, the accuracy analyses revealed several interesting facts about this unusual task. First, the majority of participants were able to make judgments about talker identity across sensory modalities. The ability to perform this task did not seem to be affected by the direction in which judgments were made. In addition, participants got better at making cross-modal judgments in the second block with which they were presented. This improvement was mainly for stimulus items with low visual intelligibility. Furthermore, participants only showed this improvement in the second block for low intelligibility items when the V-A block was presented first and the A-V block was presented second.

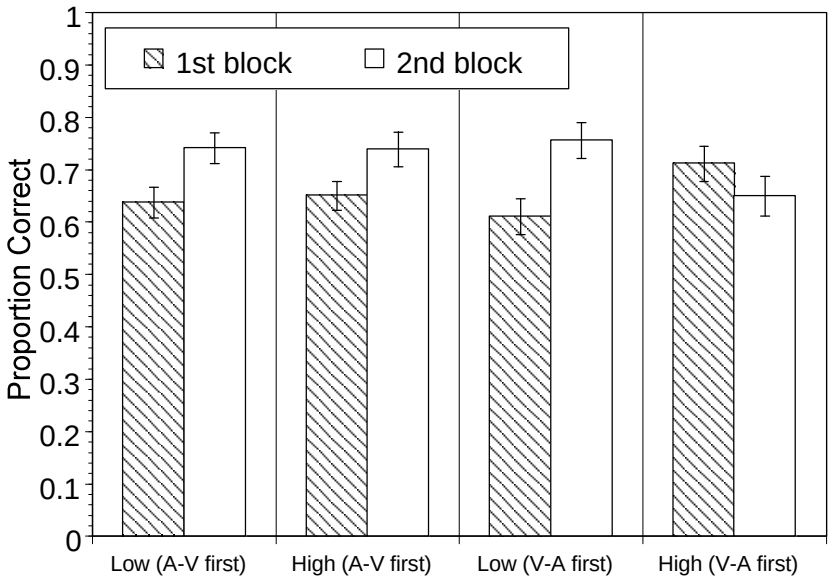


Figure 4. The interaction between Block order and visual Intelligibility. Striped bars represent performance on the first block of trials, while the open bar represents performance on the second block of trials.

Confidence Ratings

Because each participant used a different range of the 7-point confidence scale, confidence ratings were transformed relative to the scale across which each participant made his/her judgments. Before normalization, there was more variability across participants in the average confidence rating given (M = 4.47, SD = 0.74), than there was in the average variability of confidence ratings (M = 1.2, SD = 0.37). Thus, the normalized confidence ratings adjusted for differences in the "midpoint" of the scale used by different participants, without drastically changing the "range" over which confidence judgments were made.

In analyzing the confidence ratings, both correct and incorrect responses were examined. The "Score" factor had two levels: "correct" and "incorrect" and was used to determine whether there were differences in confidence based on the accuracy of responses. The normalized confidence ratings were then submitted to a 2 (Direction) x 2 (Visual Intelligibility) x 2 (Score) x 2 (Gender) x 2 (Order) repeated-measures ANOVA. A significant main effect of Score was observed, F (1, 29) = 56.068, MSE = 2.612, $p < 0.01$, $\eta^2 = 0.659$. As shown by the effect size statistic (η^2), this main effect accounted for approximately 66% of the variance in confidence ratings. Confidence ratings on correct trials (M = 0.133, SE = 0.018) were significantly higher than confidence ratings on incorrect trials (M = -0.069, SE = 0.01). Apparently, participants were aware of their ability to perform this task. However, the relatively low average values for both correct and incorrect trials indicate that participants remained equivocal on most trials. The large effect size with such low values also indicates that the extremes of each participant's confidence scale did not vary much around their average confidence rating.

The 5-way ANOVA on normalized confidence ratings also revealed a significant interaction between Direction and Visual Intelligibility, F (1, 29) = 5.985, MSE = 0.038, p = 0.021, $\eta^2 = 0.171$. This interaction is illustrated in Figure 5. Post-hoc pairwise comparisons revealed that the confidence ratings in the V-A/Hi intelligibility condition were significantly different ($p < 0.05$) from those in either A-V condition. As shown in the figure, confidence ratings in the V-A/Hi intelligibility condition were on than

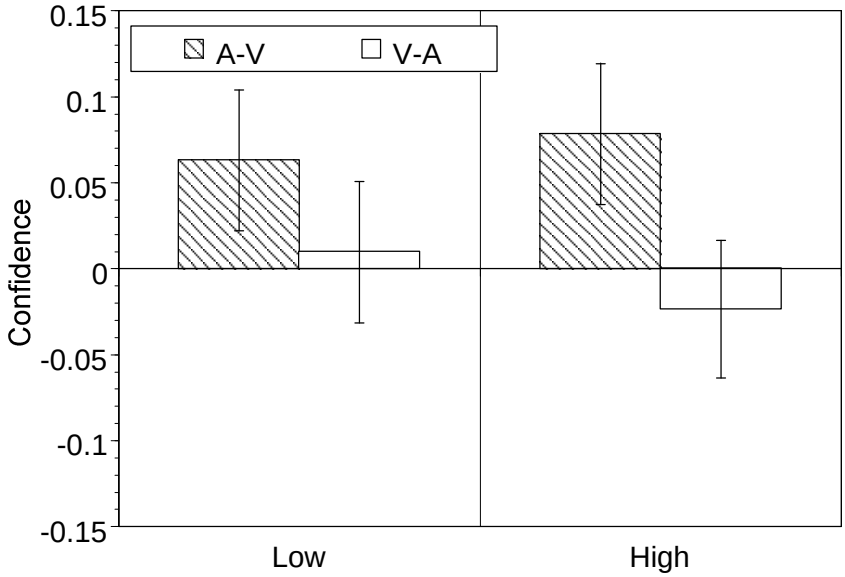


Figure 5. The interaction between Direction and Visual Intelligibility. Striped bars represent confidence on the A-V trials, while the open bar represents confidence on the V-A trials.

the average confidence ratings given in any of the other conditions, regardless of whether or not the trial was correct. Paradoxically, participants were less confident of their performance on V-A trials when they were making their decision based on highly useful visual information for spoken word recognition. This unusual result is discussed more fully in the discussion section below.

Interestingly, there was also a significant 3-way interaction between Direction, Intelligibility and Score, F (1, 29) = 5.316, MSE = 0.264, p = 0.028, η^2 = 0.155. Figure 6 shows this interaction. The left panel in the figure shows scores in the A-V conditions; the right panel shows scores in the V-A conditions. Within each panel, the left set of bars shows performance when the visual intelligibility of the stimulus was low; the right set of bars shows performance when the intelligibility was high. Shaded bars show performance on trials that were correct and open bars show performance on trials that were incorrect. It is very clear from the figure that confidence ratings for incorrect trials were generally much lower for V-A trials than they were for A-V trials. This pattern indicates that participants were more sensitive to their performance in the V-A block than in the A-V block. For trials they got incorrect, they were less confident. This pattern was more pronounced for low visual intelligibility stimuli than it was for high visual intelligibility stimuli.

In order to understand the nature of this interaction more fully, the difference in confidence ratings between correct trials and incorrect trials for each participant were computed. These difference scores were then submitted to a 4-way repeated measures ANOVA using Direction, Intelligibility, Gender and Order as factors. Figure 7 illustrates the significant interaction between Direction and Intelligibility (F (1, 29) = 5.316, MSE = 0.529, p = 0.028, η^2 = 0.155) revealed by this analysis. As shown here, the difference in average confidence ratings for correct and incorrect trials was greatest in the V-A condition when stimulus items were of low intelligibility. Post-hoc pairwise comparisons showed that the difference in this condition was significantly larger than the differences in either the V-A high intelligibility condition or the A-V low intelligibility condition.

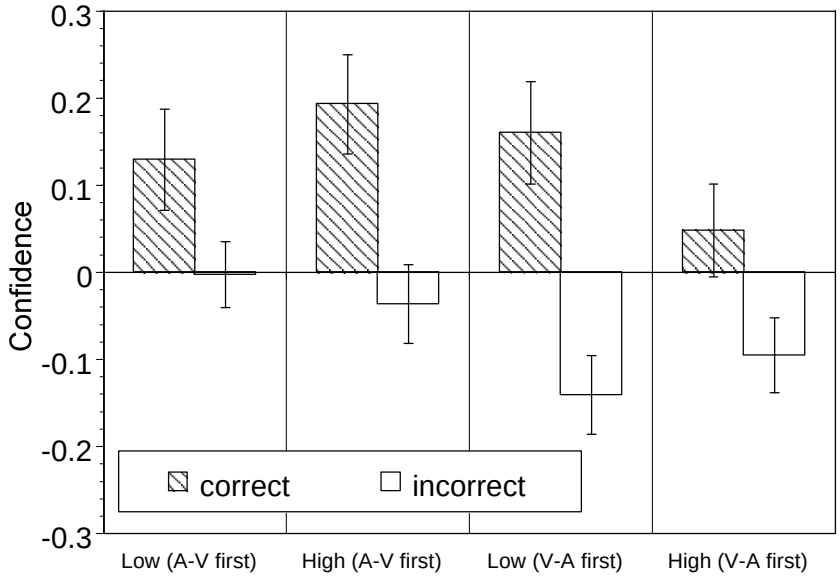


Figure 6. The interaction between Direction, Visual Intelligibility and Score. Striped bars represent confidence on correct trials, while the open bar represents confidence on incorrect trials.

Discussion

The present cross-modal source identification experiment examined the ability of participants to perceive and use auditory or visual information about articulation across sensory modalities. Participants were presented with the unimodal form of a spoken word token and required to choose which of a pair of cross-modal tokens specified the same talker. Roughly three-fourths of the participants tested were able to perform this task with better than chance performance. In addition, participants were aware of the tokens that they correctly matched, as shown by the higher confidence ratings for correct trials.

Perceptual learning was also observed in this experiment. Participants tended to perform better on the second block of trials than on the first. This learning effect was due to an increase in accuracy when the stimuli were low VO intelligibility words. Furthermore, improvement in the second block was only seen when the first block a participant experienced was in the V-A direction. Experiencing the V-A block first may have focused participants' attention on fine-grained details of optical movement in the articulator area. Stimuli with low visual intelligibility may have increased attention to these fine-grained details than those with high intelligibility. Thus, when the matching task switched to the A-V direction, participants had already learned to attend to those aspects of low visual intelligibility stimuli that would aid in the completion of the task (possibly including their acoustic correlates).

However, no such knowledge was acquired by participants who received the A-V direction first, because the low visual intelligibility of the words was only apparent *after* the test stimulus was presented. That is, the *acoustic* form of a low *visual* intelligibility word does not necessarily focus attention on fine-grained details of the stimulus, since it is not necessarily hard to perceive. Because of this, participants who experienced the A-V block first may have approached the V-A block unprepared to deal with the ambiguous or noisy information in low intelligibility words for which they were asked to make a choice.

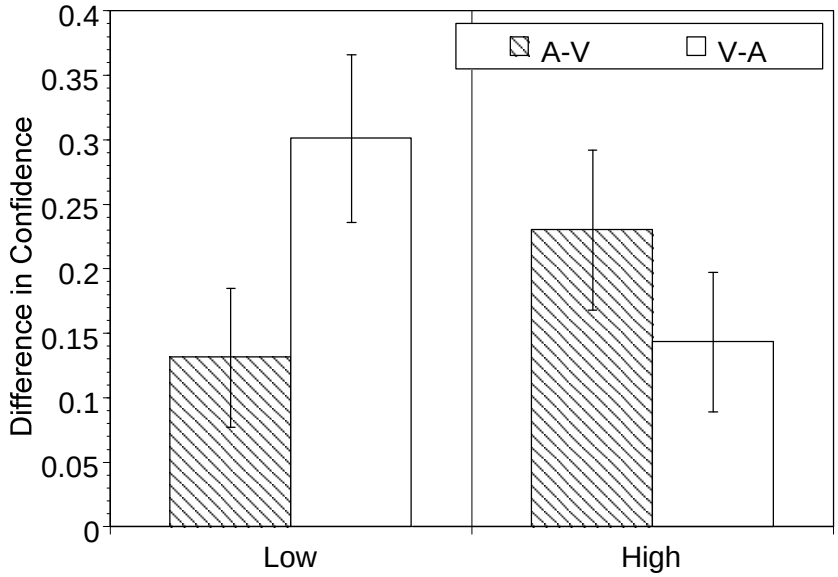


Figure 7. The interaction between Direction and visual Intelligibility for the difference in confidence ratings for correct and incorrect trials. Striped bars represent confidence on the A-V trials, while the open bars represent confidence on the V-A trials.

The patterns observed for the confidence ratings seem quite paradoxical. Although confidence was higher for trials scored correct than for those scored incorrect, overall confidence ratings in the experimental conditions did not seem to match performance levels. Despite the fact that performance was not higher in the A-V conditions than in any of the others, confidence ratings for the A-V direction of judgment were generally higher than average, with the highest ratings observed when the visual intelligibility of the tokens was high. Surprisingly, although performance was generally *best* in the V-A conditions where tokens were of high visual intelligibility, confidence ratings to these trials were on average lower than those given in any of the other conditions. Confidence ratings in the V-A/low condition were about average. This may be because participants expected that their performance in the V-A condition would be very good. If participants also expected that low intelligibility words would not be as useful as high intelligibility words, then their confidence might be placed at a "midway point" on the scale.

Interestingly, the *difference* in confidence ratings between correct vs. incorrect trials was greatest for the V-A low condition. This pattern indicates that, for this condition more than any other, there was almost a perceived dichotomy in the ability to perform the task. Remember that accuracy scores in the V-A/High intelligibility condition were significantly higher than scores in the V-A/Low intelligibility condition. Since low and high intelligibility words were interspersed throughout the blocks, it is tempting to speculate that the relatively "easy" nature of trials in the V-A/High condition set up an expectation for continued high performance on all V-A trials, exaggerating differences in confidence ratings for correct and incorrect trials.

Of course, the current experiment was not designed to test for these possibilities, so the explanations offered above concerning learning and bias in this task remain speculative at best. What is incontrovertible about the present investigation's findings, however, is that participants *can* perform this unusual matching task. Furthermore, they can do so with surprising accuracy. This provides further

support for the notion that the information about a talker's voice is inherently amodal. It can be carried by either visual or acoustic energy and it can be perceived by either the visual or auditory systems.

One surprising result was the absence of asymmetries in the direction over which the matching judgments were made. Because there *are* differences in the amount of information that acoustic and optic displays can carry about the motion of the vocal articulators, this result deserves further study. According to the Source-Filter model of speech production (Fant, 1960), the resonances in the frequency spectrum characteristic of the speech signal are directly related to the configuration of the vocal tract. The motion of the articulators causes changes in the formants' central frequencies over time. Accordingly, the acoustic form of speech can carry information about the positions and movements of the vocal articulators from the lips to the larynx. But, the same is not true for the optic form of speech. Visual displays can carry information about the configuration of the lips, tongue tip, and jaw, but it is very unlikely that they can carry information about the configuration of the velum, or show that there is a closure in the glottal area (Dodd & Campbell, 1987; Summerfield, 1987).

It is possible, however, that the effects of these asymmetries interact with the amount of phonetic information in the signal itself. After all, if the visual signal is already known to be sufficiently informative for spoken word recognition, the limitations in its ability to carry information about the non-visible articulators may be irrelevant. The data above showed that changing the visual intelligibility of the test stimulus affected performance when making judgments in the V-A direction. It is possible that adjusting the intelligibility of the audio portion of these tokens will affect performance in the A-V direction, as well.

Additional support for the proposal that the useful information in the current task is kinematic or dynamic in nature is provided by responses from a post-hoc analysis of the interview conducted at the end of each participant's session. All participants were asked to briefly describe any strategies they were using to accomplish the task. Most participants listed at least two strategies, frequently citing a preference for one or the other. A tally was made of the different types of strategies mentioned in the exit interviews and is shown in Table 1.

The “Enunciation/Emphasis” category refers to strategies that made reference to the use of information relating to the emphasis with which words were spoken or the enunciation of segments within a word. Typical responses included “I looked at their mouths to see if the way their mouths were moving was like the way I heard it”. One interesting aspect of these responses was that they almost naturally referred to the optical display as “sounding like” something. Clearly, most of the participants involved in the study used this information in forming their responses. Such information refers to the dynamics of articulatory movement: emphasis is related to the forcefulness of articulator movement, enunciation is related to the forcefulness and precision with which articulatory movements are made. Without explicit instructions, participants apparently focused their attention on what is hypothesized to be the relevant information in multimodal displays. “Duration”, the second most popular strategy used, is also a dynamic cue relating to the kinematics of articulatory motion. Responses in this category suggest that duration differences could often distinguish one candidate from another.

The remaining strategies mentioned fall in another class entirely from the two most frequent ones. While the most frequent ones rely on the use of stimulus-driven properties, the rest seem to be more “top-down” or “heuristically-driven”. “Learning” strategies specifically described how the task got easier once associations had been learned between specific voices and faces. “Expectation” strategies noted that some of the talkers seemed to “look like” they possessed certain vocal characteristics such as a high fundamental frequency. Participants who used the “self-repetition” strategy stated that they repeated the words back to themselves (silently or out-loud) as closely to the stimulus as possible.

Strategy	Number of responses
1. Enunciation/Emphasis	29
2. Duration	14
3. Learning	10
4. Expectation based on facial characteristics	7
5. Self-repetition	6
6. Process of elimination	4
7. Emotion/expression	2

Table 1. Frequency of occurrence for the various strategies mentioned in the exit interview. (See text for an explanation of the various strategy names.)

A few participants explained that they were able to figure out the correct pairings of voices and faces by the "process of elimination", i.e., by determining which pairs co-occurred most often. While this strategy is feasible, it seems like an extraordinarily complicated task, since it requires the explicit recognition of each voice and face in the experiment, and the maintenance of a frequency table of cooccurrence. Assuming that using this strategy for one block was sufficient to carry performance during the other block, a participant would have to keep track of the frequency of occurrence of 16 different response pairs (there were 4 talkers used for all participants). Unfortunately, the design of this experiment does not permit the elimination of this response strategy as a possibility, but the low number of participants who reported using it and the relatively heavy computational load it would impose cast doubts upon its usefulness. Finally, two participants reported the use of emotional expression as a cue for cross-modal identity. This is interesting because it points to the use of another, unanticipated non-linguistic aspect of the speech signal.

The ability to perceive the identity of the source of acoustic events has been demonstrated in other domains besides speech. Repp (1987) presented the sound of hand clapping for identification by participants. Some of the claps were generated by the participants themselves and the others were generated by people with whom the participants were acquainted. Perceivers performed above chance on this task, although their absolute identification accuracy was rather low (11%). Furthermore, perceivers were able to recognize the sound of their own clapping with almost 50% accuracy.

More evidence of the ability of perceivers to identify source characteristics from acoustic stimuli comes from Li, Logan, and Pastore (1991), who asked participants to identify the gender of a person whom they heard walking. Remarkably, judgments of gender were well above chance. Furthermore, anthropomorphic differences (such as weight and height) between walkers were found to be highly correlated with judgments of gender, indicating that the acoustics generated by different body-types contained information that allowed the accurate perception of these attributes.

Both the Repp (1987) and Li et al. (1991) studies indicate that detailed information about sound-producing events can be perceived and used to identify the idiosyncratic minutiae associated with the person producing them. This is also true in the domain of speech perception (Fellowes et al., 1997; Remez et al., 1997). The subtle variations exhibited by different talkers during the process of speech production can be used to identify the specific talker uttering a speech event. The current study has demonstrated that this information can be used in judgments of source variation *across sensory modalities*.

As pointed out by Vatikiotis-Bateson and his colleagues: "...the motor planning and execution associated with producing speech necessarily generates visual information as a by-product." (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzepoulos, 1997, p. 221). As a consequence, it is entirely possible that any information of relevance in the acoustic signal is also carried, in some form, by the visual signal. Because kinematic and dynamic sources of information about speech articulators are inherently amodal, they are good candidates for the form in which this information can be transmitted. Studies using point-light displays (Rosenblum & Saldaña, 1996) and sinewave replicas of speech (Remez, Rubin, Berns, Pardo, & Lang, 1994) demonstrate that speech information can be perceived from highly impoverished stimulus patterns that isolate kinematic and dynamic information. The present investigation has extended previous findings that indexical information about the source of spoken events is carried in the time-varying information about the motion of the articulators. Such information is modality-neutral, and as such can be perceived and used to make accurate judgments of identity across sensory modalities.

References

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (in press). Speech perception without hearing. *Perception & Psychophysics*.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recongition memory. *Perception & Psychophysics, 61*(2), 206 - 219.

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 - 273.

Dodd, B. E., & Campbell, R. (1987). *Hearing by eye: the psychology of lip-reading*. London; Hillsdale, N.J.: Lawrence Erlbaum Associates.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton and Co.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59*(6), 839 - 849.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3 - 28.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 17*(3), 816 - 828.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. In M. S.-K. Ignatius G. Mattingly (Ed.), *Modularity and the motor theory of speech perception.* (pp. 33-59): Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Gagné, J.-P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology, 27*, 135 - 158.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 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, & Cognition, 17*(1), 152-162.

Green, K. P., & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance, 21*(6), 1409 -1426.

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics, 45*(1), 34 - 42.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 38*(3), 269 - 276.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14*, 201 - 211.

Lachs, L., & Hernández, L. R. (1998). Update: The Hoosier Audiovisual Multitalker Database, *Research on Spoken Language Processing Progress Report 22* (pp. 377 -388). Bloomington, IN: Speech Research Laboratory, Indiana University.

Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America, 90*(6), 3036 - 3049.

Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition, 21*, 1 - 36.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*(4), 676-684.

Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science, 4*(4), 104-109.

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

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

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., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variabiliy on perception and representation of spoken words in memory. *Perception & Psychophysics, 57*, 989 - 1001.

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*(2), 309 - 328.

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.

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*(5), 651 - 666.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101*(1), 129-156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947 - 950.

Repp, B. H. (1987). The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America, 81*(4), 1100 - 1109.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research, 39*, 1159 - 1170.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 22*(2), 318 - 331.

Rosenblum, L. D., Yakel, D. A., Baseer, N., & Panchal, A. (1999). Visual speech information for face recognition. Manuscript submitted for publication.

Sheffert, S. M., Lachs, L., & Hernández, L. R. (1996). The Hoosier Audiovisual Multitalker Database. *Research on Spoken Language Processing No. 21* (pp. 578 - 583). Bloomington, IN: Speech Research Laboratory, Indiana University.

Sumby, W. H., & Pollack, I. (1954). Visual contribution of 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, NJ: Lawrence Erlbaum Associates.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzepoulos, D. (1997). The dynamics of audiovisual behavior in speech. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines* (pp. 221 - 232). Berlin: Springer-Verlag.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

New Directions in Pediatric Cochlear Implantation[1]

Karen I. Kirk,[2] Laurie S. Eisenberg[3] and Richard T. Miyamoto[2]

Indiana University
Department of Otolaryngology
DeVault Otologic Laboratory
Indianapolis, Indiana 46202

[1] This research was supported by NIH-NIDCD Grants RO1 DC00064, RO1 DC00423 and KO8 DC00126 and Psi Iota Xi.
[2] Also, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405
[3] House Ear Institute, Los Angeles, CA.

New Directions in Pediatric Cochlear Implantation

Abstract. Cochlear implantation has been an approved surgical intervention for children (2-18 years) with profound deafness for nearly 10 years. The last decade has brought technological advances in cochlear implant designs with concomitant improvements in patient performance and subsequent broadening of cochlear implant candidacy. Although average performance levels clearly establish the efficacy of pediatric cochlear implantation, individual communication abilities vary widely. This, controversy exists regarding the appropriate expansion of evolving technology into new patient populations. In this chapter, we review current implant technology, patient selection criteria and performance results for pediatric cochlear implant recipients and consider the challenges inherent in the broadening of cochlear implant candidacy.

Introduction

In 1990, the Food and Drug Administration first gave approval for cochlear implantation in children aged 2 to 18 years. Initially, children who received a cochlear implant (CI) had total, profound deafness and were usually older than five years of age. Early speech perception results demonstrated that pediatric CI users displayed substantial closed-set abilities (e.g., wherein children identify a word by selecting from a limited set of response alternatives), but only minimal open-set spoken (i.e., wherein no response alternatives are provided) word recognition abilities (Osberger et al., 1991; Staller, Beiter, Brimacombe, Mecklenberg, & Arndt, 1991). The data collected from 80 children as part of the FDA pediatric clinical trials of the Nucleus 22-channel cochlear implant system using a feature-extraction speech processing strategy, and reported by Staller (1998) illustrated the early speech recognition performance levels. The mean age at onset of deafness for this group of children was just under 3 years, and their mean age at implantation was almost 10 years. On average, children achieved monosyllabic word recognition scores of only 10% words correct and the majority of children tested could not correctly identify any individual words through listening alone. Since then, as cochlear implantation has been extended clinically to younger children, and with continued improvements in electrode design and signal processing (Wilson, 1993; Wilson et al., 1991; Wilson et al., 1993), children with CIs have achieved much higher levels of open-set word recognition (Cowan et al., 1997a; Cowan et al., 1997b; Kirk, Pisoni & Osberger, 1995; Miyamoto, Kirk, Svirsky & Sehgal, in press; Osberger, Fisher, Zimmerman-Phillips, Geier & Barker, 1998; Sehgal, Kirk, Svirsky & Miyamoto, 1998; Zimmerman-Phillips, Osberger & Robbins, 1997). For example, Eisenberg and her colleagues (Eisenberg, Martinez, Sennaroglu & Osberger, in press) reported mean PB-K scores of approximately 50% words correct for oral pediatric CI users. Open-set word recognition is an important diagnostic yardstick for determining cochlear implant success because it indicates that these children have established neural representations of words in their long-term lexical memory, a process that is fundamental to the development of spoken language (Woodward & Markman, 1997). Although these average results are very encouraging and clearly establish the efficacy of CIs, individual patients vary greatly in outcome (Osberger et al., 1991; Staller et al., 1991; Staller, 1998; Zimmerman-Phillips, Osberger & Robbins, 1997; Fryauf-Bertschy, Tyler, Kelsay & Gantz, 1992; Fryauf-Bertschy et al., 1997; Miyamoto et al., 1989; Tyler et al., 1997a; Tyler et al., 1997b; Tyler et al., 1997c). Some children can communicate extremely well using the auditory/oral modality and acquire age-appropriate language skills, whereas other children may display only minimal spoken word recognition skills and/or demonstrate severe language delays (Bollard, Chute, Popp & Parisier, 1999; Pisoni, Svirsky, Kirk & Miyamoto, submitted; Robbins, Ballard & Green, 1999; Robbins, Svirsky & Kirk, 1997; Svirsky,

Sloan, Caldwell & Miyamoto, 1998; Tyler et al., in press). Accounting for this enormous variability in the effectiveness of CIs on a wide range of outcome measures presents the most serious challenge facing cochlear implant clinicians and researchers today. Gaining an understanding of the nature of the individual differences and sources of variability in cochlear implant outcomes is crucial for predicting individual benefits prior to implantation, and for selecting appropriate intervention strategies following implantation.

Despite the variability in individual outcomes, cochlear implantation is no longer questioned as a therapeutic option for selected profoundly deaf children. However, in part because the outcomes are not guaranteed, controversy exists regarding the appropriate expansion of evolving technology into new patient populations. The current trend toward earlier implantation and the implantation of children with more residual hearing mandates careful documentation of performance limits with cochlear implants as well as with nonsurgical alternatives (e.g., hearing aids). Only through rigorous longitudinal studies will these issues be clarified. In this chapter we review current implant technology, patient selection criteria and performance results for pediatric cochlear implant recipients and consider the challenges inherent in the broadening of cochlear implant candidacy.

Background

Pediatric Cochlear Implant Selection Criteria

Current selection criteria for pediatric cochlear implantation are as follows:

- 18 months of age or greater
- Profound bilateral sensorineural hearing loss (SNHL)
- No appreciable benefit from hearing aids
- No medical contraindications
- High motivation and appropriate expectations
- Enrolled in program that emphasizes development of auditory skills

Pediatric cochlear implant recipients can be divided into three categories with significantly different expected performance outcomes:

1) Congenitally or early-deafened young children. Congenital or early-acquired deafness is the most frequently encountered type of profound sensorineural hearing loss. Cochlear implantation is performed with the anticipation that sufficient acoustic input can be provided to allow the child to perceive a speech signal linguistically. The acquisition of communication skills is typically a difficult process for these children.

2) Congenitally or early-deafened adolescents. When cochlear implantation is considered in adolescence or young adulthood for a patient who has had little or no experience with sound because of congenital or early onset deafness, caution must be exercised because this group has not demonstrated high levels of success with electrical stimulation of the auditory system.

3) Postlingually deafened children. Children who become deaf at or after age 5 are generally classified as postlingually deafened. Even though these children have developed many aspects of spoken language before the onset of their deafness, they demonstrate rapid deterioration in the intelligibility of their speech once they lose access to auditory input and feedback.

Early implantation can potentially ameliorate this rapid deterioration in speech production and perception abilities. However, a postlingual onset of deafness is an infrequent occurrence in the pediatric population.

Cochlear Implant Systems

The cochlear implant devices available for implantation, as well as the speech processing strategies used, continue to undergo technological improvements. Currently, three types of multi-channel, multi-electrode cochlear implant devices are commercially available for children in the United States. These devices have several characteristics in common. All have an electrode array that is surgically implanted into the cochlea and an external unit, consisting of a microphone that picks up sound energy and converts it to an electric signal and a signal processor that modifies the signal, depending on the processing scheme in use. The processed signal is amplified and compressed to match the narrow electrical dynamic range of the ear. (The typical response range of the ear to electrical stimulation is on the order of only 10 to 20 dB, and even less in the high frequencies.) Transmission of the electrical signal across the skin from the external unit to the implanted electrode array is most commonly accomplished by the use of electromagnetic induction or radio frequency transmission. The neural elements stimulated appear to be the spiral ganglion cells or axons. These devices use place coding to transfer frequency information in addition to providing temporal and amplitude information.

The Nucleus 22-channel (and recent 24) cochlear implant is currently the most commonly used multi-channel system. The Nucleus implantable electrode array consists of platinum-iridium band electrodes placed in a silastic carrier (Clark, 1987). Several generations of speech processors have been employed with the Nucleus multi-channel cochlear implant. The initial Nucleus speech processors used a feature-extraction scheme in which selected key features of speech were presented through the implanted electrode array. An early speech processing strategy, the F0-F1-F2 strategy, primarily conveyed vowel information, including the first and second formant frequencies and their amplitudes, as well as voice pitch. A later coding scheme, the MULTIPEAK strategy, presented these acoustic features along with additional information from three high-frequency spectral bands to aid in consonant perception. The current Nucleus speech processing strategy is the Spectral Peak (SPEAK) strategy. This strategy uses a vocoder in which a filterbank consisting of 20 filters covering the center frequencies from 200-10,000 Hz is employed. Each filter is allocated to an active electrode in the array. The filter outputs are scanned and the electrodes that are stimulated represent filters that contain speech components with the highest amplitude. Depending on the acoustic input, the number of spectral maxima detected, and thus the number of electrodes stimulated, on each scan cycle can vary from one to 10, with an average of six per cycle. The rate at which the electrodes are stimulated varies adaptively between 180-300 pulses per second.

The Clarion multichannel cochlear implant has an eight-channel electrode array that utilizes a radial bipolar configuration through electrode pairs positioned adjacent to the osseous spiral lamina in a 90-degree orientation (Schindler et al., 1986). The Clarion multi-channel cochlear implant offers two types of speech processing strategies: Simultaneous Analog Stimulation (SAS) and Continuous Interleaved Sampling (CIS). Both strategies represent the waveform or envelope of the speech signal (Wilson et al., 1991). The Clarion SAS strategy first compresses the analog signal into the restricted range for electrically-evoked hearing, and then filters the signal into a maximum of eight channels for presentation to the corresponding electrodes. Speech information is conveyed via the relative amplitudes and the temporal details contained in each channel. The CIS strategy filters the incoming speech into eight bands, obtains the speech envelope and compresses the signal for each channel. Stimulation consists of interleaved digital pulses that sweep rapidly through the channels at a rate of 833 pulses per second when using all eight

channels for a maximum pulse rate of 6,664 pulses per second (8 X 833 = 6,664). With the CIS strategy, rapid changes in the speech signal are tracked by rapid variations in pulse amplitude. The pulses are delivered to consecutive channels in sequence to avoid channel interaction.

The MED-EL COMBI 40-Cochlear Implant system utilizes the CIS (continuous interleaved sampling)-strategy that provides both spectral and temporal resolution. Up to eight active electrodes can be utilized. The electrode array used has the capability of deep insertion into the apical regions of the cochlea (Gstöttner, Baumgartner, Franz & Hamzavi, 1997). The MED-EL has the capacity to provide the most rapid stimulation rate of any of the currently available implants (maximum of 12,000 biphasic pulses per second) (Hochmair, 1996).

Surgical Implantation

Cochlear implantation in both children and adults requires meticulous attention to the delicate tissues and small dimensions. Skin incisions are designed to provide access to the mastoid process and coverage of the external portion of the implant package while preserving the blood supply of the postauricular skin. The incision employed at the Indiana University Medical Center has eliminated the need to develop a large postauricular flap. The inferior extent of the incision is made well posterior to the mastoid tip to preserve the branches of the postauricular artery. From here the incision is directed posterio-superiorly and then directed superiorly without a superior anterior limb. In children, the incision incorporates the temporalis muscle to give added thickness. A pocket is created for positioning the implant induction coil. Well anterior to the skin incision, the periosteum is incised from superior to inferior and a posterior periosteal flap is developed. At the completion of the procedure, the posterior periosteal flap is sutured to the skin flap compartmentalizing the induction coil from the skin incision. A bone well tailored to the device being implanted is created and the induction coil is fixed to the cortex with a fixation suture or periosteal flaps.

Following the development of the skin incision, a mastoidectomy is performed. The horizontal semicircular canal is identified in the depths of the mastoid antrum, and the short process of the incus is identified in the fossa incudis. The facial recess is opened using the fossa incudis as an initial landmark. The facial recess is a triangular area bounded by: 1) the fossa incudis superiorly; 2) the chorda tympani nerve laterally and anteriorly; and 3) the facial nerve medially and posteriorly. The facial nerve can usually be visualized through the bone without exposing it. The round window niche is visualized through the facial recess approximately 2 mm inferior to the stapes. Occasionally, the round window niche is posteriorly positioned and is not well visualized through the facial recess or is obscured by ossification. Particularly in these situations, it is important not to be misdirected by hypotympanic air cells. Entry into the scala tympani is best accomplished through a cochleostomy created anterior and inferior to the annulus of the round window membrane. A small fenestra slightly larger than the electrode to be implanted (usually 0.5 mm) is developed. A small diamond burr is used to “blue line” the endosteum of the scala tympani, and the endosteal membrane is removed with small picks. This approach bypasses the hook area of the scala tympani allowing direct insertion of the active electrode array. After insertion of the active electrode array, the round window is sealed with small pieces of fascia.

Special Surgical Considerations

In cases of cochlear dysplasia, a CSF gusher may be encountered. The senior author prefers to enter the cochlea through a small fenestra and tightly pack the electrode at the cochleostomy with fascia. The flow of CSF has been successfully controlled in this way. In patients with severe malformations of the

labyrinth the facial nerve may follow an aberrant course. In these cases the most direct access to a common cavity deformity may be by a transmastoid labyrinthotomy approach. The otic capsule is opened postero-superior to the second genu of the facial nerve and the common cavity is entered directly. Four patients have been treated in this way with no vestibular side effects (McElveen, Carrasco, Miyamoto, Lormore & Brown, in press).

In cases of cochlear ossification, our preference is to drill open the basal turn and create a tunnel approximately 6 mm in length and partially insert a Nucleus electrode. This allows implantation of 10 to 12 active electrodes that has yielded very satisfactory results. Gantz (1988) has described an extensive drill out procedure to gain access to the upper basal turn. The benefits of this extended procedure are under investigation. Steenerson (1990) has described the insertion of the active electrode into the scala vestibuli in cases of cochlear ossification. This procedure has merit. However, the scala vestibuli is frequently ossified when the scala tympani is completely obliterated.

Results of Cochlear Implantation in Children

The Nucleus Cochlear Implant Systems

Pediatric clinical trials with the Nucleus 22-channel cochlear implant began in 1986, and in 1990 the FDA approved this device for use in children. The children originally implanted with the Nucleus 22-channel system used the F0-F1-F2 feature extraction speech processing strategy. Children implanted after 1989 were provided with the MPEAK strategy, and the SPEAK strategy was approved in 1994. Pediatric clinical trials for the Nucleus 24-channel device with the SPEAK strategy were initiated in April 1997 and FDA approval was received in June 1998.

One of the first large-scale reports of pediatric performance with the Nucleus cochlear implant was presented by Staller et al. (1991). They presented speech perception data from 80 children with the Nucleus 22-channel cochlear implant system who were tested as part of the FDA clinical trials. The mean age at onset of deafness was 2 years, 8 months and the mean age at implantation was 9 years, 10 months for this group of children. The children's performance was classified by the highest category of speech perception achieved. Comparisons were made between their speech perception performance preimplant and again at 12-months postimplant. After 12 months of cochlear implant use, 63% of the children showed significant improvements in the closed-set speech perception tasks and 46% of the children demonstrated significant improvements on at least one open-set speech perception task. However, open-set speech abilities were still relatively modest. Similar word recognition results were reported by Osberger et al. (1991) for 28 children. Their results demonstrated that the children's speech perception abilities improved significantly after implantation with the largest gains noted when stimuli were presented in the auditory-plus-visual modality (i.e., with visual and lipreading cues). Thus, the majority of the children tested with the early Nucleus cochlear implant processing strategies demonstrated at least some open-set word recognition and performance was generally good when both auditory and visual cues were available.

The introduction of newer generation Nucleus processing strategies yielded greater speech perception benefits in children just as in adults. Osberger et al. (1996) compared the performance of six children who used the F0-F1-F2 processing strategy with that of six children who used the MPEAK strategy. The children in each group were matched by age at onset of deafness and age at implantation. After one year of implant use, the children with the MPEAK device were significantly better at discriminating vowel height and consonant place of articulation cues on the Minimal Pairs test. However, the two groups did not differ after three years of cochlear implant use. The authors concluded that children

show an accelerated rate of learning with improved speech processing strategies. Similar improvements have been noted for children who switch from the MPEAK to the SPEAK processing strategies (Cowan et al., 1997; Sehgal et al., 1998; Cowan et al., 1995). Sehgal et al. (1998) compared word recognition scores for children who switched from an earlier processing strategy to the SPEAK processing strategy. They reported mean monosyllabic word recognition scores increased from 28% words correct with the earlier strategy to 58% words correct with the SPEAK strategy.

The Clarion Cochlear Implant System

Pediatric clinical trials of the Clarion multichannel cochlear implant system began in 1995 and the device received FDA approval for use in children in 1997. Zimmerman-Phillips et al. (1997) summarized the initial results of the children's preoperative performance with hearing aids compared with their postoperative performance with the Clarion device. The mean age of the group of children implanted by 1996 was approximately five years (N=124). Data were reported for children tested at three-months postimplant (N=60) and six-months postimplant (N=23). After only three months of device use, mean scores were higher than the preimplant performance and many of the children demonstrated some open-set speech recognition. By six months postimplant, mean word recognition scores were 23% for the PB-K and 38% for a test of word recognition in a sentence context, the Glendonald Auditory Screening Procedure, or GASP (Erber, 1982). In a second study, Osberger et al. (1998) examined the performance of children implanted with the Clarion device after the age of five years who had at least six months of device experience (N=30). The children were divided into two groups based on communication method. After six months of device use, children in the Oral group correctly identified an average of 27% of the words on the PB-K. The average PB-K word score for children in the Total Communication group was 8% correct.

The Med-El Cochlear Implant System

Pediatric FDA clinical trials for the Med-El device were initiated in 1998. To date, too few children in the United States have used their devices long enough to draw conclusions regarding the benefits to be received by these children.

Summary of Pediatric Results

In summary, children with multichannel cochlear implants demonstrate significant improvements in closed-set speech discrimination, enhanced lipreading ability, and most obtain some open-set speech understanding with their devices. The rate of auditory skills development seems to be increasing as cochlear implant technology improves and cochlear implant candidacy is broadened to include younger children and children with more residual hearing. For example, early studies reported significant increases in the discrimination of nonsegmental speech cues after only six months of implant use. However, significant increases in the discrimination of vowels and consonant features were not evident until 1.5 years of cochlear implant experience and auditory-only open-set skills continued to improve long after this time period. More recent studies have shown that many children achieve open-set speech recognition within the first year of device use (Miyamoto et al., in press; Osberger et al., 1998) but these skills still continue to develop over time (Fryauf-Bertschy et al., 1992; Fryauf-Bertschy et al., 1997; Osberger et al., 1996; Miyamoto et al., 1994; Miyamoto et al., 1996). In fact, Miyamoto et al. (1994) noted continued improvements in spoken word recognition even after five years of multichannel cochlear implant use. These findings highlight the need to conduct longitudinal studies in order to determine the ultimate benefits of implant use in children.

Demographic Influences on CI Performance in Children

Age at onset of hearing loss, age at time of implantation, length of cochlear implant use, communication mode and amount of residual hearing prior to implantation are all demographic factors that have been shown to influence performance results. Early results demonstrated that age at onset and duration of deafness significantly affected speech perception performance (Zimmerman-Phillips, Osberger & Robbins, 1997; Fryauf-Bertschy et al., 1992; Osberger, Todd, Berry, Robbins & Miyamoto, 1991). That is, the children with a later onset of deafness and a shorter period of auditory deprivation prior to implantation had better speech perception skills than children who were deafened earlier and had a longer duration of deafness prior to implantation. When only children with prelingual deafness (i.e., < three years) are considered, age at onset of hearing loss is no longer a significant factor. The speech perception performance of children with congenital deafness is similar to that of children with adventitious deafness acquired prior to age three years (Osberger, Todd, Berry, Robbins & Miyamoto, 1991).

Age at Implantation

Previous studies have shown that earlier implantation yields superior cochlear implant performance. For example, Fyrauf-Bertschy et al. (1997) demonstrated that children implanted prior to age five had significantly better open-set word recognition than did those implanted at a later age. Similar results were reported by Miyamoto et al. (1997). Next, Waltzman and her colleagues conducted several studies to examine the speech perception abilities of children who were all implanted before the age of five years (between two and five years) (Waltzman & Cohen, 1998; Waltzman et al., 1997). Waltzman et al. (1995) reported the performance results of 14 children who were implanted prior to age three years and had used their device for at least three years. After one year of implant use, seven of the children demonstrated consistent open-set speech perception abilities. Following two years, this number increased to 13 children. The mean word recognition score at three-years postimplant was 47% correct. Similar performance results also were reported for a group of 11 children implanted prior to two years of age (Waltzman & Cohen, 1998).

Residual Hearing

The presence of preimplant residual hearing has also been shown to have a positive effect on postimplant speech perception performance. Zwolan and her colleagues (1997) compared the postoperative performance of 12 children who demonstrated some aided open-set speech recognition preimplant (the Borderline candidacy group) with that of 12 matched controls who had no preimplant speech recognition (the Traditional candidacy group). Candidacy for the study participation was based on preimplant binaural aided speech testing and the children were subsequently implanted in their poorer hearing ear. Thus, mean preoperative audiograms did not differ for the implanted ears in the two groups. By one year postimplant children in the Borderline group had significantly higher scores than children in the Traditional group on all six speech perception measures employed. The authors suggested that increased auditory experience prior to implantation facilitated the development of speech perception skills postimplant.

More recently, Gantz et al. (in press) demonstrated that children with greater residual hearing before implantation might achieve the highest levels of spoken word recognition with a cochlear implant. Gantz et al. suggested that children with limited preimplant residual hearing are better able to use the auditory information provided via a cochlear implant because they have more intact auditory systems, including inner hair cells, dendrites, ganglion cells, and central pathways, than their peers who have no preimplant residual hearing or word recognition.

Discussion

With the goal of universal detection of hearing loss in infants by three months of age, and appropriate intervention (e.g., amplification) by six months of age (American Academy of Pediatrics, 1994; 1999), it is likely that ever-increasing numbers of very young children will be identified as potential implant candidates. We know that early identification (i.e., by six months of age) and early intervention with HAs have a significant effect on language development in children with hearing loss (Yoshinaga-Itano, Sedey, Coulter & Mehl, 1998), but the spoken word recognition and receptive language benefits of early implantation in children with profound deafness have not been quantified and critical age limits for cochlear implantation have not been identified.

Cochlear implantation earlier than the current FDA accepted age of 18 months is feasible as the target organ, the cochlea, is adult size at birth. The small dimensions of the temporal bone must be accounted for but the facial recess and mastoid antrum that provide access to the middle ear for electrode placement are adequately developed prior to the age of one year. In fact, several centers have chosen to implant children under 18 months of age. Furthermore, implanting children under the age of eighteen months may have substantial advantages when the etiology of deafness is meningitis. Progressive intracochlear fibrosis and ossification may occur which can preclude standard electrode insertion. A relatively short time window exists during which this advancing process can be circumvented.

Nonetheless, implantation of the very young child remains controversial because the audiological assessment and management of this population is extremely challenging. As with older children, profound deafness must be substantiated and the inability to benefit from conventional hearing aids demonstrated. However, a compelling argument supporting implantation at the earliest possible time can be made because the development of speech perception, speech production, and language competence normally begins early in infancy. In addition, electrical stimulation has been shown to prevent at least some of the degenerative changes in the central auditory pathways caused by auditory deprivation (Matsushima, Shepard, Seldon, Xu & Clarl, 1991).

The extension of cochlear implantation to children with ever-higher levels of preimplant residual hearing should be approached cautiously. Surgical implantation of the electrode array results in the loss of residual hearing in that ear. Thus, cochlear implantation should not be considered unless it seems likely that a given child will receive more benefit from this device than from conventional amplification. Recently, mounting evidence has been found to suggest that some children with severe hearing loss may derive as much or even more benefit from a cochlear implant than from a well-fitted HA. In amplifying sound for an individual with hearing loss, an assumption is made that the acoustic-phonetic patterns of speech must be detected before they can be discriminated and recognized. To accomplish this goal, audibility across a broad frequency range is typically prescribed as a means of maximizing speech intelligibility (Skinner & Miller, 1983). This, in fact, has been the goal of most standard HA prescriptions. For severe-to-profound losses, however, supplying adequate gain across a broad frequency range can present a special challenge to the clinician. Moreover, achieving this amount of amplification may cause acoustic feedback, necessitating a reduction in gain and audibility (Skinner, Holden & Binzer, 1996). Another issue concerns the risk of delivering high levels of sound to the impaired ear. According to Macrae (1991a, b), the sound pressure level (SPL) required to achieve audibility for individuals with severe-to-profound hearing loss has the potential to destroy remaining hair cells due to excessive noise exposure. Thus, a trade-off may exist between providing audible speech and risking increased damage to inner ear structures. Lastly, there is some question as to the extent of benefit that may actually be realized by amplifying high frequencies to

audible levels for this magnitude of loss. Recent research has suggested that provision of adequate audibility for losses greater than 60 dB HL at 3000 Hz and higher does not improve speech recognition and may even degrade performance (Ching, Dillon & Byrne, 1998; Hogan & Turner, 1998; Turner, 1999). Preliminary research has suggested that some children with cochlear implants obtain spoken word recognition abilities that surpass those of other children with severe hearing loss (i.e., PTAs between 70-90 dB HL) who use well-fit HAs (Eisenberg et al., in press; Boothroyd, 1997; Levi, Eisenberg, Martinez & Schneider, 1998). Given the limitations imposed in providing high levels of amplified speech to children with severe-to-profound hearing loss, the evidence suggests that a cochlear implant could provide added benefit for a select population of children with this magnitude of hearing loss.

The encouraging results obtained with younger children and those who have some useful hearing prior to implantation have led investigators to push the boundaries of cochlear implantation criteria farther than ever before. With the continued evolution and expansion of cochlear implant candidacy it is crucial that we develop techniques to quantify hearing loss, to fit both hearing aids and cochlear implants, and to document the effects of implantation in these very young children.

References

American Academy of Pediatrics. (1994). Joint Committee on Infant Hearing Screening. 1994 position statement. *Pediatrics, 95,* 152-156.

American Academy of Pediatrics. (1999). Task Force on Newborn and Infant Hearing. Newborn and infant hearing loss: Detection and intervention. *Pediatrics, 103,* 527-530.

Bollard, P.M., Chute, P.M., Popp, A., & Parisier, S.C. (1999). Specific language growth in young children using the Clarion cochlear implant. *Annals of Otology, Rhinology, & Laryngology, 108 (Suppl. 117),* 119-123.

Boothroyd, A. (1997). Auditory capacity of hearing-impaired children using hearing aids and cochlear implants: Issues of efficacy and assessment. *Scandinavian Audiology, 26 (Suppl. 46),* 17-25.

Ching, T.Y.C., Dillon, H., & Byrne, D. (1998). Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification. *Journal of the Acoustical Society of America, 103,* 1128-1140.

Clark, G: The University of Melbourne Nucleus multi-electrode cochlear implant. Adv Otol Rhinol Laryngol 38:189, 1987.

Cowan RSC, Brown C, Whitford LA, Galvin KL, Sarant JZ, Barker EJ, Shaw S, King A, Skok M, Seligman P, Dowell M, Everingham C, Gibson WPR, Clark GM (1995). Speech perception in children using the advanced SPEAK speech-processing strategy. *Annals of Otology, Rhinology, and Laryngology, 104(Suppl. 166*), 318-321.

Cowan, R.S.C., DelDot, J. Barker, E.J., Sarant, J.Z., Pegg, P., Dettmen, S., Galvin, K.L., Rance, G., Hollow, R., Dowell, R.C., Pyman, B., Gibson, W.P.R., & Clark, G.M. (1997a). Speech perception results for children with implants with different levels of preoperative residual hearing. *American Journal of Otology, 18,* 125-126.

Cowan, R.S.C., Galvin, K.L., Klieve, S., Barker, E.J., Sarant, J.Z., Dettman, S., Hollow, R., Rance, G., Dowell, R.C., Pyman, B., & Clark, G.M. (1997b). Contributing factors to improved speech perception in children using the Nucleus 22-channel cochlear prosthesis. In I. Honjo & H. Takahashi (Eds.), *Cochlear implant and related sciences update.* (*Advances in Otorhinolaryngology, 52*) (pp. 193-197). Basel: Karger.

Eisenberg, L.S., Martinez, A.S., Sennaroglu, G, & Osberger, M.J. (in press). Establishing new criteria in selecting children for a cochlear implant: Performance of "platinum" hearing aid users. *Annals of Otology, Rhinology, & Laryngology.*

Erber NP (1982). *Auditory Training*. Washington, DC: Alexander Graham Bell Association for the Deaf.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D.M., & Gantz, B.J. (1992). Performance over time of congenitally deaf and postlingually deafened children using a multichannel cochlear implant. *Journal of Speech and Hearing Research, 35,* 913-920.

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 and Hearing Research, 40,* 183-199.

Gantz BJ, McCabe BF, Tyler RS: Use of multichannel cochlear implants in obstructed and obliterated cochleas. *Otolaryngol Head Neck Surg 98*:72-81, 1988.

Gantz, B., Rubinstein, J., Tyler, R., Teagle, H., Cohen, N., Waltzman, S., Miyamoto, R., & Kirk, K. (in press). Long term results of cochlear implants in children with residual hearing. *Annals of Otology, Rhinology, & Laryngology.*

Gantz BJ, Tyler RS, Woodworth G, Tye-Murray N, Fryauf-Bertschy H (1994). Results of multichannel cochlear implant in congenital and acquired prelingual deafness in children: Five-year follow-up. *American Journal of Otology, 15*(Suppl. 2), 1-8.

Gstöttner WK, Baumgartner WD, Franz P, Hamzavi J (1997). Cochlear implant deep-insertion surgery. *Laryngoscope, 107,* 544-546.

Hochmair, E.S. (1996). Clinically relevant aspects of the high-rate cis-speech coding strategy for cochlear implants. Abstracts of the *First Asia Pacific Symposium on Cochlear Implant and Related Sciences*. Abstract 19, p. 47.

Hogan, C.A., & Turner, C.W. (1998). High-frequency audibility: Benefits for hearing-impaired listeners. *Journal of the Acoustical Society of America, 104,* 432-441.

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

Levi, A., Eisenberg, L.S., Martinez, A.S., & Schneider, K. (1998, June). Performance comparisons between cochlear implant and "platinum" hearing aid user: Case Study. Paper presented at the *Seventh Symposium on Cochlear Implants in Children*, Iowa City, IA.

Macrae, J. H. (1991a). Permanent threshold shift association with overamplification by hearing aids. *Journal of Speech and Hearing Research, 34*, 403-414.

Macrae, J. H. (1991b). Prediction of deterioration in hearing due to hearing aid use. *Journal of Speech and Hearing Research, 34*, 661-660.

Matsushima, J.I., Shepard R.K., Seldon, H.L. Xu, S.A. Clarl, G.M. (1991). Electrical stimulation of the auditory nerve in deaf kittens: Effects on cochlear nucleus morphology. *Hear Res 56,* 133-42.

McElveen JT, Carrasco VN, Miyamoto RT, Lormore KA, Brown C. (in press). Surgical approaches for cochlear implantation in patients with cochlear malformations.

Miyamoto, R.T., Kirk, K.I., Robbins, A.M., Todd, S., Riley, A., & Pisoni, D.B. (1997). Speech perception and speech intelligibility in children with multichannel cochlear implants. In I. Honjo & H. Takahashi (Eds.), *Advances in Otorhinolaryngology Vol. 52: Cochlear implant and related sciences update* (pp. 198-203) . Basel: Karger.

Miyamoto, R.T., Kirk, K.I., Robbins, A.M., Todd, S., & Riley, A. (1996). Speech perception and speech production skills of children with multichannel cochlear implants. *Acta Otolaryngologica, 116*, 240-243.

Miyamoto, R.T., Kirk, K.I., Svirsky, M.A., & Sehgal, S.T. (in press). Communication skills in pediatric cochlear implant recipients. *Acta Otolaryngologica (Stockholm).*

Miyamoto, R.T., Osberger, M.J., Todd, S.L., Robbins, A.M., Stroer, B.S., Zimmerman-Phillips, S., & Carney, A.E., (1994). Variables affecting implant performance in children. *Laryngoscope, 9,* 1120-1124.

Miyamoto, R.T., Osberger, M.J., Robbins, A.M., Renshaw, J.J., Myres, W.A., Kessler, K., & Pope, M.L. (1989). Comparison of sensory aids in deaf children. *Annals of Otology, Rhinology, & Laryngology, 98,* 2-7.

Osberger, M.J., Fisher, L., Zimmerman-Phillips, S., Geier, L., & Barker, M.J. (1998). Speech recognition performance of older children with cochlear implants. *American Journal of Otology, 19,* 152-157.

Osberger, M.J., Miyamoto, R.T., Zimmerman-Phillips, S., Kemink, J.L., Stroer, B.S., Firzst, J.B., & Novak, M.A. (1991). Independent evaluation of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear and Hearing, 12(Suppl.),* 66S-80S.

Osberger, M.J., Robbins, A.M., Todd, S.L., Riley, A.I., Kirk, K.I., & Carney, A.E. (1996). Cochlear implants and tactile aids for children with profound hearing impairment. In F. Bess, J. Gravel, & A.M. Tharpe (Eds.), *Amplification for children with auditory deficits* (pp. 283-308). Nashville, TN: Bill Wilkerson Center Press.

Osberger MJ, Todd SL, Berry SW, Robbins AM, Miyamoto RT (1991). Effect of age at onset of deafness on children's speech perception abilities with a cochlear implant. *Annals of Otology, Rhinology, and Laryngology, 100,* 883-888.

Pisoni, D.B., Svirsky, M.A., Kirk, K.I., & Miyamoto, R.T. (submitted). Looking at the "stars": A first report on the intercorrelations among measures of speech perception, intelligibility and language in pediatric cochlear implant users. *Journal of Speech, Language and Hearing Research.*

Robbins, A.M., Bollard, P.M., & Green, J. (1999). Language development in children implanted with the Clarion cochlear implant. *Annals of Otology, Rhinology, & Laryngology, 108(Suppl. 117),* 113-118.

Robbins, A.M., Svirsky, M.A., & Kirk, K.I. (1997). Children with implants can speak, but can they communicate? *Otolaryngology–Head and Neck Surgery, 117,* 155-160.

Schindler RA, Kessler DK, Rebscher SJ, Yanda JL, et al. (1986). The UCSF/Storz multichannel cochlear implant: Patient results. *Laryngoscope 96,* 597.

Sehgal, S.T., Kirk, K.I., Svirsky, M.A., & Miyamoto, R.T. (1998). The effects of processor strategy on the speech perception performance of pediatric Nucleus multichannel cochlear implant users. *Ear and Hearing, 19,* 149-161.

Skinner, M.W., & Miller, J.D. (1983). Amplification bandwidth and intelligibility of speech in quiet and noise for listeners with sensorineural hearing loss. *Audiology, 22,* 253-279.

Skinner, M.W., Holden, L.K., & Binzer, S.M. (1996). Aural rehabilitation for individuals with severe and profound hearing impairment: Hearing aids, cochlear implants, counseling, and training. In M. Valente (Ed.), *Strategies of selecting and verifying hearing aid fittings* (Chapter 13). New York: Thieme Medical Publishers, Inc.

Staller, S., Beiter, 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.

Staller, S.J., (1998). Clinical trials of the Nucleus 24 in adults and children. Presented at the *Tenth Annual Convention of the American Academy of Audiology*, Los Angeles, CA.

Steenerson RL, Gary LB, Wynens MS. (1990). Scala vestibuli cochlear implantations for labyrinthine ossification. *Am J Otol 11,* 360-3.

Svirsky, M.A., Sloan, R.B., Caldwell, M., & Miyamoto, R.T. (1998, June). Speech intelligibility of prelingually deaf children with multichannel cochlear implants. Paper presented at *the Seventh Symposium on Cochlear Implants in Children*, Iowa City, IA.

Turner, C. W. (1999). The limits of high-frequency amplification. *The Hearing Journal, 52,* 10-14.

Tyler, R.S., Fryauf-Bertschy, H., Gantz, B.J., Kelsay, D.M.R., & Woodworth, G.G. (1997a). Speech perception in prelingually implanted children after four years. In I. Honjo H. Takahashi (Eds.), *Cochlear implant and related science update.* (*Advances in Otolaryngology 52*) (pp. 187-192). Basel: Karger.

Tyler, R.S., Fryauf-Bertschy, H., Kelsay, D.M.R., Gantz, B.J., Woodworth, G.G., & Parkinson, A. (1997b). Speech perception by prelingually deaf children using cochlear implants. *Otolaryngology-Head and Neck Surgery, 117,* 180-187.

Tyler, R.S., Parkinson, A.J., Fryauf-Bertschy, H., Lowder, M.W., Parkinson, W.S., Gantz, B.J., & Kelsay, D.M.R. (1997c). Speech perception by prelingually deaf children and postlingually deaf adults with cochlear implant. *Scandinavian Journal of Audiology, 26(Suppl.46),* 65-71.

Tyler, R.S., Tomblin, J.B., Spencer, L.J., Kelsay, D.M.R., & Fryauf-Bertschy, H. (in press). How speech perception through a cochlear implant affects language and education. *Otolaryngology-Head and Neck Surgery.*

Waltzman, S., & Cohen, N.L. (1998). Cochlear implantation in children younger than 2 years old. *The American Journal of Otology, 19,* 158-162.

Waltzman, S., Cohen, N.L., & Shapiro, W. (1995). Effects of cochlear implantation on the young deaf child. In A.S. Uziel, M. Mondain (Eds.), *Cochlear implants in children* (*Advances in Otorhinolaryngology, 50*) (pp. 125-128). Basel: Karger.

Waltzman S, Cohen NL, Gomolin R, Green J, Shapiro W, Brackett D, Zara C (1997). Perception and production results in children implanted between two and five years of age. In I Honjo, H Takahashi (Eds.), *Cochlear Implant and Related Sciences Update. Advances in Otorhinolaryngology, 52*, 177-180, Basel: Karger.

Wilson, B.S. (1993). Signal processing. In R.S. Tyler (Ed.), *Cochlear implants: Audiological foundations* (pp. 35-85). San Diego: Singular Publishing.

Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K., Rabinowitz, W.M. (1991). Better speech recognition with cochlear implants. *Nature, 352*, 236-237.

Wilson, B.S., Lawson, D.T., Finley, C.C., Wolford, R.D. (1991). Coding strategies for multichannel cochlear prostheses. *American Journal of Otology, 12(Suppl. 1),* 56-61.

Wilson, B.S., Lawson, D.T., Finley, C.C., Wolford, R.D. (1993). Importance of patient and processor variables in determining outcomes with cochlear implants. *Journal of Speech and Hearing Research, 36,* 373-379.

Woodward, A.L., & Markman, E.M. (1997). Early word learning. In W. Damon, D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology, Vol.2: Cognition, perception and language* (pp. 371-420). New York: Wiley.

Yoshinaga-Itano, C., Sedey, A.L., Coulter, D.K., & Mehl, A.L. (1998). Language of early- and later-identified children with hearing loss. *Pediatrics, 102*, 1161-1171.

Zimmerman-Phillips, S., Osberger, M.J., Geier, L., & Barker, M. (1997). Speech recognition performance of pediatric Clarion patients. *The American Journal of Otology, 18*, 5153-5154.

Zimmerman-Phillips, S., Osberger, M.J., & Robbins, A.M. (1997). Infant toddler meaningful auditory integration scale (IT-MAIS). Sylmar, CA: Advanced Bionics Corporation.

Zwolan, T.A., Zimmerman-Phillips, S., Asbaugh, C.J., Hieber, S.J., Kileny, P.R., & Telian, S.A. (1997). Cochlear implantation of children with minimal open-set speech recognition skills. *Ear and Hearing, 18,* 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Early Implantation and the Development of Communication Abilities in Children[1]

Richard T. Miyamoto,[2] Karen I. Kirk,[2] Susan T. Sehgal,[2] Cara Lento,[2] and Julie Wirth[2]

DeVault Otologic Research Laboratory
Department of Otolaryngology-Head & Neck Surgery
Indiana University School of Medicine
Indianapolis, IN 46202

[1] This work was supported by NIH-NIDCD Grants RO1 DC00064, RO1 DC00423 and KO8 DC00126 to Indiana University. Presented at the Annual Meeting of the American Neurotology Society, Palm Springs, CA, April 1999.
[2] Also, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405

Early Implantation and the Development of Communication Abilities in Children

Abstract. Early cochlear implantation of children with prelingual, profound deafness promotes the acquisition of communication abilities. Previously, we found that children implanted prior to age six years had superior communication abilities than children implanted after that time. This study examined the effects of age at implantation and communication mode on the development of listening skills in age-matched children with prelingual deafness who were implanted during the preschool years. Participants were deafened before age three years, implanted between the ages of two and five years, and used the Nucleus cochlear implant with the SPEAK speech processing strategy. The 28 participants were approximately five and a half years old at the time of testing. Closed- and open-set spoken word recognition and receptive and expressive language abilities were compared as a function of age at implantation and communication method. We also consider the case of a child implanted at 16 months of age and compare his performance with that of the other 28 age-matched participants. The results show that children implanted prior to age three years had higher spoken word recognition than those implanted later. Also, children who used oral communication had significantly better open-set speech understanding than children who used total communication (i.e., the combined use of spoken and signed English). This suggests that early implantation yields significant advantages in children's ability to encode, process and produce spoken language.

Introduction

The purpose of this investigation was to examine the benefits of early cochlear implantation on the development of communication skills in children with congenital or prelingual, profound deafness. Detailed longitudinal studies of speech perception, speech production and language acquisition have justified a trend toward cochlear implantation at young ages (Frauf-Bertschy, Tyler, Kelsay, Gantz & Woodworth, 1997; Meyer, Svirsky, Kirk & Miyamoto, 1998; Robbins, Svirsky, Kirk, Miyamoto, Bollard & Green, in press). We have previously shown that children with congenital deafness have the potential to derive similar speech perception benefit from a multichannel cochlear implant as do children who had some limited exposure to spoken language before the onset of their deafness (Miyamoto, Osberger, Robbins, Myres & Kessler, 1993). In addition, previous studies have shown that some children with profound hearing loss who have received a cochlear implant can attain speech perception abilities similar to other profoundly deaf children who have some residual hearing and use conventional amplification (Cowan et al., 1997; Miyamoto, Osberger, Robbins et al., 1991; Staller, Dowell, Beiter & Brimacombe, 1991; Miyamoto, Kirk, Robbins, Todd, Riley & Pisoni, 1997; Snik, Vermuelen, Brokx & van den Broek, 1997). However, the speech production (Geers & Tobey, 1992; Waltzman et al., 1997; Miyamoto, Svirsky, Kirk, Robbins, Todd & Riley, 1997) and language skills (Miyamoto, Svirsky & Robbins, 1997; Hasenstab & Tobey, 1991) of most profoundly deaf children, even with a cochlear implant, are significantly below those of their age peers with normal hearing. Early implantation offers children access to spoken language during the preschool years that are so important to language acquisition. This, in turn, may have important consequences for the development of spoken language and ultimately, academic success.

In this report, we will consider the case of the youngest child implanted to date at Indiana University School of Medicine. (We will refer to this child by his participant code, SHJ.) SHJ was implanted at 16 months of age and has used his cochlear implant for approximately four years. He is now approximately 5.5 years of age and is approaching the age at which formal schooling will begin. We will

compare his communication abilities with those of age-matched peers who were implanted between the ages of two to five years.

Case Study Participant

SHJ, a male child with congenital, profound hearing loss, was the product of a normal pregnancy and delivery. Brainstem Evoked Response Audiometry conducted when SHJ was 11 months of age suggested a bilateral, profound hearing loss. SHJ was first seen in the Department of Otolaryngology at Indiana University School of Medicine for follow-up testing when he was 12 months old, and he was fit with bilateral conventional amplification at the age of 13 months. SHJ and his parents quickly began a family-centered habilitation training program with a speech-language pathologist to help SHJ acquire oral-aural communication. After a three-month trial of amplification and appropriate intervention, audiologists concluded that SHJ derived little or no benefit from his hearing aids. SHJ was implanted with the Nucleus 22-channel cochlear implant at the age of 16 months. His Spectra speech processor was programmed with the SPEAK processing strategy one month after surgery.

SHJ's parents agreed to have him participate in the longitudinal study at Indiana University School of Medicine to assess the benefits of pediatric cochlear implantation. As part of this study, children are administered a battery of speech perception, speech production and language measures prior to implantation, again at six month-intervals during the first three years of implant use, and then annually thereafter. The test battery administered to all young children is described below.

Communication Assessments and Procedures

Speech Perception Tasks

All speech perception testing for very young children is carried out in the auditory-only modality with stimuli presented via live voice at approximately 70 dB SPL. Performance on both a closed-set and open-set speech perception measure is assessed. The Grammatical Analysis of Elicited Language—Presentence Level (GAEL-P) (Moog, Kozak & Geers, 1983) has been adapted for use as a closed-set speech perception measure. Children are first familiarized with the 30 objects in the auditory-plus-visual modality. During testing, the children must identify the target word from a set of four objects at a time, through listening alone. The test is scored as the percent of words correctly identified; chance performance is 25%.

The Mr. Potato Head Task (Robbins, 1994) is a modified open-set speech perception task. Mr. Potato Head is a toy for children consisting of a plastic "potato" body with approximately 30 body parts and accessories that can be attached. Children are asked to carry out 10 commands presented auditorily only to assemble the Mr. Potato Head toy (e.g., "Give him some green shoes"). A sentence score is generated for the number of commands correctly carried out, and a word score is generated for the number of key words (out of 20) correctly identified even if the command was not followed correctly (e.g., if the child picked up the green shoes but did not put them on the toy).

Speech Intelligibility Task

The intelligibility of the children's speech production is evaluated using a procedure developed in our laboratory (Miyamoto et al., 1997). Each child is asked to repeat 10 simple sentences from the Beginner's Intelligibility Test (Osberger, Robbins, Todd & Riley, 1994) and these productions are recorded

on audiocassette. Each child's sentences are then digitized and randomized for presentation to separate panels of three listeners with normal hearing who have no experience in listening to the speech of deaf talkers. The listeners are asked to write down each sentence as they hear it; intelligibility is measured as the mean percent of words correctly identified across the panel of three listeners.

Table 1.

Speech perception, intelligibility and language scores obtained by SHJ over time.

Testing Interval	GAEL-P (Closed-set)	Potato Head Task (Open-Set)		Speech Intelligibility	Reynell Language Quotients	
	Words	Words	Sentences		Receptive	Expressive
Preimplant	CNT*	CNT*	CNT*	0%	0.0	0.18
Postimplant						
6 months	CNT*	CNT*	CNT*	0%	0.3	0.65
12 months	67%	0%	20%	0%	0.59	0.72
18 months	93%	50%	60%	11%	0.83	0.89
24 months	87%	70%	70%	40%	0.76	0.95
30 months	93%	40%	55%	31%	0.77	0.83
36 months	100%	100%	100%	70%	0.72	1.26
48 months	100%	100%	100%	NA^	0.82	0.92

* Could not test due to vocabulary constraints

^ Intelligibility sentences collected at this interval have not yet been played to panels of listeners

Language Task

The Reynell Developmental Language Scales (Reynell & Huntley, 1985) are administered to assess independently receptive and expressive language abilities. The test is administered in the child's preferred communication mode (oral or total communication). Normative data have been obtained for this measure from more than 1,300 children with normal hearing. The Reynell Developmental Language Scales are appropriate for a broad age range (between 1-7 years) and have been used extensively with deaf children. The test requires the children to comprehend or express a hierarchy of language structures ranging from object labeling to complex instructions. This test has high face validity and reflects real-world communication demands similar to those that children might encounter in daily living situations. The Reynell Language Development Scales yield separate receptive and expressive vocabulary ages. The vocabulary age scores are then converted into separate language quotients (vocabulary age divided by chronological age). A language quotient of 1.0 indicates that the child's language age and chronological age

are equal. Language quotients <1.0 indicate that the child's vocabulary age lags behind his or her chronological age; >1.0 indicate that the vocabulary age exceeds the chronological age.

Longitudinal Results for SHJ

Table 1 presents SHJ's speech perception and speech intelligibility scores and his Reynell receptive and expressive language quotients obtained prior to implantation and then over four years of implant use. Before receiving his implant, SHJ had minimal communication abilities. Vocalizations were inconsistent and primarily consisted of undifferentiated vowel sounds. He was unable to identify or label any objects on either the speech perception tasks or the Reynell Developmental Language Scales.

After 12 months of cochlear implant use, SHJ identified words from a closed-set at significantly greater than chance levels and demonstrated some limited open-set sentence recognition. By 36 months postimplant, SHJ had reached ceiling levels of performance on both closed- and open-set word recognition tasks. Similar improvements were noted in speech intelligibility. Although none of SHJ's speech attempts were intelligible preimplant, by 36 months postimplant, 70% of the sentences he produced could be understood by naive listeners. This represents a very high level of speech intelligibility, as there were no contextual or visual cues to aid naive listeners in understanding SHJ's speech. Finally, SHJ has also shown a marked increase in his ability to comprehend and produce spoken language. Large gains in his expressive language skills were evident as early as six months postimplant with corresponding increases in receptive language noted at 12 to 18 months postimplant. Now, after three to four years of cochlear implant experience, SHJ's language skills are nearly age-appropriate. Most importantly, SHJ is able to communicate effectively in daily communication situations with both familiar and unfamiliar persons. This fall, SHJ will be fully mainstreamed in a public school kindergarten classroom. He continues to see a speech-language pathologist on a weekly basis.

A Comparison of SHJ and his Age-Matched Peers

In order to examine more carefully the effects of early implantation, SHJ's speech perception and language scores obtained at his most recent testing interval (four years postimplant) were compared with those of his age-matched peers who received a cochlear implant prior to age six years. In addition, the communication abilities of these age-matched peers were examined as a function of age at time of implantation. Separate analyses were carried out for children in the comparison group who used oral and total communication (the combined use of signed and spoken English).

Comparison Participants

The children in the comparison group were selected from the population of children with prelingual deafness who were implanted at Indiana University School of Medicine. Selection criteria included: 1) implantation prior to age 6 years; 2) use of the most recent speech processing strategies (SPEAK or CIS); and 3) chronological age similar to that of SHJ at his most recent test interval (i.e, 5.5 years) Twenty-eight children met these criteria; all children were implanted with the Nucleus cochlear implant and used the SPEAK speech processing strategy. The participants were divided into three groups based on age at implantation. The groupings consisted of children implanted before the age of 3 years (n=10), those implanted between 3 years and 3 years, 11 months (n=8), and those implanted between 4 years and 5 years, 4 months (n=10). In each age group, some of the children used oral communication and some used total communication. Tables 2 and 3 present demographic information for these oral and total communication participants, respectively.

Comparison Study Procedures

The children selected for the comparison group had been administered the same test battery at the same intervals as SHJ. For each comparison participant, we analyzed data from the test interval at which his or her chronological age was closest to 5 years, 5 months (range = 5.0-5.9 years). This was done because we wished to compare performance across groups as a function of age at implantation while controlling for chronological age at the time of testing; we also wished to compare SHJ's performance with that of his age-matched peers. Thus, the mean length of device use differed across the three age-at-implantation groups. For children within the comparison group, scores on each of the measures described below were subjected to a two-way analysis of variance with age-at-implantation group and communication mode as the independent variables.

Table 2.

Oral participant characteristics as a function of age at time of implant.

	Age at Time of Implant					
	< 3 yrs (n=7)		3.0 - 3 yrs, 11 mos (n=5)		4.0 - 5 yrs, 4 mos (n=5)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Age at Onset (yrs)	0.3	(0.7)	0.4	(0.8)	0.7	(0.5)
Length of CI Use (yrs)	2.9	(0.3)	1.98	(0.2)	0.9	(0.5)
Age at Testing (yrs)	5.4	(0.2)	5.5	(0.3)	5.7	(0.2)
Pure Tone Average (dB HL)	111	(8.2)	110	(1.6)	112	(5.9)

Table 3.

Total communication participant characteristics as a function of age at time of implant.

	Age at Time of Implant					
	< 3 yrs (n=3)		3.0 - 3 yrs, 11 mos (n=3)		4.0 - 5 yrs, 4 mos (n=5)	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
Age at Onset (yrs)	0.0	(0.0)	0.2	(0.4)	0.0	(0.0)

Length of CI Use (yrs)	3.0	(0.1)	1.9	(0.2)	0.8	(0.3)
Age at Testing (yrs)	5.7	(0.2)	5.3	(0.1)	5.3	(0.1)
Pure Tone Average (dB HL)	116	(3.6)	110	(1.6)	113	(7.3)

Comparison Study Results

Figures 1-5 illustrate the communication abilities of SHJ compared with the mean scores for children in the three age-at-implantation comparison groups. In each of these figures, the score for SHJ is plotted to the left of the mean score for children who were implanted prior to age 3 years.

Speech Perception Results

Figure 1 illustrates the percent of words correctly identified on the closed-set measure, the GAEL-P. For the comparison group, age at time of implantation significantly influenced closed-set word identification ($F(2,20) = 4.25$; $p = .03$). Mean GAEL-P scores were 96%, 86%, and 62%, respectively, for children in the <3 years, 3-4 years, and 4-5 years age-at-implant groups. Children implanted before the age of three years had significantly better closed-set word recognition than the children in the two remaining groups. Finally, there was a trend for the oral children to have higher GAEL-P scores than the children who used total communication (88% vs. 76% words correct, respectively) but this difference was not significant. SHJ's score of 100% was similar to the mean performance of the oral children in the <3 years age at implantation group.

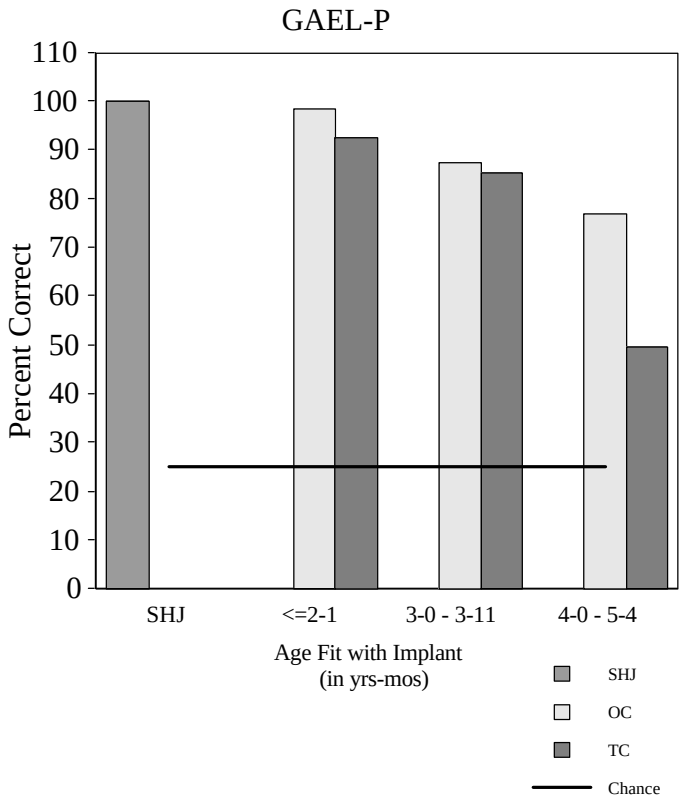


Figure 1. Percentage of words correctly identified on the GAEL-P, a closed-set measure of speech perception, as a function of age fit with an implant.

Figures 2 and 3 present the percent of words and sentences correctly identified, respectively, on the open-set Mr. Potato Head Task. Differences in both open-set word and sentence recognition performance among the three age-at-implantation groups were marginally significant. Mean word recognition scores were 84%, 63%, and 49%, respectively, for children in the <3 years, 3-4 years, and 4-5 years age-at-implantation groups ($F(2,22) = 3.25$; $p = .058$). For the same three groups, mean open-set sentence recognition scores were 80%, 60%, and 44% ($F(2,22) = 2.71$; $p = .089$). Finally, the oral children had significantly higher open-set word recognition ($F(2,22) = 6.42$; $p = .02$) and sentence recognition ($F(2,22) = 5.60$; $p = .03$) than did children who used total communication. SHJ's score of 100% for both word and sentence recognition exceeded the mean score of the oral children in the earliest age-at-implantation group. In fact, only one of the 28 children in the three comparison groups also achieved a perfect score on both of these open-set measures.

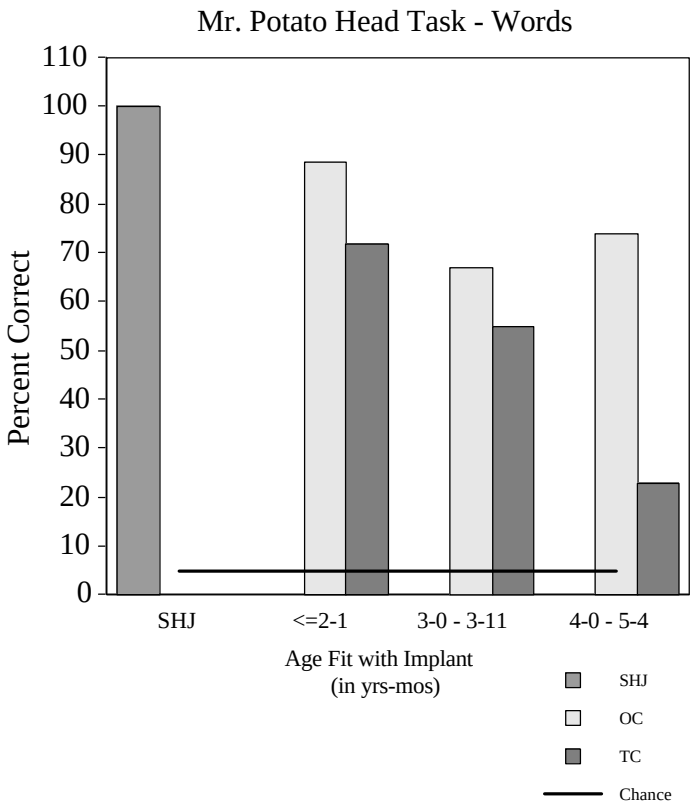


Figure 2. Percentage of words correctly identified on the Mr. Potato Head Task, an open-set measure of speech perception, as a function of age fit with an implant.

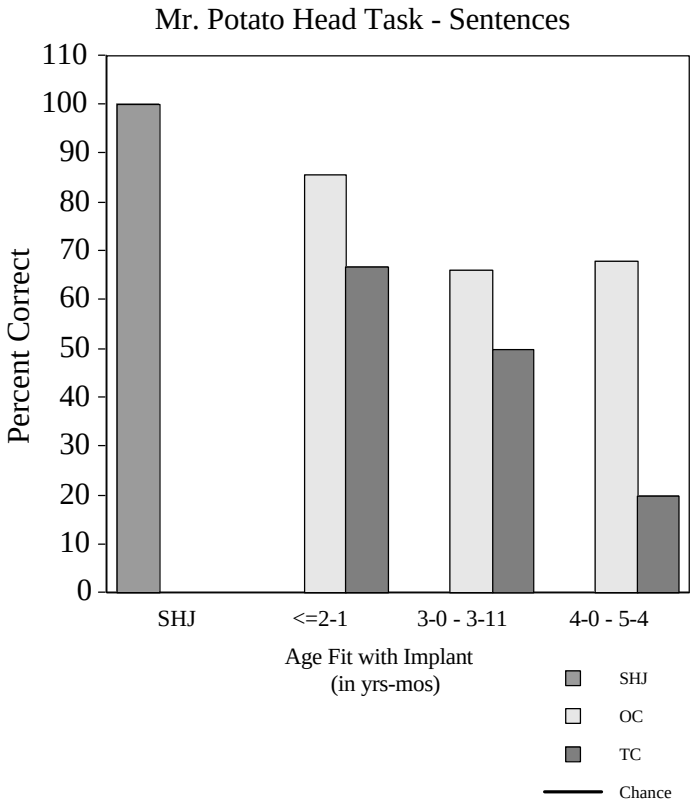


Figure 3. Percentage of sentences correctly identified on the Mr. Potato Head Task, an open-set measure of speech perception, as a function of age fit with an implant.

Receptive and Expressive Language Results

Figure 4 illustrates the receptive language quotients obtained on the Reynell Developmental Language Scales. Recall that a language quotient of 1.0 indicates that the child's receptive language age was equal to his or her chronological age. On this measure, children in the three comparison groups did not differ significantly as a function of age at time of implantation. Mean receptive language quotients for children in the comparison groups ranged from 0.57-0.60, indicating a mean vocabulary age that is somewhere near half of the mean chronological age. In addition, there were no significant differences in receptive language abilities as a function of communication method. Individual variability within the three comparison groups was noted, however. Across the 28 age-matched children, receptive language quotients ranged from 0.25 to 1.23. There was a trend for children who were implanted by around four years of age and used oral communication to demonstrate the highest language quotients. SHJ's receptive language quotient of 0.82 was higher than all but two of the children across the three comparison groups.

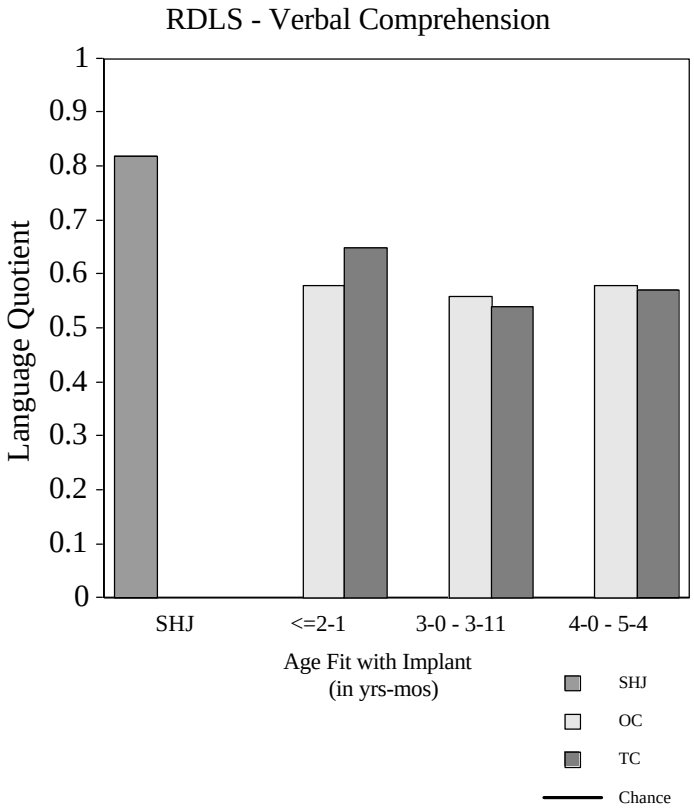


Figure 4. Receptive language quotient on the Reynell Developmental Language Scales as a function of age fit with an implant.

Figure 5 presents the expressive language quotients from the Reynell. Across age-at-implant groups, expressive language quotients ranged from 0.52-0.63. For all of the children in the comparison group, individual expressive language quotients ranged from 0.23-1.27. There were no significant effects of age at implantation or communication mode. Once again, there was a trend for children implanted prior to age 4 years to have higher language quotients, but these differences were not significant. SHJ's expressive language quotient of 0.92 was higher than that obtained by the majority of children in the comparison groups. Four children in the comparison groups achieved expressive language quotients that were similar to SHJ's (range = 0.91-1.27). Of these four, three received their cochlear implant before the age of four years.

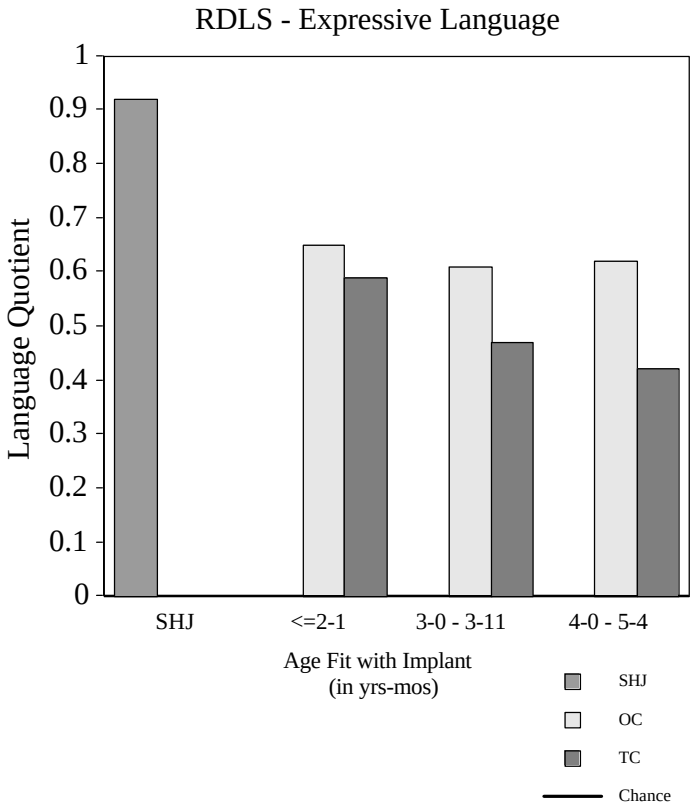


Figure 5. Expressive language quotient on the Reynell Developmental Language Scales as a function of age fit with an implant.

General Discussion

The current results demonstrated that both age at implantation and communication mode have significant effects on the development of a number of communication skills in profoundly deaf children. Our comparisons of communication abilities among age-matched peers who differed in the age at implantation revealed that children implanted by around three years of age generally had higher spoken word recognition than children who were implanted after that time. In contrast, there were minimal differences among the average language quotients of the three age-at-implantation groups on the receptive or expressive language measures. The effects of early implantation on receptive and expressive language abilities may emerge over a longer time period than that examined here.

Children in this study who used oral communication had significantly better spoken word recognition abilities compared to children who used total communication. This finding should be interpreted with some caution as children in the two communication groups were not matched on any variables other than chronological age. However, our results support earlier findings from our laboratory and suggest that a strong emphasis on the development of speaking and listening abilities promotes the acquisition of these skills.

The importance of cochlear implantation for children with profound hearing loss is exemplified by the superior communication abilities demonstrated by our youngest implant recipient, SHJ. His

communication abilities were superior to the majority of his age-matched peers in this study. All of these children are approaching the age at which formal education begins; their communication abilities will impact greatly on their academic placement and success.

In summary, the present study demonstrates that early implantation for children with profound deafness promotes the acquisition of speaking and listening skills during critical periods of development. Children implanted prior to age three years showed significant advantages in their ability to encode, process and produce speech compared to age-matched peers who were implanted at later ages. It should be noted that the children in this study had used their cochlear implant for differing lengths of time; it is not clear yet whether their skills will ultimately plateau at the same level. Nonetheless, the increased early auditory experience provided by implanting children as young as possible should have important consequences for the development of reading and other academic skills. We plan to continue following these children over time to evaluate the ultimate benefits of early implantation.

References

Cowan RSC, DelDot J, Barker EJ, Sarant JZ, Pegg P, Dettman S, Galvin KL, Rance G, Hollow R, Dowell, RC, Pyman B, Gibson WPR, Clark GM. (1997). Speech perception results for children with implants with different levels of preoperative residual hearing. *Am J Otol*, 18(suppl):S125-S126.

Fryauf-Bertschy H, Tyler RS, Kelsay DMR, Gantz BJ, Woodworth GG. (1997). Cochlear implant use by prelingually deafened children: The influences of age at implant and length of device use. *J Speech Lang Hear Res*, 40:183-199.

Geers A, Toby E. (1992). Effects of cochlear implants and tactile aids on the development of speech production skills in children with profound hearing impairment. *Volta Rev*, 94:135-163.

Hasenstab M, Tobey E. (1991). Language development in children receiving Nucleus multichannel cochlear implants. *Ear Hear*, 12(suppl):S55-S65.

Meyer TA, Svirsky MA, Kirk KI, Miyamoto RT. (1998). Improvements in speech perception in prelingually-deafened children: Effects of device, communication mode, and chronological age. *J Speech Lang Hear Res*,41:846-858.

Miyamoto R, Kirk K, Robbins AM, Todd SL, Riley AI, Pisoni DB. (1997). Speech perception and speech intelligibility in children with multichannel cochlear implants. *Acta Otolaryngol*, 117:198-203.

Miyamoto RT, Osberger MJ, Robbins AM Myres WA, Kessler K. (1993). Prelingually deafened children's performance with the Nucleus multichannel cochlear implant. *Am J Otol*, 14:437-445.

Miyamoto R, Osberger MJ, Robbins AM, et al. (1991). Comparison of speech perception abilities in deaf children with hearing aids or cochlear implants. *Otol—Head & Neck Surg*, 104:42-46.

Miyamoto RT, Svirsky M, Kirk KI, Robbins, AM, Todd, S, Riley, AI. (1997). Speech intelligibility of children with multichannel cochlear implants. *Ann Otorhinolaryngol*, 106:35-36.

Miyamoto RT, Svirsky M, Robbins AM. (1997). Enhancement of expressive language in prelingually deafened children with cochlear implants. *Acta Otolaryngol*, 154-157.

Moog JS, Kozak VJ, Geers AE. (1983). *Grammatical analysis of elicited language—Pre-sentence level*. St. Louis, MO: Central Institute for the Deaf.

Osberger MJ, Robbins AM, Todd SL, Riley AI. (1994). Speech intelligibility of children with cochlear implants. *Volta Rev*, 9:169-180.

Reynell JK, Huntley M. (1985). *Reynell developmental language scales* (Revised 2nd Edition). Windsor, England: NFER-Nelson Publishing Company Ltd.

Robbins AM. (1994). The Mr. Potato Head Task. Indianapolis, IN: Indiana University School of Medicine.

Robbins AM, Svirsky MA, Kirk KI, Miyamoto RT, Bollard P, Green J. (In press). Enhancement of language performance in children with cochlear implants. IN AH Morgon, E Truy (eds.) *Audiology, speech, language and deafness*. London: Whurr Publishers.

Snik A, Vermuelen A, Brokx J, van den Broek P. (1997). Long term speech perception in children with cochlear implants compared with children with conventional hearing aids. *Am J Otol*, 18(suppl):S129-S130.

Staller SJ, Dowell RC, Beiter AL, Brimacombe, JA. (1991). Perceptual abilities of children with the Nucleus 22-channel cochlear implant. *Ear Hear*, 12(suppl):34S-47S.

Waltzman S, Cohen N, Gomolin R, Green, J, Shapiro, W, Brackett, D, Zara, C. (1997). Perception and production results in children implanted between two and five years of age. *Adv Otorhinolaryngol*, 52:177-180.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Lexical Neighborhoods and Release from Proactive Interference: A First Report[1]

Winston D. Goh[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This research was supported by NIH-NIDCD Research Grant DC00111 to Indiana University. I would like to thank David Pisoni and Miranda Cleary for helpful comments during the preparation of this article.
[2] Also, Department of Social Work & Psychology, National University of Singapore.

Lexical Neighborhoods and Release from Proactive Interference: A First Report

Abstract. The effect of lexical competition on recall performance was studied using the release from proactive interference paradigm. Word triads were equated in terms of local neighborhood density but varied on global neighborhood density, neighborhood frequency, and word frequency to create sets of lexically "easy" and "hard" words. Proactive interference was built up across four trials of either "easy" or "hard" words and "released" on the fifth trial with words from the opposite lexical category. The results failed to support the prediction that greater release from proactive interference would be observed when switching from "hard" to "easy" words, although some evidence for a reversal of the release effect was found when switching from "easy" to "hard" words. Extensions to the current design and future studies using this experimental paradigm are discussed.

Introduction

In the Brown-Peterson paradigm (Brown, 1958; Peterson & Peterson, 1959), participants are presented with a short list of three items, followed by a retention interval in which they engage in some distracting activity – counting backwards by threes, backward naming, stroop color naming etc. – to prevent active rehearsal. Participants are then required to recall the list. Keppel and Underwood (1962) demonstrated that recall performance declined quickly over the first few trials. They attributed this effect to the build-up of proactive interference (PI). PI refers to the forgetting of current material because of interference from previously learned material. Items from earlier trials interfere with the recall of items from the current trial.

Recall performance can be raised again if some attribute of the items in the current trial is changed. Wickens, Born, and Allen (1963) demonstrated this "release" from PI effect by changing the semantic category of the items to be remembered in the critical trial; for example, switching from a consonant triad to a number triad or vice-versa. Participants are "released" from the effects of PI and recall performance is restored to the level of the first trial.

The typical procedure in this paradigm is as follows: in the first three trials, the experimental and control groups receive identical triads sharing a similar attribute, or drawn from a single category. These initial trials comprise the PI build-up phase. Recall performance declines as a function of increasing interference from earlier trials. On trial 4 – the critical or release phase – the experimental group receives a triad that differs from the earlier triads on some attribute, or the items are drawn from a different category. The control group continues to receive a triad that is similar to the trials in the PI build-up phase. Recall performance gets a dramatic boost in the experimental group, while it continues to decline in the control group.

This "release" from PI effect has been shown to be very robust using a wide variety of attribute changes. Wickens (1970) has argued that the paradigm is a powerful tool for exploring the extent to which various dimensions are encoded and stored in memory. The amount of "release" obtained with different dimensional changes is assumed to reflect the salience of that dimension in memory. Generally, semantic attributes such as category membership produce a greater amount of release from PI than physical attributes such as word length, phonological similarity, and visual configurations of the presentation (Wickens, 1970).

Although the Brown-Peterson technique is a short-term memory (STM) task, the nature of the release from PI effect is generally regarded as a secondary memory phenomenon (Craik & Birtwistle, 1971). The finding that semantic dimensions are more effective in producing release effects is consistent with the view that semantic dimensions are the domain of long-term memory (LTM). Phonological similarity and word length affect performance in STM span tasks (Baddeley, 1966; Baddeley, Thompson, & Buchanan, 1975), but both of these dimensions have little effect in producing release from PI (Wickens, 1970). This suggests that the locus of PI release is not in the STM processes involved in immediate memory span tasks but is in some other memory process that is dependent on long-term knowledge.

The Locus of the "Release"

Three different explanations have been proposed to account for the release from PI effect. First, the attentional or encoding hypothesis proposes that participants are perceptually alerted by the change in material, and consequently, the items are better registered in memory (Wickens, 1970).

Second, the storage hypothesis posits that PI reflects the spontaneous interaction during storage between traces of current items and those of similar items stored from preceding trials. "Release" items are less vulnerable to inter-trial interference (Posner, 1967). This view of PI release assumes the library metaphor of memory, in which items are "automatically" stored in their proper "shelves", with similar items being stored "closer" together. A change in material will therefore be less vulnerable to interference because the items will no longer be stored in the vicinity of the earlier items.

Third, the retrieval hypothesis proposes that the build-up of PI reflects the declining effectiveness of a participant-generated retrieval cue, which is common to the past few trials. A change in stimulus materials supplies a novel, and thus, more effective retrieval cue (Wickens, 1970).

The debate about PI release in the 1970s was largely between the encoding and retrieval hypotheses. Although one can imagine that some aspect of both explanations are operating in reality, these alternative hypotheses were often pitted against each other as mutually exclusive explanations. The resolution has largely been in favour of the retrieval cue hypothesis. I will describe one study that has often been cited in this regard.

Gardiner, Craik, and Birtwistle (1972) attempted to elucidate the locus of the PI release effect by making the category shift less obvious. Instead of shifting from one category to another on the critical trial, they made the change subtler by shifting from one subset of a larger category to another subset within that same category. For example, the items shifted from *wild* flowers to *garden* flowers.

Participants were randomly placed in one of three conditions. In the "cue at presentation" (CP) condition, a subset cue, "garden", was provided *before* the critical triad of words was presented. This manipulation should induce participants to use the subcategory cue during encoding of the triad, and presumably during recall as well. In the "cue at recall" (CR) condition, the subset cue was only provided at the *recall* phase. This should induce participants to use the subcategory cue during recall, but not at encoding. In the control condition, participants did not receive any subset cue at all. Table 1 summarizes the conditions in the study:

Gardiner et al. (1972) found that participants in *both* the CP and CR conditions showed a significant release from PI effect. No release was obtained in the control condition. At first glance, the results seem to support both the encoding and retrieval hypotheses. However, the authors argued for one alternative – the retrieval hypothesis. If the encoding hypothesis is correct, performance in the CR

condition should be no better than the control condition. This is because there is no processing difference between the CR and control conditions until the recall phase. In both conditions, no attention was drawn to the fact that there was a subcategory shift in the last trial until the recall phase, so the participants in the CR condition should not have encoded the "release" trial any differently from the participants in the control condition. Since the cue was only provided during the recall phase in the CR condition, the facilitation in performance suggests that the cue was useful in the *retrieval* process and not the *encoding* process. In contrast, for the CP condition, attention to the cue was induced prior to encoding the "release" triad, so participants could encode the "release" triad differently. The encoding hypothesis can explain the release in the CP condition, but not the release in the CR condition. According to the encoding hypothesis, there should not have been any release in the CR condition, but the findings showed that there was a release effect.

Condition	PI Build-up Phase		Release Phase	Recall
Cue at presentation (CP)	wild flowers	cue	garden flowers	???
Cue at recall (CR)	wild flowers		garden flowers	cue ???
Control	wild flowers		garden flowers	???

Table 1. Conditions in the Gardiner et al. (1972) study.

This finding is also incompatible with the storage explanation because it rules out any "automatic" encoding of items in memory – otherwise any change in the material should lead to storage in a different location and thus better recall. The control group clearly showed that if the cue is not made salient, participants do not "automatically" use it. Because neither the encoding or storage hypotheses can fully explain the data, we are left, by elimination, with the retrieval hypothesis.

These results have been replicated in a more recent study (Wixted & Rohrer, 1993). Several investigations using other paradigms have also found support for a retrieval cue locus for the release from PI effect (e.g., Watkins & Watkins, 1975; Wickens, Moody, & Dow, 1981). The prevailing view is that subjects are unable to restrict their search space to just the current list, if the retrieval cue is also relevant to items from previous lists. Watkins and Watkins (1975) argued that the build-up of PI can be viewed as cue "overload", which describes the declining efficiency of a functional retrieval cue as the number of items it subsumes increases. This is consistent with other memory phenomena such as the list length effect (the decrement in performance as the length of the list increases) in free recall (see Raaijmakers & Shiffrin, 1992). If the context of the experiment or the presented list acts as a functional retrieval cue, then the cue's effectiveness for retrieval declines as list membership grows.

The use of a cue (completely new material, a subcategory, or other discriminating cue) that allows an effective restriction of the search space should result in the "release" from the interference of previous items.

Lexical Neighborhoods and Memory

Since the mid 1980s, there has only been a handful of research using the release from PI paradigm (e.g., Wixted & Rohrer, 1993). Given the robust nature of the effects that have been elicited by this paradigm, it remains a potentially useful task that can be utilized to explore the organization of lexical entries in LTM.

One aspect of lexical organization that has been intensively investigated in the spoken word recognition literature is the concept of lexical neighborhoods as defined by their phonological properties (Landauer & Streeter, 1973; Treisman, 1978; Luce, 1986; Luce & Pisoni, 1998). The neighborhood density of a word can be computationally derived by considering the number of words that can be obtained by a single phoneme substitution, addition, or deletion. The neighborhood frequency of a word can be derived from the average frequency of the word's neighbors.

Using these properties, words can be classified into those that would be theoretically "easier" or "harder" to recognize. "Easy" words are high frequency words with low neighborhood density and low neighborhood frequency. Thus, they have less competition from similar sounding words during the recognition process and so should be more easily recognized relative to "hard" words, which are low frequency words that come from high density and high frequency neighborhoods.

Previous studies have demonstrated that these lexical neighbourhood properties influence word recognition as shown in the accuracy of perceptual identification, latencies in naming and lexical decision tasks, and priming effects (Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998; Luce, Pisoni, & Goldinger, 1990). Specifically, "easy" words are recognised more accurately and more quickly than "hard" words. These results have been incorporated in the Neighborhood Activation Model of word recognition (Luce & Pisoni, 1998), and the connectionist version, PARSYN (Luce, Goldinger, Auer, & Vitevitch, in press). The key assumption of both models is that words are recognised "relationally" through a process of lexical discrimination. The organization of the mental lexicon, specifically the phonological properties of these words' LTM traces, plays a critical role in the recognition process (see Luce & Pisoni, 1998).

While the effects of lexical neighborhoods on word recognition tasks have been well established, several recent studies have begun to explore the extent to which these properties affect other tasks such as serial recall. Goldinger, Pisoni, and Logan (1991) showed that serial recall of ten-word lists made up of lexically "easy" words was superior to lists made up of lexically "hard" words. Goh and Pisoni (1998) further showed that the difference between "easy" word-spans and "hard" word-spans were not related to participants' short-term memory capacity, as measured by digit-span performance, and that the serial recall differences between "easy" and "hard" words can be eliminated when a small set of words are repeatedly used over and over compared to using novel words for each new list. These results suggest that when considering the effects of lexical neighbourhoods on memory tasks such as serial recall, the locus of these effects is likely to be in the LTM component of the tasks, and not in the STM processes. STM capacity does not reliably predict the performance differences between "easy" and "hard" word lists (see Goh & Pisoni, 1998).

As previously described, the release from PI effect appears to be localized in LTM retrieval processes. The release from PI paradigm may thus provide another way to test the proposition that lexical neighbourhood effects stem from LTM processes. The use of different tasks to test hypotheses will provide valuable converging evidence or alternatively, disconfirmatory evidence for theoretical claims. The present pilot study was an attempt to explore the influence of lexical neighbourhoods on the release from PI effect.

The basic question of interest is whether the "release" effect will be of the same magnitude when the critical triad switches from "easy" words to "hard" words, compared to the reverse condition. Lexically "hard" words should result in a greater build up of PI because their dense neighbourhoods will result in interference from many similar sounding words in the lexicon. Any retrieval cues that participants may utilize to retrieve the target words will become increasingly less useful for discriminating between the context of the current trial and the interference from the neighbourhood

activations of "hard" words and the context of previous trials. Hence, switching from "hard" to "easy" words may result in a significant "release" effect because the "easy" words will have to contend with fewer competing neighbours, and the retrieval cues may more easily discriminate the context of the current trial from the previous trials. In the reverse situation (switching from "easy" to "hard" words), the PI build up is likely to be less severe because the sparse neighborhoods may produce less interference. Switching from a context of less interference to one of more interference may not produce a "release" effect at all, and in fact, may even produce a greater decline in performance.

One important issue that needs to be discussed is the distinction between *local* and *global* neighborhoods. Global neighborhood properties refer to what has been traditionally computed in selecting the set of experimental stimuli – density and neighbourhood frequency indices derived from computations using the entire mental lexicon, or an estimate thereof. In contrast, local neighborhood properties refer to indices computed only from the set of experimental stimuli. For example, the local neighborhood density of a particular word will be derived from the number of other words in the experimental corpus that can be obtained using the phoneme substitution rule, not from the *total* number of words that can be obtained. Ideally, if one is investigating the influence of LTM structure and not STM, one should vary the global neighbourhood properties of the words and control for the local neighborhood properties. In other words, the "easy-hard" distinction should be based on *global* properties, while *local* properties are equated. Whatever effects are found can then be solely attributed to the global properties in LTM organization. This control is difficult to implement in studies that use a large number of words that are not repeated from trial to trial (e.g., Goh & Pisoni, 1998; Goldinger et al., 1991). However, in the release from PI paradigm, only a handful of words are required for the whole experiment, and hence it is easier to establish a local neighborhood control for the neighborhood density index. To establish local controls for the frequency and neighborhood frequency indices will require experimentally induced frequencies, which is not within the scope of the current paradigm.

The specific hypothesis of this investigation is that a greater release from PI will be observed when the critical triad switches from "hard" words to "easy" words than the reverse. There may not even be a reliable "release" effect in the "easy-hard" switch relative to performance on the first trial. In fact, the opposite effect may be observed, where performance declines even more sharply relative to the decline in performance in the trials during the PI build up phase.

Method

Participants

Forty-three Indiana University psychology undergraduates participated for course credit. All were native English speakers with no self-reported speech or hearing disorder at the time of testing.

Materials

Tokens of 30 spoken words (15 "easy" and 15 "hard") were drawn from a pre-recorded digital database (see Torretta, 1995, for a detailed description). All of the test items were monosyllabic consonant-vowel-consonant words that were rated as highly familiar (> 6.7) on the 7-point Hoosier Mental Lexicon scale (Nusbaum, Pisoni, & Davis, 1984). The tokens recorded by the most intelligible male talker (M9, who had a mean intelligibility score of > 95%) at a "medium" speech rate were used (see Torretta, 1995, for details). The mean statistical properties of the tokens are listed in the appendix.

Apparatus

Gateway 2000 Pentium 133 MHz IBM compatible computers equipped with SoundBlaster AWE32 sound cards were used to control the experiment. The stimuli were presented to participants via Beyer Dynamic DT100 headphones at approximately 75 dB SPL.

Design

The experiment employed a 2 x 5 mixed factorial design. The Release independent variable – "easy" words build-up, "hard" words release (EH) vs. "hard" words build-up, "easy" words release (HE) – was run as a between-subjects variable. The Trial independent variable (4 build-up trials and a single release on the 5th trial) was run as a within-subjects variable.

Procedure

Participants were tested individually or in small groups of five or fewer. There were three phases in the experiment, which took approximately 15 minutes to complete. In the first phase, participants were given practice on the filler task, which was simply to mentally keep a running sum of a sequence of digits presented on the computer monitor. Each digit appeared at the center of the monitor for a duration of 1750 msecs, after which it was cleared from the screen and the next digit appeared at the same location. The digits were randomly picked from the numbers 5 through 9, with the proviso that the same digit did not appear consecutively. A total of ten digits appeared before participants were prompted to enter the sum of the digits. Feedback was given on the accuracy of the responses. After five filler trials, the computer determined if participants achieved a criterion of four correct trials. If not, another five trials were given to the participants and so on. After the third set of five filler trials, participants were allowed to move on to the next stage of the experiment, regardless of their performance on the practice filler trials. Performance accuracy was recorded by the computer.

In the second phase, participants heard a random sequence of three digits spoken by a male talker over the headphones at an inter-stimulus interval of 250 msecs. This male talker was not the same as the talker who produced the word tokens that were used in the final phase. After the presentation of the last digit, the same filler task that participants practiced in the first phase appeared. After a single trial of the filler task, participants were prompted to recall the three digits that they heard over the headphones. The total retention interval was 20 seconds. Participants had 15 seconds to type their responses using the keyboard before the next trial began. The computer recorded the responses for both the filler and recall tasks. Participants completed two more trials consisting of a digit triad presented auditorily followed by the filler task before moving on to the next stage of the experiment. The purpose of this second phase was to give participants some practice on doing the recall and filler tasks together.

In the final phase of the experiment, the procedure was exactly the same as the second phase, except that a triad of spoken word tokens was randomly selected and presented auditorily instead of three digit tokens. Words selected for the current trial were not used again in any subsequent trials. The first four trials constituted the PI build-up trials in which the triads were comprised of only "easy" words in the EH condition, or only "hard" words in the HE condition. The fifth trial was the release trial, where "hard" words were presented in the EH condition, and "easy" words were presented in the HE condition.

Results

Analyses by Subjects

I had initially intended to analyse only the data from participants who were able to meet the criterion of four out of five correct trials in the filler task practice phase. This was deemed necessary to ensure that participants were actually doing the filler task and not rehearsing the auditory triad during the retention interval. However, the filler task proved to be rather difficult and only six participants met this strict criterion. Therefore, I will report two sets of analyses – one using all participants, regardless of filler task performance, and one using a more relaxed criterion, in which those participants who gave an accurate response to three out of the five filler trials in the final phase of the experiment were included in the analyses. The latter criterion yielded 20 admissible participants.

Following Kincaid and Wickens (1970), a point was given for each word of the triad that was correctly recalled, regardless of sequence position. An extra point was given if all three words were recalled in the correct order. So the maximum score for any single trial is four – three points for each word, and one extra point if the triad was recalled in the correct order.

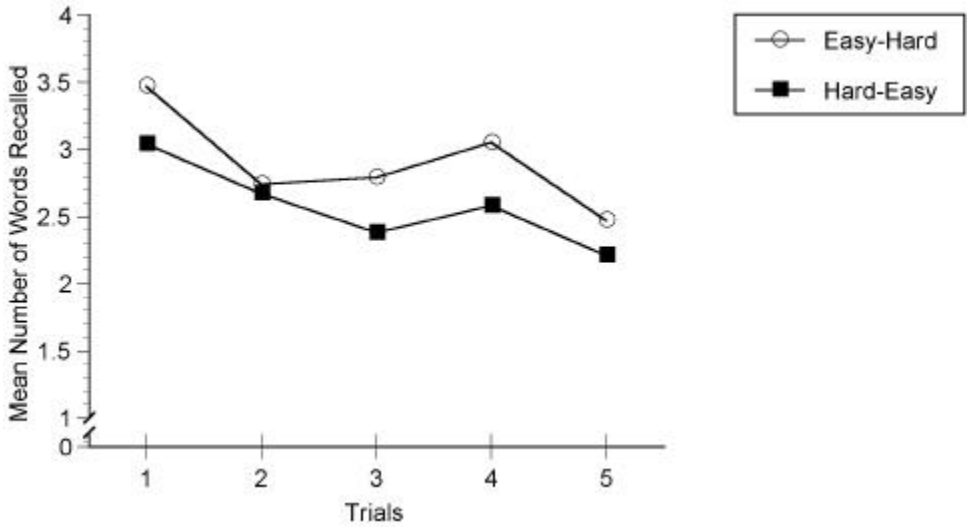


Figure 1. Mean recall for all participants.

Figure 1 shows the pattern of results when the data from all participants were included, regardless of whether they met the performance criterion in the filler task. A 2 x 5 mixed factorial analysis of variance (ANOVA) revealed a reliable main effect of Trials, $F(4, 164) = 3.27$, $MSe = 1.49$, $p < .05$, indicating that performance decreased as the number of trials increased. Figure 1 shows the presence of an increasing build-up of PI from the first to the fifth trial. All other effects were not significant, which indicates that when the performance of all participants was considered, the hypothesised interaction between the type of word ("easy" vs. "hard") and the effect on the release from PI (EH vs. HE at trial 5) was not supported. The mean performance on each trial is given in Table 2.

Release Condition	Trial				
	PI Build-up Phase				Release
	1	2	3	4	5
Easy-Hard					
M	3.47	2.74	2.79	3.05	2.47
SD	0.90	1.33	1.08	1.18	1.39
Hard-Easy					
M	3.04	2.67	2.38	2.58	2.21
SD	1.08	1.31	1.41	1.53	1.28

Table 2. Mean recall performance across trials (all participants).

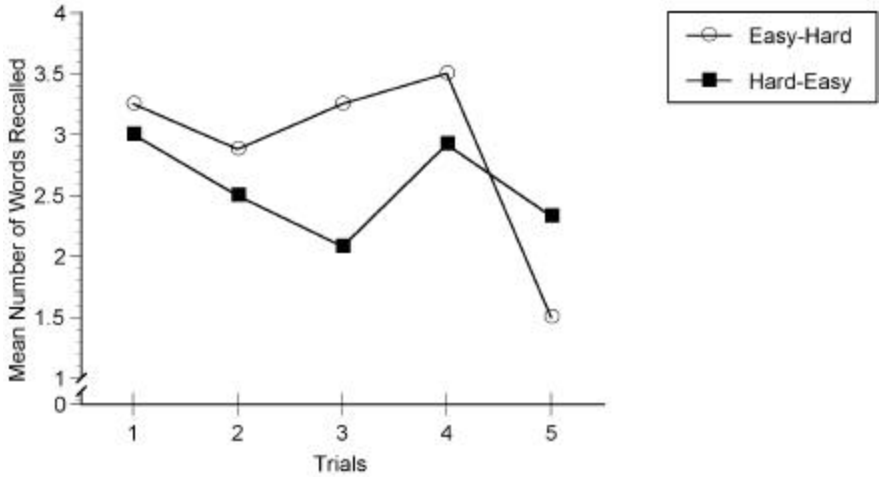


Figure 2. Mean recall for participants who were accurate on at least 3 filler trials.

Figure 2 shows the pattern of results when only the data from the participants who met the relaxed criterion of 3 out of 5 filler trials correct were examined. A 2 x 5 mixed factorial ANOVA again showed a reliable main effect of Trials, $F(4, 72) = 4.51$, $MSe = 1.12$, $p < .01$, indicating a general build-up of PI as performance decreased across trials. This time, however, there was a marginal Trials x Release interaction, $F(4, 72) = 2.55$, $MSe = 1.12$, $p < .07$. Tests of simple effects showed that the source of the interaction was the significant simple effect of Trials in the EH condition, $F(4, 72) = 4.57$, $MSe = 1.12$, $p < .01$; all other simple effects were not significant. Post-hoc pairwise comparisons of the recall performance at each trial in the EH condition showed that the "release" trial (trial 5) was significantly lower than all the trials in the PI build-up phase (trials 1 to 4). This result suggests that there was a reversal of the release effect in the EH condition – switching to "hard" words caused even more interference for recall than switching to "easy" words. This supports one of the earlier predictions. However, the data do not support the prediction that switching from "hard" to "easy" words should result in a reliable release effect in the HE condition. The mean performance at each trial is shown in Table 3.

Release Condition	Trial				
	PI Build-up Phase				Release
	1	2	3	4	5
Easy-Hard					
M	3.25	2.88	3.25	3.50	1.50
SD	1.04	1.25	1.04	0.93	1.20
Hard-Easy					
M	3.00	2.50	2.08	2.92	2.33
SD	1.04	1.17	1.31	1.38	1.37

Table 3. Mean recall performance across trials (at least 3 fillers correct).

Analyses by Trials

The data were also subjected to a trials analysis. In this analysis, for each of the five trials across every participant, only those trials where a correct answer was given for filler task were admitted. Although this meant that there would probably be unequal numbers of participants for each trial (to be conservative, Trials would be considered a between-subjects factor), this method of analysis may provide additional converging evidence for the pattern of results obtained in the analyses by subjects.

Figure 3 shows the pattern of results observed after using the above criterion. A 2 x 5 between-subjects ANOVA showed a reliable main effect of Trials, $F(4, 84) = 3.11$, $MSe = 1.54$, $p < .05$, again indicating a general build-up of PI across trials. All other effects were not significant. Although there was no significant interaction between Trials and Release, the observed pattern is similar to the pattern displayed in Figure 2. Post-hoc analyses using Tukey's HSD procedure for each level of the Release independent variable showed that, in the EH condition, recall at Trial 5 was significantly lower than recall at Trials 1 and 4. None of the other pairwise comparisons were significant. In the HE condition, all pairwise comparisons were not significant. Again, the results of this analysis suggest that there was some indication of a reversal of the release effect in the EH condition, but there was no evidence of any release effect in the HE condition.

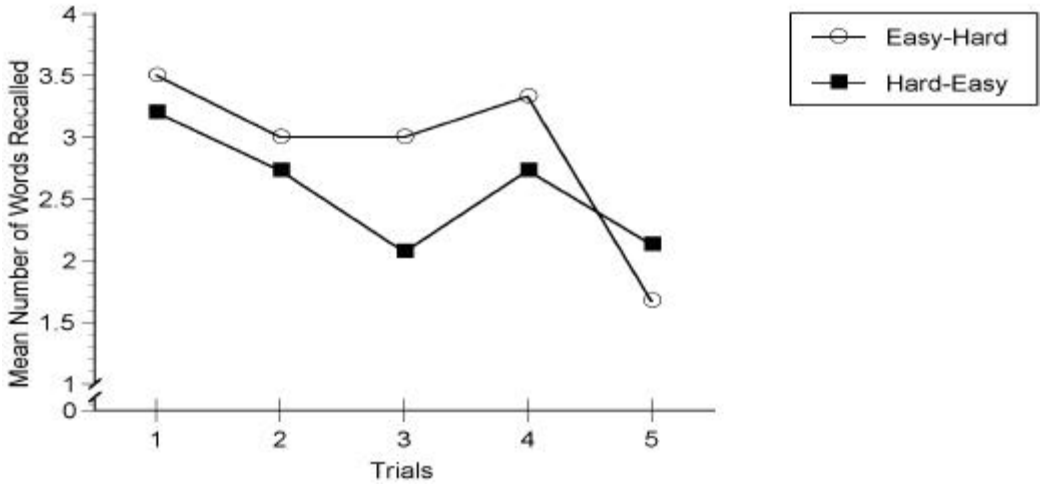


Figure 3. Mean recall for trials with accurate filler answers.

Figure 4 shows the pattern of results obtained when a less stringent criteria was used to do the trials analyses. This time, trials in which the participants' filler task response did not deviate by more than 5 from the correct sum were included. The 2 x 5 between-subjects ANOVA revealed a reliable main effect of Trials, $F(4, 173) = 3.62$, $MSe = 1.54$, $p < .01$, replicating previous analyses, and a reliable main effect of Release, $F(1, 173) = 6.13$, $MSe = 1.54$, $p < .05$, indicating that there was generally better recall in the EH condition relative to the HE condition. There was no significant interaction between Trials and Release. Again, however, Tukey's HSD analyses showed that there was a significant difference between recall at Trials 1 and 5, but only in the EH condition. As before, the "release" with the "hard" words was actually more detrimental to performance, but again, there was no evidence of any release in the predicted direction in the HE condition.

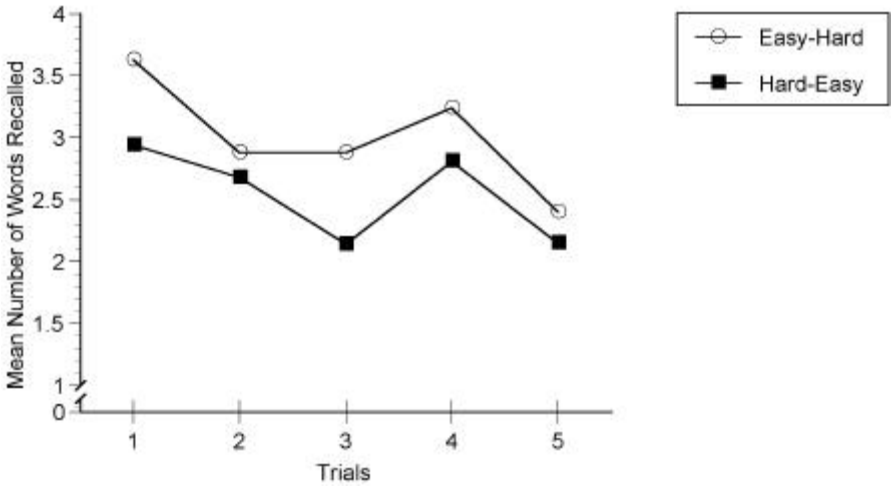


Figure 4. Mean recall for trials with filler answers that did not deviate by more than 5 from the correct answer.

Discussion

The observed results did not support the main hypothesis that there would be a greater release from PI effect when participants were switched from "easy" word triads to "hard" word triads than the reverse. Some evidence was obtained that supported the predicted detrimental effect on performance in the release trial of the "hard-easy" switch.

It should be pointed out that several important control conditions were not run in the present pilot study, namely, the conditions where all five trials remained "easy" or "hard" word triads. The extent of a "release" effect would normally be determined by comparing the recall performance on the "release" trial with the recall performance on the fifth trial on these control conditions. However, since the critical prediction of the original hypothesis was a *relative difference* between recall on the first and fifth trials across the two different release conditions, the lack of the "no switch" controls does not detract from the interpretation of the major results.

The overall lack of a release from PI effect in this study suggests that the lexical neighborhood properties of the word triads are not affecting the retrieval processes used in recalling the words in this paradigm. However, there are a number of issues that warrant further investigation. First, there does seem to be some detrimental effect when switching from "easy" to "hard" triads, so the greater lexical competition in "hard" neighborhoods may in fact be affecting the efficacy of retrieval cues. Second, it is

possible that a longer retention interval may be required to reveal a release effect in the "hard-easy" condition – the activation levels of the "hard" neighborhoods in the PI build-up phase may still be at a level that is high enough to cause interference in the "easy" release trial. Increasing the retention interval may serve to reduce the activation levels of the "hard" neighborhoods sufficiently for the "easy" release trial to show the predicted effect. Neighborhood activity becomes attenuated when the inter-stimulus-interval between trials is increased in a priming paradigm (Goldinger et al., 1989). Finally, Wickens (1970) noted that the release from PI phenomenon is most robust when the dimension that is switched is semantic in nature. Surface level features such as orthography and phonology show much weaker release from PI effects. It may thus be inherently more difficult to elicit an effect with phonological neighborhood properties, especially when it is the global and not local properties that were manipulated – phonological neighborhoods may constitute a dimension that is not "salient" enough to produce a robust effect in this procedure.

In summary, while the present results did not provide conclusive evidence that phonological neighborhood properties affect the efficacy of retrieval cues that prior studies have shown to be the locus of the release from PI phenomenon, some evidence was observed for an asymmetry across the two conditions tested. There seemed to a reversal of the release from PI effect in the EH condition, but the typical release from PI effect was not obtained in the HE condition. Extensions of the current design using some of the control conditions that were not implemented in the present study, or increasing the retention interval, may provide more reliable results.

References

Baddeley, A.D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, *18*, 363-365.

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.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10, 12-21.

Craik, F.I.M., & Birtwistle, J. (1971). Proactive inhibition in free recall. *Journal of Experimental Psychology*, *91*, 120-123.

Gardiner, J.M., Craik, F.I.M., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior*, *11*, 778-783.

Goh, W.D., & Pisoni, D.B. (1998). Effects of lexical neighborhoods on immediate memory span for spoken words: A first report. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 195-213). Bloomington, IN: Speech Research Laboratory, Indiana University.

Goldinger, S.D., Luce, P.A., & Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, *28*, 501-518.

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.

Keppel, G., & Underwood, B.J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior, 1*, 153-161.

Kincaid, J.P., & Wickens, D.D. (1970). Temporal gradient of release from proactive inhibition. *Journal of Experimental Psychology, 86*, 313-316.

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. (1986). *Neighborhoods of words in the mental lexicon* (Research on Speech Perception Technical Report No. 6). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A., Goldinger, S.D., Auer, E.T., & Vitevitch, M.S. (in press). Phonetic priming, neighborhood activation and PARSYN. *Perception & Psychophysics*.

Luce, P.A., & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1-36.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods for spoken words. In G.T.M. Altmann, (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 142-147). Cambridge: MIT Press.

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, Indiana University.

Peterson, L.R., & Peterson, M.J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*, 193-198.

Posner, M.I. (1967). Short term memory systems in human information processing. In A.F. Sanders (Ed.), *Attention and performance* (pp. 267-284). Amsterdam: North Holland.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1992). Models for recall and recognition. *Annual Review of Psychology, 43*, 205-234.

Torretta, G.M. (1995). The “easy-hard” word multi-talker speech database: An initial report. In *Research on Spoken Language Processing Progress Report No. 20* (pp. 321-334). Bloomington, IN: Speech Research Laboratory, Indiana University.

Treisman, M. (1978). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior, 17*, 37-59.

Watkins, O.C., & Watkins, M.J. (1975). Build up of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory, 104*, 442-452.

Wickens, D.D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review, 77*, 1-15.

Wickens, D.D., Born, D.G., & Allen, C.K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 2*, 440-445.

Wickens, D.D., Moody, M.J., & Dow, R. (1981). The nature and timing of the retrieval process and of interference effects. *Journal of Experimental Psychology: General, 110*, 1-20.

Wixted, J.T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1024-1039.

Appendix

"Easy" Words

curve	dirt	dog	fig	join
judge	league	leg	lose	noise
page	roof	soil	theme	wash

"Hard" Words

cheer	chore	comb	hid	hoot
hurl	lad	mall	pawn	pup
sill	soak	tan	wail	white

Note. Except for cheer/chore, league/leg and sill/soil, which are neighbors of each other, the local neighborhood density for the other 24 words is 0 – none of these words are neighbors of each other. The mean global neighborhood density, mean log neighborhood frequency, and log frequency for the "easy" words are 11.87 (4.63), 1.93 (0.25), and 2.75 (0.10) respectively; and for the "hard words, they are 25.53 (5.94), 2.22 (0.25), and 1.80 (0.54) respectively. Numbers in parentheses are the standard deviations. Frequency counts are based on the Kucera and Francis (1967) counts.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Perception and Production of Intonational Contrasts in an Adult Cochlear Implant User[1]

Rebecca Herman and Cynthia Clopper

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH-NIDCD T32 training grant DC00012. We would like to thank our speakers for their participation in this experiment. We are also grateful to David Pisoni for help with this work. We would also like to thank Luis Hernandez and Carlos Colon for technical assistance in every phase of this project. Thanks also to Jimmy Harnsberger for helpful editorial comments on an earlier draft of this paper.

Perception and Production of Intonational Contrasts in an Adult Cochlear Implant User

Abstract. This paper presents a case study on the linguistic use of fundamental frequency (F0), amplitude, and duration by a post-lingually deafened adult speaker with a cochlear implant. Although multi-channel cochlear implants do not have good resolution of F0, it has been reported anecdotally that the intonation of cochlear implant users sounds "fine" in conversations. This observation implies that post-lingually deafened adults may be relying on very robust phonological representations of intonational contrasts, even in the face of degraded phonetic input. To address this issue, we developed an imitation task for eliciting intonational contrasts, which was piloted with a post-lingually deafened adult who had received a cochlear implant and with a matched, normal-hearing subject. Recordings from the imitation task were used as stimuli in a set of perception experiments using normal-hearing, college-aged listeners. The results of these perception experiments showed, as expected, no statistical difference between listeners' judgments of utterances as produced by the cochlear implant user and by the normal hearing speaker. These findings provide a good baseline study of the perception and production of intonational contrasts by a skilled, post-lingually deafened cochlear implant user.

Introduction

Anecdotal reports on users of cochlear implants indicate that such speakers sound "fine" in conversations, and that their intonational use is within normal limits. These informal reports are intriguing because cochlear implants with multiple channels do not have good resolution in the frequency range of the fundamental frequency of voices. Rosen, Walliker, Brimacombe, and Edgerton (1989) mention that multi-channel implant systems do not necessarily signal some of the basic information on F0 very well.[2] Cochlear implant users are apparently receiving a degraded phonetic input, without good resolution in the F0 range. However, even if this is true, they seem to be functioning normally in conversations, in which it is crucial to signal pragmatic meanings using intonational conventions. If it can be shown experimentally that cochlear implant users are in fact performing similarly to normal-hearing speakers in producing and perceiving a range of intonational contrasts, this would be very useful to know and would suggest several implications.

The first implication of normal performance despite degraded phonetic input is that the phonetic signal is so robust and so full of co-varying acoustic cues that the degradation of one aspect of the signal is not enough to prevent the phonological parsing of the intended message by the listener. For example, intonational events are cued not just by movements of F0, but also by factors such as the harmonic structure of vowels (see Sluijter & van Heuven, 1996, for evidence of this in stressed syllables in Dutch) amplitude variations over time, durational differences (see Beckman, 1986), and even perhaps differences in "strength" of consonantal articulations. (See, e.g., Fougeron & Keating, 1997, on the different "strength" of articulation of segments at different locations in the prosodic hierarchy.) In order to explore this implication, it would be necessary to first show that cochlear implant users are in fact performing comparably to normal hearing speakers in tests of intonation perception and production. We would then have to show that other cues to intonational contrasts have good resolution with a multi-channel cochlear implant.

[2] Note that this is different from the earlier single-channel cochlear implants, whose primary function was to convey voicing information, including F0.

Another implication of normal performance in conversations despite insufficient input is that, in the case of post-lingually deafened individuals in particular, the speaker's phonological representations of intonational contrasts are sufficiently robust to allow for phonological parsing of the intended message. In order to explore this issue, it would again be necessary to first show normal performance by cochlear implant users in the perception and production of intonational contrasts. Then, it would be necessary to find speakers who might be conjectured to have less robust phonological representations[3], either because they were pre-lingually deafened or because they had longer periods of profound hearing loss before implantation. In such cases, we might expect that these speakers would have less robust phonological representations, and thus might be more hampered in production and perception by a degraded phonetic input than speakers with more robust phonological representations.

There have been many studies of F0 perception and production by cochlear implant users in the past. These past studies have in large part focused on two opposing issues. On the one hand, some studies investigating audiological aspects of cochlear implant use have tended to examine micro-level psychophysical aspects of the perception of pitch differences. For example, past studies have examined intensity difference limens and frequency difference limens. (See, e.g., Brimacombe, Edgerton, Doyle, Erratt, & Danhauer, 1984. Other studies on the psychoacoustics of cochlear implant use, as cited by Richardson, Busby, Blamey, & Clark, 1998, include Lim & Tong, 1989; Pfingst, Holloway, Poopat, Subramanya, Warren, & Zwolan, 1994; Shannon, 1983; Tong, Clark, Blamey, Busby, & Dowell, 1982; Tong, Blamey, Dowell, & Clark, 1983; and Townshend, Cotter, van Compernolle, & White, 1987.) Such studies, while crucial in understanding cochlear implant users' hearing, have not focused on the linguistic uses of pitch and intensity.

On the other hand, a number of studies focusing on more global aspects of cochlear implant use have tended to examine macro-level, gross overall features of voice such as pitch range, intensity, and overall voice quality. Such studies tend to find somewhat abnormal voice parameters during deafness, with an approximation to more normal voice features after implantation. (See, e.g., Kirk & Edgerton, 1983; Lane, Wozniak, Matthies, Svirsky, Perkell, O'Conell, & Manzella, 1997; Leder & Spitzer, 1990; Monini, Banci, Barbara, Argiro, & Filipo, 1997; and Perkell, Lane, Svirsky, & Webster, 1992.) Such studies also do not examine linguistic use of F0 per se, but focus on concerns about pathological features of vocal use in profoundly deaf speakers.

However, a few studies have in fact examined the use of pitch, amplitude, and duration to convey linguistic information. Such experiments have tended to examine either the production and perception of contrastive stress by cochlear implant users or else the production and perception of pitch rises vs. pitch falls, such as those found at the ends of question vs. statement intonation in English. (See, e.g., Bolinger, 1986 on rises vs. falls utterance-finally in English.) Such studies have tended to use the sections of standardized perception test batteries that address these factors.

For example, Jones, McDermott, Seligman and Milar (1995), in an examination of different strategies for coding F0 in cochlear implants, used the "roving stress test"[4] and question-statement tests

[3] "Less robust phonological representation" might be analogous to Quillet, Wright, and Pisoni's (1998) ideas about "coarse coding" of place contrasts in cochlear implant users, in which cochlear implant users may actually use fewer categories of place of articulation than normal-hearing listeners. This idea of "robustness" of phonological representations needs further exploration.

[4] Note that in the "roving stress test," the patient indicates which word in a three-word phrase such as "two large trees" or "few weak parts" is stressed. In the first version of this test, the stress appeared on any of the three words, but in the second version of the test, because of confounds from final lengthening, the stress appeared only on the

from the Speech Pattern and Contrasts battery (Boothroyd, 1988). They also used a "specific test of the detection of rising and falling intonation," which "consisted of nine repetitions of four distinct tokens of a rising, falling, or steady intonation contour imposed on a sustained /m/" (Jones et al., 1995, p. 365). They found group means between 60% and 70% for both coding strategies tested.[5] Thus, Jones et al. tested/assessed the perception of stress and the perception of rises vs. falls in cochlear implant users.

As another example, Leder, Spitzer, Milner, Flevaris-Phillips, Richardson, and Kirchner (1986) examined the longitudinal "re-acquisition" of contrastive stress in one subject with a single-channel cochlear implant, using as stimuli pairs of words with stress shift from noun to verb. The words used in their study were "contrast," "conflict," "contest," "contract," and "construct." They found that although the subject was unable to produce contrastive stress correctly pre-cochlear implant, his productions of contrastive stress were always correct post-cochlear implant.

Richardson et al. (1998) selected three tests drawn from the Speech Pattern Contrast Battery (Boothroyd, 1984) in order to compare different processing strategies in the cochlear implant. These included the "roving stress test" in which the patient indicates which word in the sentence was stressed (see footnote 4), the "rise-fall" test in which the patient indicates whether the sentence was a statement or a question, and the "pitch and intonation tests" in which the patient indicates whether the sentence was spoken by a male or a female and whether the tone of the voice was "natural" or "monotone." Richardson et al. also used two tests from the Minimal Auditory Capabilities Battery (Owens, Kessler, Telleen, & Schubert, 1982) the "accent" test in which the patient indicates which word in the sentence is stressed and the "question-statement" tests in which the patient indicates whether the sentence was a question or a statement. They found scores ranging between 55% and 91%, varying by test, processing strategy, and patient. Thus, Richardson et al. tested/assessed the perception of location of "stress," rises vs. falls, and more gross overall features of pitch perception.

As another example, Rosen et al. (1989) also used the question-statement subtest of the Minimal Auditory Capabilities Battery (Owens, Kessler, Telleen, & Schubert, 1981). In this test, "the same list of twenty sentences was used but the order in which they appeared as questions or statements varied." (Rosen et al., 1989, p. 95) They found that all subjects were performing significantly better than chance at the task.

In a similar study conducted by Huang, Wang, and Liu (1995), Mandarin-speaking cochlear implant patients were tested using the Mandarin tone perception test, a four-alternative forced choice test using 10 words with identical phonological structure but different tone features. Patients were also tested on the Mandarin version of the monosyllable, trochee, and spondee (MTS) test, but with tones (Wang & Huang, 1988). In this test, the patient heard two repetitions of four monosyllables with either a flat, rising, falling-then-rising, or falling tone; four bisyllabic words with a falling tone on the first syllable and a rising tone on the second syllable; and four bisyllabic words with falling tones on both syllables. They found that "patients who had increased scores on tone perception after implantation also had improvements on other test batteries." (Huang et al., 1995, p. 294) Thus, the tests used by Huang et al. (1995) assessed the perception of lexical tone in Mandarin.

The tests and procedures summarized above provide important insights into the use of dynamic pitch contrasts by cochlear implant users, and their acquisition or re-acquisition of such contrastive pitch changes. However, these studies have been limited by the small set of prosodic phenomena investigated.

---

first or second word. From the descriptions in Boothroyd (1988) it seems like what is being called "stress" should be interpreted according to modern intonational theory as "pitch-accented." (See, e.g., Beckman and Ayers, 1994.)

[5] They do not show comparisons with normal-hearing listeners.

For example, contrastive stress is not a particularly productive aspect of English phonology, being limited to a handful of words. Moreover, the focus of past research on sentence-final rises and falls leaves out many of the linguistic uses of F0 that have been uncovered in research on the intonational phonology of English. A long history of work on English intonation originating in the "British School" can be traced back to at least Palmer (1922).[6] Research on the phonemic system of English intonation has been developed throughout this century, as summarized by Ladd (1996). Despite some disagreement as to the exact inventory of tones in English intonation and the nature of those tones,[7] the consensus on English intonation, drawing on work from acoustic phonetics, phonology, speech technology, and speech perception, seems to be that there is a level of tones mediating between form and function. What is meant by the "form" here is the phonetic factors of F0, amplitude and duration and what is meant by the "function" is the conveying of information structure. Walker, Joshi, and Prince (1998, p. 15), in defining information structure, write that "The function of these choices [the syntactic choices that a speaker makes] is to package the information in the utterance for particular pragmatic and semantic effects." It can be argued that not only the syntactic choices but also the intonational choices made by the speaker serve to organize the information for pragmatic purposes. Thus, in English, F0, amplitude, and duration can be manipulated to convey pragmatic meanings. For example, there has been much research on how "focus" is conveyed by intonational means. (See, e.g., Jackendoff, 1972 and Selkirk, 1984, to name but a few of the many works on this issue.) As another example, there has also been work on how discourse structure is conveyed through intonational means. (See, e.g. Hirschberg & Pierrehumbert, 1986 and Pierrehumbert & Hirschberg, 1990, to name two such works.) Similarly, work on "given" vs. "new" information has also examined how intonational form conveys pragmatic function. (See, e.g., Clark & Haviland, 1977, on "given vs. new.") Thus, a whole range of pragmatic functions can be communicated, via intonational tones, using the phonetic factors of F0, amplitude, and duration.

The intonational model used here draws heavily on the model of English intonational phonology developed by Pierrehumbert (1980) and the subsequent development of an intonational transcription standard developed on the basis of this model. The AE-ToBI[8] transcription standard for mainstream varieties of American English (Beckman & Ayers, 1994) describes movements between relatively low and high F0. There are three types of tones. The first type of tone, pitch accents, are pitch movements associated with a stressed syllable. There are two mono-tonal pitch accents, represented by H* and L*, where the "*" indicates association with a stressed syllable. Pitch accents can also be bitonal, with another tone leading or trailing the associated tone. There are two rising bitonal pitch accents, represented by L+H*, in which the "L+" indicates a leading low tone, and L*+H, in which the "+H" indicates a trailing high tone. Downstep, or compression of the pitch range, also occurs in English intonation, represented in the set of pitch accents by !H*, L+!H*, L*+!H, and H+!H*, where the "!" indicates a downstepped tone. The second type of tone, phrase accents, are pitch movements aligned with the edge of an intermediate phrase. There are two-phase accents in English, H- and L-. It is also possible to have a downstepped phrase accent in English, represented as !H-. The third type of tone, boundary tones, are pitch movements aligned with the edge of an intonational phrase. An intonational phrase contains one or more intermediate phrases. The two boundary tones in English are H% and L%. Each of the tones, according to Pierrehumbert and Hirschberg (1990, p. 308), contributes to discourse interpretation in a compositional manner. Pitch accents "convey information about the status of discourse referents, modifiers, predicates, and relationships

[6] Ladd (1996) explains that early works on English intonation which did treat intonation as composed of a small number of distinct elements, were focused mainly on descriptions useful for either teaching English or developing phonemic theory.

[7] For example, there is disagreement about whether tones can be divided up into separate units for phrase-medial vs. phrase-edge tones or not. See, e.g., Ladd (1996, pp. 44-45) for discussion.

[8] Where AE-ToBI stands for American English - Tones and Break Indices.

specified by accented lexical items," while phrase accents convey information about "the relatedness of intermediate phrases" and boundary tones "convey information about the directionality of interpretation for the current intonational phrase."

Given the high functional load of intonational contrasts in English, it is important in testing cochlear implant users' use of pitch, amplitude, and duration to go beyond the limited use of contrastive stress and rises vs. falls to a fuller range of intonational contrasts. Thus, it is necessary to empirically test the perception and production of a wide range of intonational contrasts by cochlear implant users, and to compare those tests with tests of matched normal hearing speakers, to ensure that any level of functioning below expectation is not due to the experimental task but rather to the use of a cochlear implant.

Methodologies for studying the perception of intonational contrasts have not been as fully developed as those for studying the production of intonational contrasts. Discrimination tasks between intonational contours have sometimes been carried out (Remijsen & van Heuven, 1999; Ladd & Morton, 1997). Identification tasks, on the other hand, are somewhat more difficult to carry out, given the more abstract nature of the function of intonational contrasts, and the difficulties inherent in labeling such contrasts when using naïve listeners as subjects. Th us, while listeners can generally label "question" vs. "statement" with ease, more subtle aspects of intonational function may be difficult for untrained listeners to give labels to. Given these difficulties with intonational contrasts in perceptual experimentation, we developed a new method for exploring speakers' capabilities of perceiving and producing intonational contrasts.

The method developed here for eliciting a range of intonational contrasts was an imitation task, in which the experimenter briefly set up a context for the speaker and then produced an utterance with a specific intonation contour. The speaker was then asked to imitate the utterance exactly as it had been produced. As an initial pilot study, we used one highly-skilled cochlear implant user and one matched normal-hearing speaker. Modeling, an intonation contour, is a convenient way to elicit the desired range of contrasts. Imitation as an elicitation method was used, for example, in Liberman and Pierrehumbert's (1984) study, in which "for subjects other than the authors, the desired intonation patterns were demonstrated by example before the experiment" (p. 172). Imitation as an experimental tool is nonetheless problematic. As Markham (1997, p. 142) points out, "there is no unequivocal body of evidence for or against the validity of imitative productions as a reflection of normal performance potential," although he does use imitation as an experimental method. A lack of ability to imitate an intonational contour could be due to several factors. One factor causing a poor imitation could be a poor perception of the stimulus. Alternatively, the stimulus could be accurately perceived, but, as Markham (1997, p. 141) suggests, the imitation could still be poor because of "strong influences from within the imitators' phonetic or linguistic processing systems, resulting in considerable deviation from the model." Furthermore, Markham (1997) suggests that people vary in their level of skill at imitations. Thus, there are some problems inherent in using an imitation task as a measure of perception and production. However, imitation is a simple way to elicit a range of intonational contrasts, and the fact that it involves both perception and production is one of its strengths as well as one of its weaknesses. The strength of this method is that if there is a successful imitation, then that implies successful perception, processing, and production of the stimulus. The weakness of this method is that if there is an unsuccessful imitation, it is not clear whether the lack of success is due to problems in perception, processing, or production. With these caveats in mind, an imitation task was used here as a way to roughly assess a cochlear implant user's production and perception of intonational contrasts. More details about the task are provided in the "Production Experiment" section below.

The utterances collected in the imitation task were assessed in two perception experiments with naïve native speakers of English. In "Perception Experiment 1" (see b elow) listeners were asked to rate the goodness of the speakers' imitation in the context of the experimenter's model. In "Perception Experiment 2," listeners were asked to pick which pragmatic context the utterance could have been uttered in, given a choice of four contexts.

Production Experiment

Subjects

Two speakers were recorded. The first speaker (referred to below as "CI-1") was a 35 year old male. His age at onset of deafness was 29, and his age at onset of profound deafness was 30, due to cryoglobal anemia and autoimmune syndrome. He was implanted with a Clarion 8-channel cochlear implant at age 31. At the time the recordings were made, he had used his cochlear implant for 4 years. This subject lived in Indiana his whole life. The second speaker (referred to henceforth as "NH-1") was a normal-hearing 50 year old male who has spent most of his life (except for 6 years during his 20s) in Indiana. Both speakers were paid $10 for participation in the experiment.

Intonational contour	ToBI transcription
"calling contour"	H* !H-L% Marianna!
"yes-no question contour"	L* H-H% Marianna?
"declarative contour"	H* L-L% Marianna.
"surprise-redundancy contour"	L+H* L-H% Marianna?!
"calling contour"	H* !H-L% Anna!
"yes-no question contour"	L* H-H% Anna?
"broad focus"	H* H* L-L% I didn't like taking all those tests.
"narrow focus: like"	H* L+H* L-H% I didn't <u>like</u> taking all those tests,
"narrow focus: tests"	H* L+H* L-L% I didn't like taking all those <u>tests</u>,
"narrow focus: vitamins"	H* H* L-H% Legumes are a good source of <u>vitamins</u>,
"narrow focus: good"	H* L+H* L-H% Legumes are a <u>good</u> source of vitamins,
"narrow focus: legumes"	H* L-H% <u>Legumes</u> are a good source of vitamins,

Table 1. The ToBI transcriptions of the utterances used in the experiment.

Materials

There were four different sets of utterances used. In each set, the "same" utterance, segmentally and lexically speaking, was uttered with several different intonations. One set of target intonation contours used the name "Marianna," referred to below as the "Marianna" sentences. Another set used the name "Anna," referred to below as the "Anna" sentences. A third set used the sentence "I didn't like taking all those tests," referred to below as the "Tests" sentences. The final set used the sentence "Legumes are a good source of vitamins," referred to below as the "Legumes" sentences.

The intonational contours used in this task were selected in order to have a range of functionally important intonational contrasts. The ToBI transcriptions of the utterances are shown aligned with the text in Table 1.

The "Marianna" sentences were originally based on the "calling3" example from the ToBI training guide, in which the name "Marianna" is spoken with a calling contour. The ToBI training guide was taken as a starting point in developing the corpus for this experiment because it presents a standard which has been subjected to inter-transcriber reliability tests and which has emerged as a consensus among many researchers in the field. In the "Marianna" sentences, the first contour was the "calling contour," described by Pierrehumbert (1980) and Ladd (1978). This sentence has a high pitch accent followed by a level tail. Since it is the "calling contour," the context was "Marianna! Come to dinner!" The next contour for the "Marianna" sentences was a yes-no question contour, which has been described in English as having a low pitch accent followed by a high rising tail. In fact, this contour is one of the few contrasts used in batteries of speech perception tests that do assess prosodic factors (Boothroyd, 1988; Boothroyd, 1984; Owens et al., 1981; Owens et al., 1982). The context used for this sentence was "Marianna? Are you there?" The third contour used with the "Marianna" sentences was a simple falling contour, in which there is a high pitch accent followed by a falling tail. Since this contour is often used in "declarative" intonations, the context used in this sentence was "Who ran the experiment? Marianna." The final contour used with the "Marianna" sentences was the so-called "surprise-redundancy" contour, (Sag & Liberman, 1975), in which there is a rise followed by a fall and another rise at the end. Since this contour, as explained by Ladd (1996, p. 140), expresses the sentence "as an incredulous or disapproving reply to someone else's statements," the context used here was "Marianna?! I thought she was away." Thus, the "Marianna" set of sentences included a range of functionally useful combinations of pitch accents, phrase accents, and boundary tones, including both high and low pitch accents, high, downstepped high, and low phrase accents, and both high and low boundary tones.

The "Anna" sentences were borrowed from the example "calling2" in the ToBI training guide. The first "Anna" sentence was spoken with the calling contour (Pierrehumbert, 1980; Ladd, 1978), which has a high pitch accent and a level tail. The context for this sentence was "Anna! Come to dinner!" The second "Anna" sentence was spoken with the yes-no question contour, which has a low pitch accent and a rising tail. The context for this sentence was "Anna? Are you there?" Thus, the "Anna" sentences included two rather stylized contours that are often used in day-to-day conversation.

The "Tests" sentences were developed for this experiment. These sentences were designed to assess the perception and production of contrastive focus on various elements of the sentence. The first "Tests" sentence was spoken with a broad focus, as if in answer to the very general question of "How was your day today?" with the answer being "I didn't like taking all those tests." The second "Tests" sentence was supposed to focus the verb, such that the context was "I didn't <u>like</u> taking all those tests, I loved it!" This sentence was generally uttered with a L+H* and an expanded pitch range on the verb "like." The third

"Tests" sentence was supposed to focus the object noun phrase, such that the context was "I didn't like taking all those tests, but I did enjoy the recording session." In this case, the sentence was generally uttered with a L+H* on the object noun "tests." Thus, this set of sentences was meant to assess the use of contrastive focus, changing the location of the focus from trial to trial.

The "Legumes" sentences were based on the examples "legumes1," "legumes2," and "legumes3" from the ToBI training guide. These sentences derive from Pierrehumbert (1980), in which they are used as a demonstration of nuclear accent, which is the final pitch accent in a phrase, shifting from location to location in the sentence. The first "Legumes" sentence reads "Legumes are a good source of vitamins, and of protein as well." In this case, the nuclear pitch accent falls on "vitamins." The second "Legumes" sentence reads "Legumes are a good source of vitamins, but not the best." In this case, the nuclear pitch accent falls on "good." Although the ToBI training guide has this with a H* on "good," in this experiment the pitch accent used with "good" was generally a L+H*. The third "Legumes" sentence reads "Legumes are a good source of vitamins, and so are greens." In this case, the nuclear pitch accent falls on the first word of the phrase, "legumes." Thus, this set of sentences was meant to assess the perception and production of nuclear pitch accents that shift location from trial to trial and which are used for varying pragmatic effects.[9]

Procedure

The experimenter said a sentence, and the speaker was instructed to imitate the sentence exactly as it was produced. The speakers were told that the emphasis or focus in the sentences would be changing from trial to trial, and that they should imitate the sentences exactly, including the emphasis. Recordings were made in a quiet room using a DAT recorder (recording continuously during the session) with a microphone on the table in front of the speaker, approximately 20 cm. from the speaker's mouth. The stimuli were presented using live voice, and speakers could see the experimenter's face during the session. (The speaker with the cochlear implant is also a skilled speech-reader.) The contexts for each sentence were included, either as a brief description or as a continuation of the target sentence.

At the end of each speaker's session, the speaker was asked to speak spontaneously for a few minutes while being recorded, in order to get an idea of the general characteristics of the voice of the speaker. Suggested topics were job descriptions, where the speakers lived, and where they had lived in the past, all topics intended to be neutral and non-invasive.

The experimenter's voice was also recorded onto the DAT tape from the microphone, but the quality of the recording of the cue phrases as spoken by the experimenter was very poor, because of the greater distance from the microphone. Therefore, because of the need to compare the speakers' renditions of the utterances to the experimenter's renditions in the first perception experiment, the experimenter's cue phrases were re-recorded under the same conditions as the original recording session. Each cue phrase from the original recording session was played out on a cassette tape and immediately mimicked by the experimenter for the new recording, attempting to make as close of an approximation to the original recording as possible in terms of intonation, voice quality, and rate. Unfortunately, this means that in the perception tests, the listeners were not able to compare the speakers' utterances to the original target that they were imitating. Re-recording provided a better sound quality than boosting the sound level of the

[9] Depending on whether the "stress" in the "roving stress" test of the Speech Pattern and Contrasts battery (Boothroyd, 1988) is interpreted as nuclear accent or not, this set of sentences may be somewhat analogous to the "roving stress" test

poor-quality original recordings would have, because boosting the amplitude would have boosted the noise level as well. In the future, changes will be made to improve this procedure.

Recordings were digitized at 16 bit 16 kHz using ESPS *Waves*™ (Entropic Research Laboratory, 1993). Target sentences were extracted using *Waves*™. The target sentences were down sampled to 8 kHz and converted to .wav format for use on a PC by the "Sound Exchange" program, SoX.

Results and Discussion

In general, the two speakers have similar overall pitch ranges, voice quality and accent, as determined subjectively by the authors. Sample F0 traces for the experimenter's model and the speaker's imitation are shown in Figure 1. In the top panel of the figure, the F0 of the experimenter's model for the "Anna, calling" utterance is shown on the left, and the F0 of speaker NH-1's imitation is shown on the right. In the bottom panel of the figure, the F0 of the experimenter's model for the "Anna, calling" utterance is shown on the left, and the F0 of speaker CI-1's imitation is shown on the right. In Figure 1, it can be seen that not only are the imitations fairly good reproductions of the experimenter's F0 contour (relatively speaking, within the speaker's own pitch range), but also that the two speakers have fairly similar F0 ranges.

The speakers' success at imitating the experimenter's model during the recording session was assessed in two perception experiments. The first perception experiment obtained listener judgments of how good of an imitation the speaker's rendition is of the experimenter's model. Listeners heard both the experimenter's model and the speaker's imitation and provided a similarity rating. The second perception experiment assessed listeners' responses as to which context they thought the target sentence was drawn from. This experiment is based on the assumption that if the intonational contour was adequately reproduced by the speaker, it should be a retrieval cue as to which context would be most appropriate. This experimental task was much more difficult for subjects, as will be discussed below.

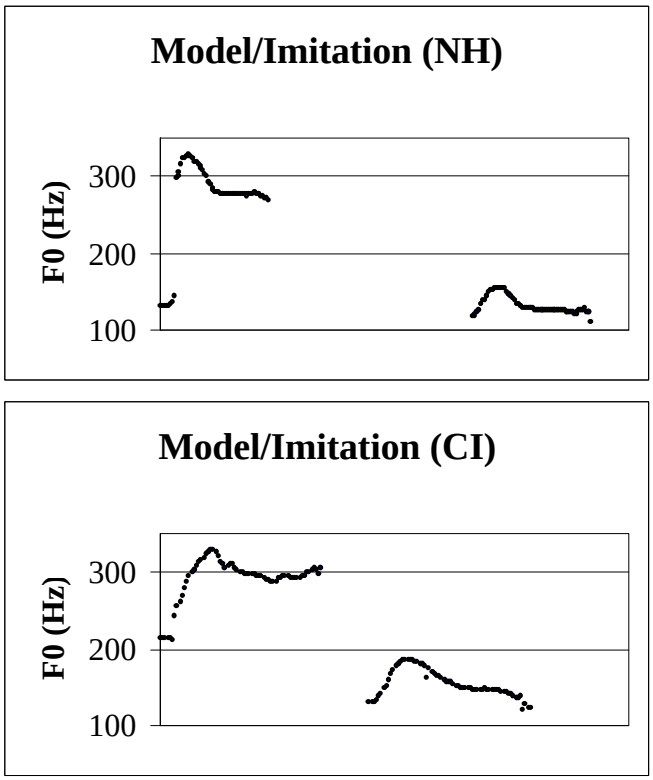


Figure 1. The F0 traces for the "Anna, calling" utterances. The top panel shows the experimenter's model followed by speaker NH-1's imitation. The bottom panel shows the experimenter's model followed by speaker CI-1's imitation.

Perception Experiment 1

Subjects

Thirty-three Indiana University undergraduates, 7 males and 26 females, all native speakers of English with no reported speech or hearing disorders, participated in this experiment as partial fulfillment of a course requirement.

Materials

The materials used in this task were the original 52 target sentences[10] as spoken by the two speakers, along with the (re-recorded) cue sentences as spoken by the experimenter.

[10] This included two repetitions of each of the two "Anna" sentences (4), three repetitions of each of the three "Legumes" sentences (9), two repetitions of each of the four "Marianna" sentences except for one which was not produced correctly by the experimenter during recording (8-1=7), and two repetitions of each of the three "Tests" sentences (6). There were 26 utterances for each speaker, resulting in 52 total utterances.

Procedure

Listeners were seated at a computer monitor equipped with "Beyerdynamic DT100" headphones. The utterances were presented in pairs to listeners over headphones at a comfortable listening level. Each pair of utterances consisted of the experimenter's utterance followed by the speaker's rendition of the utterance. As each pair of utterances was presented, a rating scale from 1 to 7 appeared in a dialog box on the computer monitor. Listeners used the computer mouse to click on the rating which best characterized that imitation. Seven was considered "a very good imitation" and one was "an unsuccessful imitation." Listeners also had the option of listening to each sentence again, for as many repetitions as they wanted to. This rating task was self-paced.

The experiment was preceded by a sample session in which listeners heard three pairs of utterances. In each pair, the listeners heard two utterances spoken by the experimenter (making this slightly different from the actual experimental task). In one case, the two utterances in the pair were spoken with the same intonational contour, and in two of the cases, the two utterances in the pair were spoken with different intonational contours. This practice session was intended to familiarize the listeners with the task and the procedures to be used.

In the main experiment, the pairs of utterances were presented in blocks according to speaker. Within each block, the order of the utterances was randomized. Thus, the order of presentation of speaker CI-1's utterances was different than the order of presentation of speaker NH-1's utterances. Group A (consisting of 16 listeners) heard speaker CI-1's imitations in random order first and then speaker NH-1's imitations in random order. Listeners heard these two blocks twice, in the same order each time. Group B (consisting of 17 listeners) heard speaker NH-1's imitations in random order first (but in the same random order as Group A had heard them), followed by speaker CI-1's imitations in random order (again, in the same random order as heard by Group A). Listeners heard these two blocks of utterances twice, in the same order each time. All listeners heard the utterances in the same order of presentation.

Results and Discussion

The mean rating used by each listener in the main experiment (excluding the practice trials, but including both repetitions of the task) was calculated. These ratings are displayed in Figure 2. Although a few of the individual listeners' ratings tended to be skewed towards one end of the scale or the other, it can be seen that there is an approximately normal distribution of the 33 listeners' mean ratings centered around 4.0-4.5.

The analysis here deals only with the first block of the task, consisting of a block of utterances as produced by speaker NH-1 and a block of utterances produced by speaker CI-1. The second block of the task, consisting of the second time that listeners heard speaker NH-1 and the second time that listeners heard speaker CI-1, will not be analyzed here. The mean rating for each utterance type (the "Anna, calling" sentences, the "Anna, yes/no" sentences, etc.) was calculated separately, for each group and for each speaker.

An unpaired, 2-tailed[11] t-test was conducted using the mean ratings for Group A vs. Group B for speaker CI-1. Another t-test was conducted using the mean ratings of Group A vs. Group B for speaker

[11] This t-test was 2-tailed because we did not have an expectation of which speaker would be rated as the "better" imitator in this task. If we had expected NH-1 to be rated as "better," then we would have used a 1-tailed t-test.

NH-1. The t-tests showed no significant difference between Group A and Group B for either speaker ($\underline{t}(22)=.32$, n.s., for speaker NH-1; $\underline{t}(22)=.14$, n.s., for speaker CI-1). This rules out a bias in mean ratings by order of presentation of speakers. Therefore, since the order of presentation of speakers did not result in a significant difference, the two groups of subjects (those who heard speaker NH-1 first and those who heard speaker CI-1 first), were pooled together for all subsequent analyses.

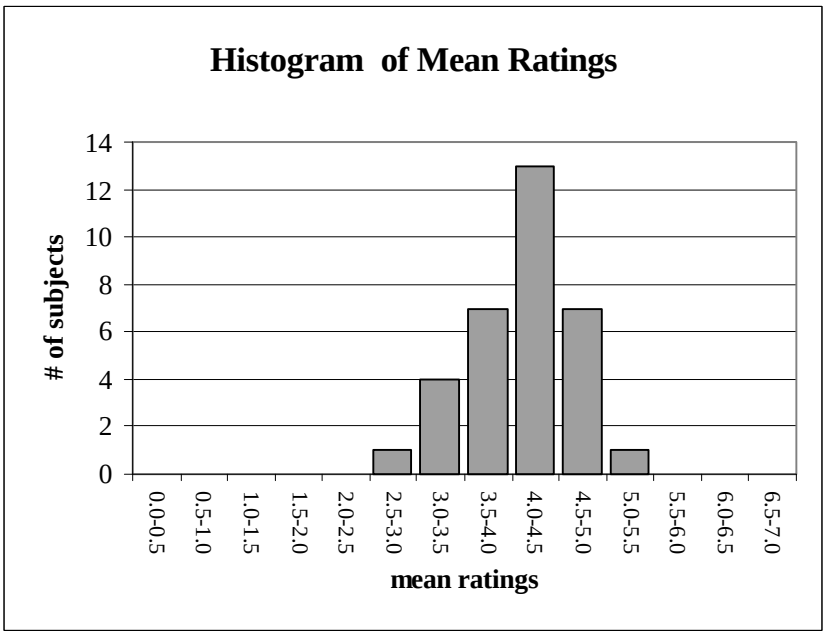


Figure 2. The mean ratings for all subjects in the "goodness-of-imitation" judging task.

Next, the means for all 33 listeners together were calculated, separated out by utterance type and by speaker. Only the means from the first repetition of the task are considered here. These means are plotted in Figure 3. Notice that there are overall differences among listener ratings for utterance types, corresponding to subjective impressions that the imitators did not do as well on the "legumes" utterances, for instance. These differences are indicative that the rating task is serving as a diagnostic measure, since it was sensitive enough to find differences in performance among the four different utterance types.

A paired, 2-tailed t-test between the set of means for all listeners for speaker NH-1 and the set of means for all listeners for speaker CI-1 was conducted. This t-test showed no significant difference between the mean ratings for the two speakers. ($t(11)=-.38$, n.s.) This suggests that listeners are not giving statistically different "goodness of imitation" judgments for the two speakers, implying that the two speakers' performances on the imitation task yield perceptually similar results. Recall that speaker CI-1 is a very skilled cochlear implant user, and sounds "fine" in conversation. We expect that the same sequence of imitation task using a less skilled cochlear implant user as a speaker followed by the same type of perception experiment would produce statistically significant differences between the judgments for such a speaker and for a normal-hearing speaker.

Statistical power for this experiment is low due to a small effect size (d=0.19). This small effect size is indicative of the similarity between the imitative abilities of speakers CI-1 and NH-1 as judged by

listeners in this task. Future experiments will need to be more sensitive to differences between the speakers in order to adequately increase power.

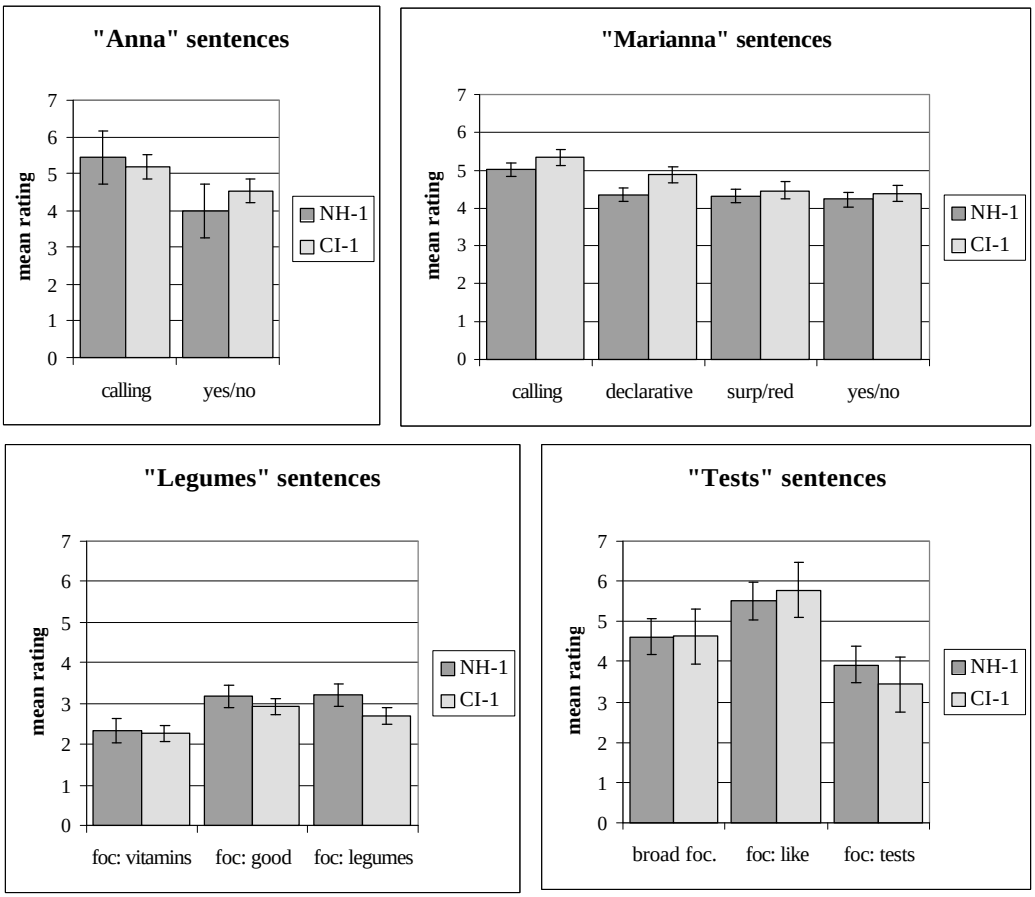


Figure 3. Mean ratings for all subjects for the each utterance type, comparing speaker NH-1 and speaker CI-1, including standard error bars.

Thus, in general the results of this task show that the two speakers, CI-1 and NH-1, are not performing statistically differently from each other on the imitation task, as judged by naïve listeners. While there are differences between the four sentence types, perhaps due to differences in familiarity of words used in the sentences, in general, the two speakers are performing at a similar level to each other. The pattern of results suggests that CI-1 is producing and perhaps perceiving intonational contours as well as NH-1 in this task, despite the degraded phonetic input that he is receiving.

Perception Experiment 2

Subjects

Thirty-two Indiana University undergraduates, 8 males and 24 females, all native speakers of English with no reported speech or hearing disorders, participated in this experiment as partial fulfillment of a course requirement. These subjects had not participated in Perception Experiment 1.

Materials

The materials used in this task were the original 24 target sentences[12] as spoken by the two talkers, NH-1 and CI-1. One sentence token of each type was selected for use in this task. For example, one "Anna, calling" token was chosen for speaker CI-1, one "Anna, yes/no" token was chosen for speaker CI-1, and so forth for all of the different utterance types for each speaker. The "best" utterance token for each sentence type was selected by the experimenters based on repeated listening and consensus. In future research, it might be better to select the "best" utterance using goodness of imitation ratings such as those from Perception Experiment 1.

Procedure

Listeners were seated at a computer monitor equipped with Beyerdynamic DT100 headphones. The target utterances, extracted from the context in which they were uttered, were presented one at a time to listeners over headphones at a comfortable listening level. As each utterance was presented, the sentence and four possible contexts or continuations appeared on the computer monitor. The contexts used in this experiment are shown in Table 2. On each trial, the listener saw four alternatives on the computer screen. Each alternative contained both the original utterance (which was heard by the listeners) and a potential continuation for that utterance, which was not heard by the listeners but was meant to be inferred on the basis of the "way that the utterance was spoken." Each utterance had 4 potential responses presented to the listeners, even the utterances that were only spoken with 2 or 3 distinct intonational contours in the recording session. Listeners used the computer mouse to click on the best possible context for the utterance of the sentence. In addition, listeners indicated a confidence rating on a scale of 1 to 7, with 7 being "most confident" that this was the best choice of context or continuation for the utterance and 1 being "least confident." They also had the option to listen to each sentence again, as many times as they wanted to. Thus, this was a self-paced forced choice categorization task.

The pairs of utterances (the experimenter's model and the speaker's imitation) were presented in blocks according to speaker. Within each block, the order of the utterances was randomized. Thus, the order of presentation of speaker CI-1's utterances was different than the order of presentation of speaker NH-1's utterances. Group A (consisting of 16 listeners) heard speaker CI-1's utterances in random order first and then speaker NH-1's utterances in random order. Listeners heard these 2 blocks three times, in the same order each time. Group B (also consisting of 16 listeners) heard speaker NH-1's utterances in random order first (but in the same random order as Group A), followed by speaker CI-1's utterances in random order (again, in the same random order as heard by Group A). Listeners heard these two blocks of utterances 3 times, in the same order each time.

Results and Discussion

Since each utterance had in fact been uttered in a particular context and was supposed to indicate a particular pragmatic message, that meant that each utterance could be assigned to a specific "correct" context. All of the listeners' responses were scored for correctness according to the expected response for that context condition.

[12] There was 1 token of each type of utterance. There were 2 "Anna" sentences, 3 "Legumes" sentences, 4 "Marianna" sentences, and 3 "tests" sentences. Thus, there were 12 utterances for each speaker, resulting in 24 total utterances.

"Anna" sentence contexts:
1. Anna! come to dinner.
2. Anna? Are you there?
3. Who ran the experiment? Anna.
4. Anna?! But I thought she was away last week!
"Legumes" sentence contexts:
1. Legumes are a good source of vitamins (and of protein as well).
2. Legumes are a good source of vitamins (but not the best).
3. Legumes are a good source of vitamins (and so are greens).
4. Legumes are a good source of vitamins?! (I didn't know that!)
"Marianna" sentence contexts:
1. Marianna! Come to dinner!
2. Marianna? Are you there?
3. Who ran the experiment? Marianna.
4. Marianna?! I thought she was away.
"Tests" sentence contexts:
1. (I had kind of a bad day today.) I didn't like taking all those tests.
2. I didn't like taking all those tests. (I loved it!)
3. I didn't like taking all those tests. (But I did enjoy the recording session.)
4. I didn't like taking all those tests. (But my roommate did.)

Table 2. The options for the contexts/continuations presented to the subjects during the perception experiment.

A probability distribution was calculated for each group of utterances. For the "Anna" group of utterances, there were 2 utterance types, heard 3 times each. Thus, there were 6 trials in this group. Since there were 2 possible correct responses, the probability of getting a correct answer was .5. The probability distribution is shown in Table 3 for each possible number of correct responses. According to the distribution, in order for there to be a probability of less than .05 ($p<.05$) that the number correct is due to chance, the listener would have to get 6/6 correct. This cell is shaded in the table. For the "Legumes" and "Tests" groups of utterances, there were 3 utterance types, heard 3 times each. Thus, there were 9 trials in this group. Since there were 3 possible correct responses, the probability of getting a correct answer was .33. According to the probability distribution for these, in order for there to be a probability of less than .05 ($p<.05$) that the number correct is due to chance, the listener would have to get at least 6/9 correct. These cells are shaded in the table. For the "Marianna" groups of utterances, there were 4 utterance types, heard 3 times each. Thus, there were 12 trials in this group. Since there were 4 possible correct responses, the probability of getting a correct answer was .25. According to the probability distribution for these, in order for there to be $p<.05$ that the number correct is due to chance, the listener would have to get at least 6/12 correct. These cells are shaded in the table.

For the two groups of listeners Group A (who heard speaker CI-1 first) and Group B (who heard speaker NH-1 first), the number of listeners performing above chance level was tallied (pooling together the results for speaker NH-1 and speaker CI-1). The tallying of listeners performing above chance level was done separately for the different utterance types. Thus, if a listener scored above chance level for the "Marianna" utterances but not for the "Legumes" utterances, that listener would be counted as performing above chance for the "Marianna" sentences only. Listeners who did not perform above chance level for a particular sentence type were excluded from the final analysis.

"Anna"	
# correct	Probability that # correct is due to chance
1	.09
2	.23
3	.31
4	.23
5	.09
6	.02

"Tests" & "Legumes"	
# correct	Probability that # correct is due to chance
1	.12
2	.24
3	.27
4	.20
5	.10
6	.03
7	.007
8	.0008
9	.00005

"Marianna"	
# correct	Probability that # correct is due to chance
1	.13
2	.23
3	.26
4	.19
5	.10
6	.04
7	.01
8	.002
9	.0003
10	.00004
11	.000002
12	.00000006

Table 3. The probability distributions for the different groups of utterances. The shaded cells indicate how many correct responses are needed for the probability to be less than .05 that the number correct is due to chance.

As it turns out, this was a very difficult task for listeners to perform, and many of the listeners had to be excluded from the final analysis because they did not perform above chance level. In some cases, only 4, 5, or 6 out of the 16 listeners in a group performed above chance for a particular utterance type. There are several explanations for the poor performance of so many of the listeners. First, these utterances were presented in isolation out of context, without a conversational framework, which is an unnatural situation. Second, this task involved reading and evaluating four contexts, which may have involved too much attention and effort for some of the participants. Interestingly, though, there were differences between the groups of utterances, with many more subjects performing above chance for the "Marianna" utterances than for the other three types of utterances. Thus, this indicates that this task is indeed serving as a diagnostic measure, measuring differences among the different groups of utterances and types of intonation contours, even though it was a very difficult task.

The number of listeners who did display above chance performance levels were tallied. A paired, 2-tailed t-test on the average scores for subjects performing above chance, comparing Group A to Group B (those who had heard speaker CI-1 first and those who had heard speaker NH-1 first), showed no significant difference between the two groups ($t(6)=-.096$, n.s.) Thus, the data for the two groups were combined together in all further analyses. Next, the number of subjects performing above chance level for each type of sentence, separated out by speaker NH-1 vs. speaker CI-1, were tallied. These figures are displayed in Figure 4.

A paired 2-tailed t-test was calculated, comparing average scores for listeners performing above chance level for speaker NH-1 vs. for speaker CI-1. ($t(3)=-.97$, n.s.) This t-test indicated no significant difference between the number of listeners performing above chance level on the two different speakers'

utterances. Power for this experiment is low due to the exclusion of so many of the subjects as a result of the difficulty of the task. However, the null result is consistent with the finding from the first perception experiment.

The task used in the second experiment was difficult to perform and many of the listeners had to be excluded from analysis of one or another sets of utterances due to not performing above chance level.

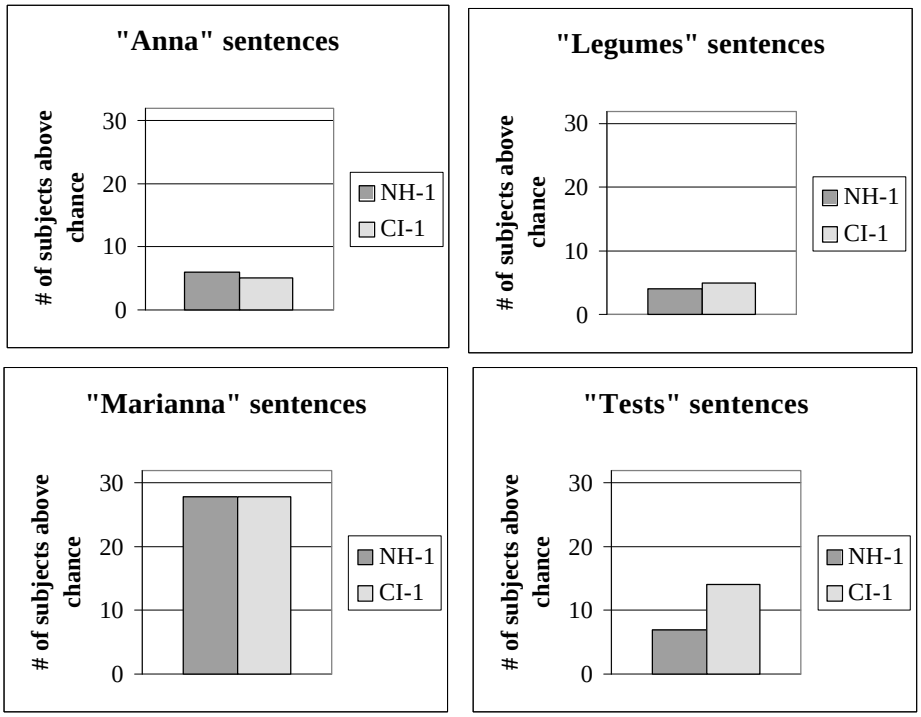


Figure 4. The number of subjects who did perform above chance for sentences uttered by Speaker NH-1 and Speaker CI-1 (out of 32 total listeners), separated out by utterance type.

Nonetheless, when considering only subjects performing above chance level for a particular group of utterances, it did turn out that there was no significant difference between the average scores for listeners of speaker CI-1 and of speaker NH-1. This result indicates that the two speakers are conveying the intonational meaning with approximately the same level of adequacy, as far as the listeners are concerned. This confirms the earlier results and demonstrates that speaker CI-1 is performing as well as speaker NH-1 on the imitation task, despite the degraded phonetic input which he receives.

General Discussion

In general, the sentence imitation task used here served as a good way of eliciting a fuller range of intonational contours than had been tested in the past. However, several issues arose during recording and evaluation which should be changed to improve the procedure. For example, it might be better to have pre-recorded stimuli for imitation rather than live voice, to eliminate the possibility of speech-reading of prosodic cues from the experimenter's face and also to assure consistency of the model for all speakers.

Also, it might be better to have a male voice serve as the model for male speakers, so as to provide a closer model for imitation. In addition, the subjects did seem to have more trouble producing the "legumes" utterances than the other utterances, perhaps due to the unfamiliarity of the word "legumes." Thus, it might be better to use some standardized utterances in this type of task. Furthermore, it would have been better to use a head-mounted microphone rather than a microphone on the table. In terms of the perception tasks, performance might have been much better if the listeners received some training and prior experience on the context-judging task from Perception Experiment 2. Then, only listeners who could perform the task would be included in the experiment. These are a few of the changes that should be made if this pilot experiment is extended to studies of other speakers.

As expected, the results of both perception tasks showed that there were no significant differences in judgments of speaker CI-1's utterances and speaker NH-1's utterances. Rather than simply interpreting this as a null result, this can be interpreted as confirming the informal observation that speaker CI-1's intonation sounds "fine" in conversation. However, this is interesting because the cochlear implant user is performing as well as the normal hearing speaker, even though he is receiving degraded phonetic input (because the cochlear implant does not have good resolution of F0). Thus, we can infer that he is using the richness of the phonetic input and the co-variance of acoustic cues such as harmonic structure, amplitude, and duration to phonologically parse the input into a grammatical utterance which he can then reproduce. Also, given the short length of time that he was profoundly deaf, we can conclude that he must be using a robust phonological representation of intonation in parsing degraded input.

In summary, this pilot study of the perception and production of intonational contrasts in an adult patient with a cochlear implant provides a good baseline study of a highly skilled user. It would be interesting in the future to examine more closely the phonetic inputs that are being transmitted by the cochlear implant, as well as examining less skilled users and individuals with pre-lingual profound hearing loss. This line of research will provide insights into the interface between phonetics and phonology by examining the abilities of speakers with incomplete phonetic inputs to perceive and produce linguistically meaningful pitch changes.

References

Beckman, M. (1986). *Stress and Non-Stress Accent.* Foris: Dordrecht.

Beckman, M. and G. Ayers. (1994). *ToBI Labelling Guide*. The Ohio State University.

Bolinger, D. (1986). *Intonation and its Parts: Melody in Spoken English.* Stanford, CA: Stanford University Press.

Boothroyd, A. (1988). Perception of speech pattern contrasts from auditory presentation of fundamental frequency. *Ear and Hearing, 9*, 313-321.

Boothroyd, A. (1984). Auditory perception of speech contrasts by subjects with sensorineural hearing loss. *Journal of Speech and Hearing Research, 27,* 134-144.

Brimacombe, J. A., B. J. Edgerton, K.J. Doyle, J. D. Erratt, and J. L. Danhauer. (1984). Auditory capabilities of patients implanted with the House single-channel cochlear implant. *Acta Otolaryngologica, Supplement 411,* 204-216.

Clark, H. and S. Haviland. (1977). Comprehension and the Given-New Contract. In R. Freedle (ed.) *Discourse Production and Comprehension.* Ablex: Norwood, NJ. pp. 1-40.

Entopic Research Laboratory, Inc. (1993). *Waves+* 5.0. Washington, D.C.

Fougeron, C. and P. Keating. (1997). Articulatory strengthening at the edges of prosodic domains. *Journal of the Acoustical Society of America, 101,* 3728-3740.

Hirschberg, J. and J. Pierrehumbert. (1986). The intonational structuring of discourse. *Proceedings of the 24th Annual Meeting for the Association for Computational Linguistics,* 136-144.

Huang, T.-S., N.-M. Wang, and S.-Y. Liu. (1995). Tone perception of Mandarin-speaking postlingually deaf implantees using the Nucleus 22-channel cochlear mini system. *The Annals of Otology, Rhinology, and Laryngology, supplement 166,* 294-298.

Jackendoff, R. S. (1972). *Semantic Interpretation in Generative Grammar.* MIT Press.

Jones, P.A., H. J. McDermott, P. M. Seligman, and J. B. Millar. (1995). Coding of voice source information in the Nucleus cochlear implant system. *The Annals of Otology, Rhinology, and Laryngology, supplement, 166,* 363-365.

Kirk, K. I. and B. J. Edgerton. (1983). The effects of cochlear implant use on voice parameters. *The Otolaryngologic Clinic of North America, vol. 16,* 281-292.

Ladd, D. R. (1996). *Intonational Phonology.* Cambridge University Press.

Ladd, D. R. (1978). Stylized intonation. *Language, 54,* 517-539.

Ladd, D. R. and R. Morton. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics, 25,* 313-342.

Lane, H., J. Wozniak, M. Matthies, M. Svirsky, J. Perkell, M. O'Conell, and J. Manzella. (1997). Changes in sound pressure and fundamental frequency contours following changes in hearing status. *The Otolaryngologic Journal of the Acoustical Society of America, 101,* 2244-2252.

Leder, S. B. and J. B. Spitzer. (1990). Longitudinal effects of single-channel cochlear implantation on voice quality. *Laryngoscope, 100,* 395-398.

Leder, S.B., J. B. Spitzer, P. Milner, C. Flevaris-Phillips, F. Richardson, and J. C. Kirchner. (1986). Reacquisition of contrastive stress in an adventitiously deaf speaker using a single-channel cochlear implant. *Journal of the Acoustical Society of America, 79,* 1957-1974.

Liberman, M. and J. Pierrehumbert. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oerhle (eds.) *Language Sound Structure.* Cambridge, MA: MIT Press. pp. 157-233.

Lim, H. H. and Y. C. Tong. (1989). Pitch scaling, difference limens, and identification of electrical pulse rate on multi-channel cochlear implant patients—Implications on temporal speech processing. *Proceedings of the Annual International Conferences of the IEEE, 11,* 1065-1066.

Markham, Duncan. (1997). *Phonetic Imitation, Accent, and the Learner*. Lund University Press.

Monini, S., G. Banci, M. Barbara, M. T. Argiro, and R. Filipo. (1997). Clarion cochlear implant: Short term effects on voice parameters. *The American Journal of Otology, 18,* 719-725.

Owens, E., D. K. Kessler, C. C. Telleen, and E. D. Schubert. (1981). The Minimal Auditory Capabilities Battery (Instruction Manual). St. Louis: Auditec.

Owens, E., D. K. Kessler, C. C. Telleen, and E. D. Schubert. (1982). The Minimal Auditory Capabilities (MAC) Battery. *Hear Aid Journal, 34,* 9-34.

Palmer, H. (1922). *English Intonation, with Systematic Exercises*. Cambridge: Heffer.

Perkell, J., H. Lane, M. Svirsky, and J. Webster. (1992). Speech of cochlear implant patients: A longitudinal study of vowel production. *Journal of the Acoustical Society of America, 91,* 2961-2978.

Pfingst, B. E., L. A. Holloway, N. Poopat, A. R. Subramanya, M. F. Warren, and T. A. Zwolan. (1994). Effects of stimulus level on non-spectral frequency discrimination by human subjects. *Hearing Research, 78,* 197-209.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation.* Bloomington, Indiana: Indiana University Linguistics Club.

Pierrehumbert, J. and J. Hirschberg. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack (eds.) *Intentions in Communication.* The MIT Press: Cambridge, Massachusetts. pp. 271-311.

Quillet, C., R. A. Wright, and D. B. Pisoni. (1998). Perception of "place-degraded speech" by normal-hearing listeners: Some preliminary findings. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 353-375). Bloomington, IN: Speech Research Laboratory, Indiana University.

Remijsen, B. and V. J. van Heuven. (1999). Gradient and categorical pitch dimensions in Dutch: Diagnostic test. *Proceedings of the 14th International Congress of Phonetic Sciences.* 1865-1868.

Richardson, L. M., P. A. Busby, P. J. Blamey, and G. M. Clark. (1998). Studies of prosody perception by cochlear implant patients. *Audiology, 37,* 231-245.

Rosen, S., J. Walliker, J. A. Brimacombe, and B. J. Edgerton. (1989). Prosodic and segmental aspects of speech perception with the House/3M single-channel implant. *Journal of Speech and Hearing Research, 32,* 93-111.

Sag, I. and M. Liberman. (1975). The intonational disambiguation of indirect speech acts. *Proceedings of the Chicago Linguistic Society, 11,* 487-497.

Selkirk, E. O. (1984). *Phonology and Syntax: The Relation Between Sound and Structure.* MIT Press.

Shannon, R. V. (1983). Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics. *Hearing Research, 11,* 157-189.

Sluijter, A. and V. van Heuven. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America, 100,* 2471-2485.

Tong, Y. C., G. M. Clark, P. J. Blamey, P. A. Busby, and R. C. Dowell. (1982). Psychophysical studies for two multiple-channel cochlear implant patients. *Journal of the Acoustical Society of America, 71,* 153-160.

Tong, Y. C., P. J. Blamey, R. C. Dowell, and G. M. Clark. (1983). Psychophysical studies evaluating the feasibility of a speech processing strategy for a multiple-channel cochlear implant. *Journal of the Acoustical Society of America, 74,* 73-80.

Townshend, B., N. Cotter, D. van Compernolle, and R. L. White. (1987). Pitch perception by cochlear implant subjects. *Journal of the Acoustical Society of America, 82,* 106-115.

Walker, M. A., A. K. Joshi, and E. F. Prince. (1998). Centering in naturally occurring discourse: An overview. In M. A. Walker, A. K. Joshi, and E. F. Prince (eds.) *Centering Theory in Discourse.* Oxford: Clarendon Press. pp. 1-28

Wang, B. K. and T. S. Huang. (1988). Current clinical results of the cochlear implant program conducted on Mandarin-speaking patients. *American Journal of Otology, 9,* 44-51.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Speech Intelligibility of Pediatric Hearing Aid Users[1]

Mario A. Svirsky, Steven B. Chin, Matthew D. Caldwell and Richard T. Miyamoto[2]

DeVault Otologic Research Laboratory
Department of Otolaryngology-Head & Neck Surgery
Indiana University School of Medicine
Indianapolis, IN 46202

[1] This study was supported by NIH-NIDCD grants DC00064 and DC00423. Terri Kerr assisted with data management; Allyson Riley, Amy Robbins, and Susan Sehgal collected the data, and Karen I. Kirk and Susan Sehgal gave useful suggestions for the manuscript. All this help is gratefully acknowledged.
[2] All of the authors are also affiliated with the Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405.

Speech Intelligibility of Pediatric Hearing Aid Users

Abstract. This study examined the speech intelligibility of profoundly, prelingually or congenitally deaf children who use hearing aids**.** Children were one to fifteen years old and they were classified into subgroups according to residual hearing (PTA between 90 and 100, 100 and 110 or greater than 110 dB HL) and communication mode (either oral or total communication). They read lists of standard sentences which were played back to panels of three "naive" listeners who were not familiar with the speech of the deaf and who did not know which subgroup the children belonged to. The data revealed a strong significant trend toward higher intelligibility for children with more residual hearing, and a significant trend toward higher intelligibility for users of oral communication than for total communication users. However, the latter trend was much more pronounced for some ranges of residual hearing than for others, and it may have been partly due to a sampling effect. A third trend showed significantly higher intelligibility levels at older ages, but this was particularly pronounced for children with PTA's between 90-100 dB HL, and for the majority of oral communication users (and only a few total communication users) with PTA's between 100-110 dB. These results suggest that the amount of residual hearing (possibly in interaction with the communication mode used by the child) may be an important factor in the development of intelligible speech.

Introduction

A connection between prelingual hearing impairment and atypical speech production has been observed for millennia (e.g., Hippocrates, 1853, p. 609), and it is known that the speech of prelingually deafened children can be very difficult to understand. Nevertheless, throughout history, people have attempted, with varying degrees of success, to impart to children with hearing impairment the skills necessary to allow them to communicate successfully in a world so dependent on spoken language. Among the myriad measures used to assess speech production in children (e.g., phonemes correct, suprasegmentals correct, aerodynamic deviance), the one with the highest face validity as regards the need to communicate using spoken language is overall speech intelligibility. Measures of overall speech intelligibility cut to the quick, addressing the important question, "Can this child be understood?"

The question of whether a child with hearing impairment can be understood appears a simple question, but stating the question, and, perhaps more importantly, answering the question, do not by any means lack scientific subtlety. The question "Can this child be understood?" immediately raises obvious follow-up questions: "By whom?" and "How much of the time?" and "At what level?" The scientific literature on speech intelligibility among children with hearing impairment is not uniform in its approaches to these questions, and how best to address the question of speech intelligibility can be as much a matter of practical considerations as of theory.

Measuring the speech intelligibility of children with hearing impairment has generally relied on two types of tasks: (1) rating scales and (2) write-down procedures (Samar & Metz, 1988). In the first type of task, listeners make explicit judgments about the talker's (overall) speech intelligibility by assigning numerical values to samples of speech (e.g., the NTID rating scale: Subtelny, 1977), whereas in the second type of task, listeners "write down what they thought each child said" (Monsen, 1981). As Metz, Schiavetti, & Sitler (1980) point out, write-down procedures appear to have higher face validity than a rating task, because results depend on the listener actually and literally understanding what has been said. Moreover, write-down procedures are relatively insensitive to vocal qualities of speech; these

may contaminate responses on a rating scale, but they need not of necessity degrade message intelligibility (Samar & Metz, 1988). On the other hand, write-down protocols are time consuming and labor intensive and therefore expensive (but see Samar & Metz, 1988), whereas rating tasks are viewed as relatively "quick and easy" (Metz et al., 1980). Some considerations making write-down intelligibility assessments difficult to administer are that listeners must respond to every word heard, and that individual listeners should not hear the same word or sentence more than once, to avoid an order effect from increased familiarity with the spoken material. Disadvantages of rating scale tasks are the relatively lower face validity of the protocol, as noted above, and the lack of material for phonemic analysis of listener misidentifications.

A second consideration in speech intelligibility assessment is the type of listener judge, roughly, experienced vs. inexperienced (naive). "Experienced" generally means experienced with the speech of people with hearing-impairments, and such experienced listeners may include educators of the deaf, audiologists, and speech-language pathologists (e.g., McGarr, 1983). Naive listeners are those who have had little or no experience with the speech of the hearing impaired. There is often a connection between the type of task involved in speech intelligibility assessment and the type of listeners involved in the protocol. Rating scales, especially equal-interval scales, are often administered to experienced listeners (e.g., Subtelny, 1977), whereas inexperienced listeners very often participate in write-down tasks (e.g., John & Howarth, 1965).

A third parameter in the assessment of speech intelligibility is the type of speech material used, including both elicitation protocols and the linguistic level of analysis. One parameter in elicitation procedures is that of spontaneous speech vs. read speech (Smith, 1982). If spontaneous speech is elicited, a semantic/pragmatic context is generally established, such as a set of pictures either for description or as basis for a story (on the other hand, context can be an independent variable within a single investigation e.g., Becker, Schildhammer, & Ruoß, 1994). Read speech is usually elicited as a set of full sentences (e.g., Hudgins & Numbers, 1942; Maasen & Povel, 1985; Monsen, 1978, 1983; Osberger & Levitt, 1979), although the SPINE (Monsen, 1981) elicits read versions of isolated words. One variation of read speech is used with preliterate children, who may be asked to repeat sentences spoken to them by an examiner (e.g., Osberger, Robbins, Todd, & Riley, 1994). The level of analysis for measuring speech intelligibility also varies across studies and instruments. Rating scales generally apply to linguistic units at least as large as sentences and ranging up to entire passages (e.g., The Rainbox Passage, from Fairbanks, 1960). The NTID Speech Intelligibility Scale (Subtelny, 1977), for instance, rates connected spontaneous discourse on a 5-point scale ranging from "Speech cannot be understood" (1) to "Speech is completely intelligible" (5). Analysis units for identification tests are generally either words or sentences. Words can be tested either in isolation (e.g., the SPINE test: Monsen, 1981) or in phrasal or sentential contexts (e.g., John & Howarth, 1965; Maassen & Povel, 1985; Markides, 1970; McGarr, 1983). Words in phrases or sentences can be scored on the basis of keywords only (e.g., McGarr, 1983; Smith, 1975; or CID Everyday Sentences [Hirsch et al., 1952]); all words equally (e.g., John & Howarth, 1965); or all words, with weighting (e.g., Monsen, 1978, 1983, with different scores for individual words based roughly on frequency and predictability). When whole sentences or phrases are the units of analysis, listeners must perforce hear each word correctly; in the scoring procedure of Hudgins and Numbers (1942), for example, "no credit was allowed for partially correct auditions" (p. 302).

Aside from construct variables in speech intelligibility assessment, research on speech intelligibility may also differ according to subject characteristics. Some important subject variables in studies of the speech intelligibility of children with hearing impairments are chronological age, degree of hearing loss (conversely, amount of residual hearing), age at onset of hearing loss, duration of hearing loss, and communication mode (oral communication vs. total communication [i.e., spoken and signed]).

In the present study, we assessed the speech intelligibility of hearing aid users in a write-down task (because of its higher face validity) with naive listeners. Write-down procedures date to at least Hudgins and Numbers (1942), an early study using this procedure to examine the speech intelligibility of deaf children. In that study, experienced, rather than naive, listeners served as judges, because the authors believed that inexperienced listeners would be distracted by atypical voice characteristics and would "thus lose much of the content" (p. 301). One-hundred ninety-two children participated and were divided into three groups according to the degree of hearing loss, using the classification of Guilder and Hopkins (1936). The children were between 8 and 20 years of age, and all were being educated at oral schools. Phonographic recordings were made of the children reading 10 sentences (e.g., "Sally likes to swim"). Listeners either were teachers of the deaf or were training to be teachers of the deaf. An average of seven listeners heard each of the sentences. Listeners heard each sentence three times and were instructed to "write down what you think the child says after each reproduction"; listeners were allowed to correct what they had written on previous trials. Responses were scored according to whole-sentence-correctness, that is, regardless of the number of words correctly or incorrectly transcribed. Mean intelligibility across subjects with all degrees of hearing loss was 29.2%.

Whereas Hudgins and Numbers (1942) used experienced listeners and scored whole sentences, John and Howarth (1965) used inexperienced listeners and scored words in a study of the effects of time distortions on the speech intelligibility of deaf children. Twenty-nine children each read a sentence both before and after intensive speech training. Five children had hearing losses below 80 dB, five with losses of 80-89 dB, five with losses of 90-99 dB, and fourteen with losses of 100 dB or more. Twenty "lay" listeners listened to tapes of the 29 sentences both before and after training and were instructed to "write down as much as they could of each phrase" (p. 131). Listeners' responses were scored according to the number of words heard correctly. Across the entire group of 29 children, which included 10 with less than profound hearing losses, the percent words correctly understood by listeners (the measure of intelligibility) was 19% before training. Using methods adapted from John and Howarth (1965), Markides (1970) examined the speech intelligibility of 58 deaf children (mean hearing loss = 95 dB). Stimulus materials were five unrelated pictures for description by the participating children, who were tape-recorded. These recordings were played to panels of listeners consisting of three university students "who were totally naive with regard to speech of deaf children" (p. 128). Listeners were instructed to write down as much as they could of what the child said; scores for each child were means of the number of words correctly transcribed by the three judges. The mean score across children for responses by these naive listeners was 160 words correctly transcribed of 825 words produced, or 19.4%.

Smith (1975) examined the relationship between residual hearing and speech production in deaf children. Forty children participated in the study; all but three had pure tone average (PTA) thresholds of at least 92 dB in the better ear. Each child recorded a list of 20 sentences, and each sentence was later played for three listeners, "without significant previous experience in hearing the speech of deaf persons" (p. 797). Listeners were allowed to hear sentences up to twice each and then wrote down what they thought the children had said. Identification responses were scored word by word; scores for each child were the percentage of keywords correctly understood. Across children and listeners, percent correct scores ranged from 0 to 76.1%, with a mean of 18.7% (SE = 3.2). Monsen (1978) compared intelligibility and acoustic measures in the speech of hearing-impaired adolescents. Thirty-seven adolescents ranging in age from 12;6 (years; months) to 16;6 participated; 27 of the subjects had PTA thresholds better than 95 dB in the better ear. Speech materials were simple sentences containing only common monosyllables and spondees; these were read by the subjects from typewritten copies and audio-recorded. Two successive repetitions of each sentence were played through a loudspeaker to naive listeners "who had never before knowingly heard the speech of a hearing-impaired person" (p. 202). Listeners were instructed "to write down in normal English orthography what he though each subject said, to guess if necessary..." Scoring was based on a maximum value of 10 for each sentence, with scores for individual words assigned

according to frequency in the language (low scores = high frequency and predictability; high scores = low frequency and predictability). There were no partial scores for partially correct word responses. Across talkers and listeners, mean intelligibility ranged from 31.3% correct to 99.9% correct, with an average intelligibility score of 76.7%.

Monsen (1981) described "an easy way to use an accurate test for the intelligibility of the speech of severely hearing-impared speakers" (p. 845), called the Speech Intelligibility Evaluation (SPINE). The full protocol as described requires pretesting rehearsal with the examiner, although the test phase itself does not appear to require an experienced listener. During testing, children are shown cards on which are printed one item of a four-item set (e.g., feel, fill, fail, fell). The child is asked to say the word on the card, and the listener must decide which of the four words in the set the child has said; the procedure is thus a four-alternative closed-set task. The entire test consists of 10 such sets; some are true minimal sets (as the one cited above), whereas others are overlapping minimal pairs (e.g., ten, den, ton, done). To validate the SPINE, results were compared with results from a write-down procedure using inexperienced listeners. Testing and scoring of the latter protocol were as in Monsen (1978). The SPINE was administered to 42 hearing-impaired children, including 34 classified as profoundly deaf and 8 as severely deaf. Additionally, the same children produced sentences, which were audiotape-recorded and played for 15 listeners. Scores on the SPINE test, which was admnistered by the author, ranged from 43% correct to 93% correct (mean = 77.9%, median = 79.0%), and scores from the write-down procedure ranged from 27% to 100% (mean = 78.7%, median = 82.3%); correlation (Pearson's product-moment) between the two sets of scores was $r = +.86$.

A study by McGarr (1983) compared the intelligibility of deaf speech to experienced and inexperienced listeners. Twenty profoundly deaf (group mean PTA = 98.6 dB) children, aged 8-10 years and 13-15 years, participated in the study. Test materials were 36 monosyllabic words taken from Smith (1975). The children produced these words both in isolation and embedded in sentences; productions were audio-recorded for subsequent hearing by listeners. In a third listening condition, the words in sentences were excised and presented to listeners as isolated words. Listeners were 60 experienced and 60 inexperienced listeners, the latter defined as persons "with no previous experience in hearing the speech of the deaf" (p. 452). For words in sentential contexts, listeners wrote down the entire sentence, although only the test words were used in scoring. For words produced and heard in sentences, the mean percent correct score for experienced listeners was 41% and for inexperienced listeners 30%. For test words produced and heard in isolation, the mean percent correct score was 29% for experienced listeners and 23% for inexperienced listeners. For words both in sentences and in isolation, differences in scores between experienced and inexperienced listeners were statistically significant.

Monsen (1983) examined a number of variables relevant to the study of speech intelligibility in hearing-impaired talkers: presence or absence of a verbal context, auditory-only or audio-visual presentation, number of presentations, grammatical complexity, and listener experience. Subjects in the study were 10 hearing-impaired adolescents, ages 11;7 (years; months) to 15;3. Eight of the subjects had PTAs worse than 95 dB (mean = 104 dB); PTAs for the remaining two subjects were 83 and 88 dB. Subjects recorded 160 sentences on both audio- and videotape, which were later played to both experienced and inexperienced listeners. Listeners were instructed to "write down as much of each sentence as could be understood" (p. 289), and scoring was based on a total of 100 points for each sentence, individual words being assigned points according to frequency of occurrence in the language (cf. Monsen, 1978). Across all listeners and talkers, percent correct was 79%. Across all talkers, experienced listeners scored 84% correct, whereas inexperienced listeners scored 74% correct; this difference was statistically significant ($p < .05$).

Becker, Schildhammer, and Ruoß (1994) examined speech intelligibility in 23 children (ages 9;6 [years;months] to 10;6) and adolescents (ages 12;6 to 15;4). Eighteen of the subjects had PTA thresholds above 90 dB, and five had PTAs between 85 and 90 dB. Speech materials consisted of descriptions of a picture story, which were audiotape-recorded and later played for two groups of listeners. Experienced listeners were six project personnel, and inexperienced listeners were 48 university students. The recordings of the stories were first transcribed by the experienced listeners and later by the inexperienced listeners. Because the specific speech material produced could vary from talker to talker, the transcriptions from the experienced listeners were considered to be "correct"; transcriptions by the inexperienced listeners were subsequently compared to these to generate outcome measures of intelligibility. Results compared percent correct words as a function of talker age and presence/absence of context. Additionally, results were categorized according to both type of word involved (noun, main verb, function word) and type of response (correct, substitution with same-class word, substitution with different-class word, unintelligible). Results indicated that the naive listeners understood more if there was a context. In general, the adolescents were more intelligible than the children. For example, mean percent correct nouns was 40% for the adolescents and 27% for the children; this difference was statistically significant (F (1, 31) = 4.55, $p < .05$).

Various authors (e.g., Gold, 1980; Osberger, Maso, & Sam, 1993) have cited 20% intelligibility as a consistent finding in the literature on the speech intelligiblity of children with hearing impairment. This estimate is based on, among others, Brannon (1964), John and Howarth (1965), Markides (1970), and Smith (1972, 1975), all of whom used inexperienced listeners. Other studies, also using inexperienced listeners, have reported slightly higher intelligibility (e.g., Becker et al., 1994; Maassen & Povel, 1985; McGarr, 1983). The by-now classic figure of 29% from Hudgins and Numbers (1942) is nevertheless for a study involving experienced listeners. The largest discrepancy in the literature, however, is the one existing between Monsen (1978, 1981, 1983) and just about everyone else. Monsen (1978) notes this discrepancy, citing three studies (John & Howarth, 1965; Markides, 1970; Smith, 1973) as representative of the common result of approximately 20% intelligibility. In the case of Smith (1973), for instance, intelligibility scores ranged from 0% to 76.1%, with a mean of 18.7%. The average in Monsen (1978) across severely and profoundly hearing impaired subjects was 76.7%, that is, more than the maximum score from Smith (1973).

Monsen (1978) further notes "marked differences in the subjects' ages, hearing levels, the recording techniques, materials spoken, listeners, scoring techniques, and so forth" (p. 215). Because of differences from study to study in such basic characteristics as independent and dependent variables, he says that "in one sense, the notion of an average intelligibility figure for such speakers is rather meaningless" (p. 215), yet notes further that the concept of "average intelligibility" can be useful for specific purposes. In attempting to explain the discrepancy between his own work (Monsen, 1978, specifically, but also the later reports, Monsen, 1981, 1983) and that of others, Monsen concentrates on differences in the sentences produced by the subjects. He notes that the sentences in Smith (1973) were more than twice as long on average (10.5 vs. 4.5 syllables) as those used in Monsen (1978), contained more polysyllabic words, and were syntactically more complex. In Markides (1970), there were no standard sentences (children described five pictures), and in John and Howarth (1965), speech materials consisted of children's spontaneous utterances. In the present study, both BIT and Monsen sentences (see below, Methods) were used, which contain on average 4.5 and 5.2 syllables per sentence, respectively.

As indicated above, the communication mode employed by children with hearing impairment is a variable with potentially important effects on the development of intelligible speech. "Oral communication" (OC) educational programs emphasize speech and auditory skill development. All children use hearing aids, and the use of manual signs is not encouraged. "Total communication" (TC) programs use a simultaneous combination of speech and signs. Although much of the manual lexicon is

borrowed from American Sign Language (ASL), the manual language used in TC programs encodes the words, morphology, and syntax used orally, which is why it is called "signed English". Studies of the oral communication skills of children in OC and TC programs have yielded strikingly contradictory results. Extensive literature reviews (Caccamise, Hatfield & Brewer, 1978; Wilbur, 1979) conclude that the use of TC is not detrimental to the development of speech skills. Indeed, Caccamise et al. propose that the use of TC may indeed facilitate the development of speech skills. On the other hand, other studies found a significant advantage in speech intelligibility for children enrolled in OC programs. A comparison of the performance of adolescents with hearing impairment from oral and total communication education settings on the SPINE (Monsen, 1981) was reported in Geers and Moog (1992). This study is of particular interest because the data were analyzed according to two different independent variables: communication mode (i.e., oral or total communication) and degree of hearing loss. The sample consisted of 227 16- and 17-year-olds with PTAs greater than or equal to 80 dB HL, including 100 subjects educated in oral programs and 127 subjects in total communication programs; the latter group included 64 subjects with deaf parents (TC-DP) and 63 with normally hearing parents (TC-HP). Additionally all subjects were divided into four groups according to best binaural PTA threshold: 80-90 dB HL, 91-100 dB HL, 101-110 dB HL, and >110 dB HL. For subjects in oral communication programs, group mean SPINE scores were as follows: 80-90 dB: 93%, 91-100 dB: 87%, 101-110 dB: 83%, >110 dB: 73%. The group mean score for TC-DP students with thresholds in the 80-90 dB range was 69%; for TC-HP students in the same threshold range, the group mean score was 77%. All other TC subjects had mean scores below 60%. Post-hoc comparisons of means showed a significant advantage for students in oral programs over those in total communication programs, and no difference between the two total communication groups.

There are two factors that complicate the interpretation of some of these studies and that may explain some of the discrepancies among them. Firstly, the listeners may be aware of the child's status and this knowledge may influence the outcome of the study. This is particularly important when the dependent variable is a rating determined by a listener who has a strong belief in a particular hypothesis under study, for example, the superiority of total communication over oral communication training methods, or vice versa. To avoid this problem, in the present study we used naive listeners who were not aware of the speaker's communication mode or amount of residual hearing. Likewise, the experimenters who conducted the listening sessions were unaware of these details. Second, the choice of total communication or oral communication for a particular child is not an entirely random process. In some parts of the country only one type of program may be available, with the result that geography plays an important role in the selection of communication mode for a deaf child, and becomes a confounding factor. Additionally, children may be steered towards one type of program depending on their aural/oral skills as perceived by their clinicians. The use of TC may be recommended more often in the case of children with poorer aural/oral skills, to ensure that at least some linguistic communication occurs, via the use of sign. In addition, children who start using one type of approach may switch to a different one depending on their success (or lack thereof) in developing aural/oral skills. Children who perform very poorly in OC programs may switch to TC programs and, conversely, children in TC programs who develop very good speech skills may switch to OC. This may introduce a bias in any study that compares the skills of randomly selected children in OC and TC programs, even when the investigators attempt to control for potentially confounding parameters. The most satisfactory way to address this problem would be to study children assigned to OC or TC programs in a prospective, randomized way. However, this would be very costly, impractical, and many clinicians would contend that such a study would lack clinical equipoise, making it ethically questionable. Thus, like those who have preceded us, we have not attempted to randomize assignment of the children under study to OC or TC mode: we will simply be cautious when interpreting measured differences in speech intelligibility between the two groups. In summary, the goal of this study is to assess the speech intelligibility of children with profound deafness, either congenital or acquired before the age of three, as a function of three independent variables: age, residual hearing, and communication mode.

Table I.

Total number of subjects and total number of data points (in parentheses) in each subject group.

	TC	ORAL	TOTAL
HA $_{90\text{-}100}$	10 (21)	11 (28)	21 (49)
HA$_{100\text{-}110}$	23 (47)	18 (30)	41 (77)
HA $_{110+}$	41 (64)	23 (29)	64 (93)

Methods

Subjects

We tested 126 profoundly and prelingually deaf children, who were divided into three groups according to their residual hearing, following the classification proposed by Osberger et al. (1993). This classification is based on unaided hearing losses at three frequencies: 500, 1000 and 2000 Hz. $HA_{90\text{-}100}$ is the group of hearing aid users with most residual hearing among the profoundly deaf: they have unaided losses of 90 to 100 dB HL (inclusive). The $HA_{100\text{-}110}$ group includes children with average unaided losses greater than 100 dB HL but less than (or equal to) 110 dB HL. The HA_{110+} group had the least residual hearing, with losses greater than 110 dB HL. All these children were potential candidates for cochlear implantation, and they were tested as part of a longitudinal study of cochlear implant users. Subjects were classified as users of Oral Communication (i.e., their therapy and formal education takes place without the use of signs) or Total Communication, which is the simultaneous use of signs and oral speech. All subjects were in educational programs that emphasized the development of oral skills. Table I shows some characteristics of each group of subjects. Subjects were tested between 1 and 5 times. When a subject was tested more than once, the testing sessions were at least six months apart.

Procedures

Each subject produced 10 sentences. Children under 6 were administered one list from the BIT (Beginner's Intelligibility Test), which uses objects and pictures to convey the target sentence, and an imitative response was elicited after an examiner's spoken model (Osberger, Robbins, Todd, & Riley, 1994). Older children (>6 y.o.) who could read were given the Monsen Sentences Test (Monsen, 1983). Sentences were recorded on cassette tape and digitized. The sentences were played back, in random order, to panels of three listeners with no experience listening to deaf speech The listening sessions were conducted in a double-blind fashion; that is, neither the listeners nor the experimenter knew the communication mode or the amount of residual hearing of the subjects being tested. Listeners heard more than one set of sentences, but these sentences were never from the same list or produced by the same talker. Following Monsen's procedure, each sentence was presented twice and no contextual information was provided (i.e., listeners had no information about the topics related to each sentence). The listeners transcribed what they heard, and their responses were tabulated to determine the percentage of transcribed keywords that were the intended target. Scores for the three judges were then averaged.

Data analysis

The goal of this study was to assess the effect of residual hearing and communication mode on the speech intelligibility of children who have profound, prelingual deafness and who do not use cochlear

implants. The effect of residual hearing was analyzed separately for the OC and the TC group, and the effect of communication mode was analyzed separately for each one of the three groups sorted according to residual hearing: $HA_{90\text{-}100}$, $HA_{100\text{-}110}$, and HA_{110+}. To assess the effect of communication mode on speech intelligibility, multivariate nonlinear regressions were done for each one of the three HA groups using the formula:

$$\text{INTELLIGIBILITY} = a_0 + a_1 \times \text{AGE} + a_2 \times \text{MODE} + a_3 \times \text{AGE} \times \text{MODE} \quad [1]$$

MODE refers to communication mode (coded as a discrete variable, with 0 indicating the use of Oral Communication and 1 indicating the use of Total Communication); AGE is the chronological age of each subject when his/her speech was recorded, and INTELLIGIBILITY is the mean intelligibility measured from each recording as explained above. The best-fit regression line for users of Oral Communication (i.e., MODE=0) as a function of age was

$$\text{INTELLIGIBILITY} = a_0 + a_1 \times \text{AGE} \quad [2]$$

and the best-fit regression line for users of Total Communication (i.e., MODE=1) was

$$\text{INTELLIGIBILITY} = (a_0 + a_2) + (a_1 + a_3) \times \text{AGE} \quad [3]$$

When neither a_2 nor a_3 were statistically significant for a particular HA group, this indicated that the regression lines for users of OC and TC were not significantly different. In other words, two separate regression lines (one for OC users and another one for TC users) did not fit the data any better than a single regression line When a_2, a_3, or both were statistically significant, the best-fit lines for the OC and TC groups were examined to see which group had an advantage in speech intelligibility.

The analysis of the effect of residual hearing was conducted in a similar fashion, and it was done separately for the OC and TC groups. Multivariate nonlinear regressions were done using this formula:

$$\text{INTELLIGIBILITY} = (a \times \text{AGE} + b) + (c \times \text{AGE} \times \text{PTA1}) + (d \times \text{AGE} \times \text{PTA 2}) + (e \times \text{PTA1}) + (f \times \text{PTA2}) \quad [4]$$

PTA1 and PTA2 are discrete variables that code the subject's PTA. For group $HA_{90\text{-}100}$, both PTA1 and PTA2 are 1; for group $HA_{100\text{-}110}$ PTA1 is 0 and PTA2 is 1; and for group HA_{110+} both variables are 0. The meaning of the variables AGE and INTELLIGIBILITY has already been explained. The best-fit regression lines for members of each residual hearing group were:

$$\text{INTELLIGIBILITY} = a \times \text{AGE} + b \text{ (for the } HA_{90\text{-}100} \text{ group)}, \quad [5]$$

$$\text{INTELLIGIBILITY} = (a + d) \times \text{AGE} + (b + f) \text{ (for the } HA_{100\text{-}110} \text{ group), and} \quad [6]$$

$$\text{INTELLIGIBILITY} = (a + c + d) \times \text{AGE} + (b + e + f) \text{ (for the } HA_{110+} \text{ group).} \quad [7]$$

The significance of the different regression parameters was examined to determine whether the data were better fit with separate regression lines for each subgroup than with a single line. All the regression analyses were conducted twice: once with all the data points, and another time using only the first data point for each subject (i.e., using a strictly cross-sectional data set). Finally, all the data points were plotted in a single graph for visual inspection of overall trends.

Results

Figure 1 shows speech intelligibility scores as a function of age at testing for children with the least amount of residual hearing: those in the HA_{110+} group. Intelligibility scores were quite low regardless of age. The slopes of the regression lines were very shallow, indicating only small differences as a function of age: 1.2 %/year for TC users and 1.5 %/year for OC users. However, there were significant correlations between speech intelligibility and age, $r = +.64$ for the OC users and $r = +.63$ for TC users ($p < .001$ in both cases). The a_2 and a_3 parameters in the multiple regression were not significant ($p = .77$ and .41, respectively), indicating that the intelligibility data for children in the HA_{110+} group was not fit any better by two regression lines (one for OC users, one for TC users) than by a single line.

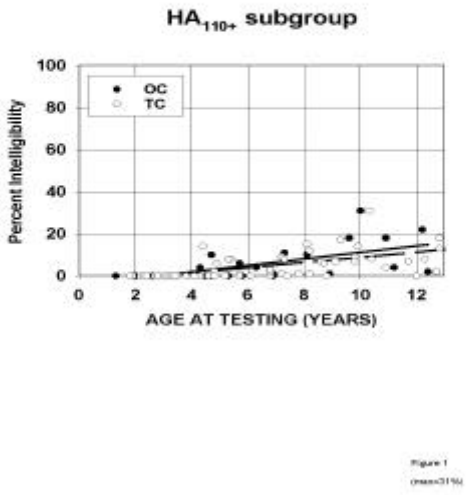


Figure 1. Percent intelligibility as a function of age at testing, for HA_{110+} subjects (PTA>110 dB HL). The black circles represent users of Oral Communication and the black line is the corresponding regression line. Total Communication users are represented by white circles and the corresponding regression line is dashed.

Results were more positive in the $HA_{90\text{-}100}$ group, which included the children with most residual hearing (always within the profoundly hearing impaired range) As Figure 2 shows, there is a very clear trend towards higher intelligibility scores at older ages, both for OC and TC users. The slopes of the regression lines are quite steep, 10.4%/year for the TC group and 7.6%/year for the OC group. The difference between these two slopes was not statistically significant (p=.11 for the a_3 parameter), but the difference between the intercepts of the regression lines for OC and TC users in this $HA_{90\text{-}100}$ group was significant ($p < .001$ for the a_2 parameter). Correlations between speech intelligibility and age were significant ($r = +.81$ for the OC group and $r = +.88$ for the TC group, $p < .001$ in both cases). To the extent that the regression lines are representative of the data, they point to an advantage for the OC group, particularly at the younger ages. All children older than 9.5 years in this sample achieved intelligibility scores higher than 64%. This is a level that would probably allow most of them to have reasonably fluent face-to-face conversations, when the listener can see their face as they talk. In summary, there is a stark contrast between children in the $HA_{90\text{-}100}$ group, who have a relatively good prognosis for their ability to learn how to speak intelligibly (provided that they receive appropriate training) and the children in the HA_{110+} group, whose intelligibility levels are uniformly low.

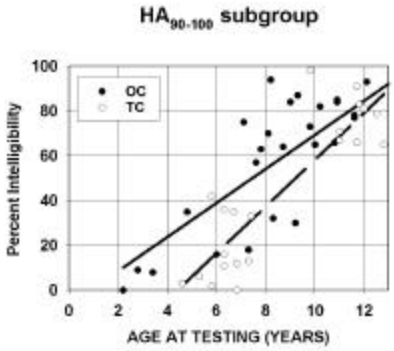


Figure 2. Percent intelligibility as a function of age at testing, for $HA_{90\text{-}100}$ subjects (90PTA100 dB HL). The black circles represent users of Oral Communication and the black line is the corresponding regression line. Total Communication users are represented by white circles and the corresponding regression line is dashed.

Results for the group with intermediate amounts of residual hearing, $HA_{100\text{-}110}$, are shown in Figure 3. In this group there was a dramatic difference between OC and TC users: the slopes of the regression lines were 7.8%/year for the OC group and only 0.9%/year for the TC group. The coefficients that were significant in the regression were a_2 (p = .002) and a_3 (p < .001), indicating that the speech intelligibility of OC users in this subgroup is significantly different from (and substantially superior to) that of TC users. The regression line for OC users in this group was only slightly lower than the regression lines of OC or TC users in the $HA_{90\text{-}100}$ group, and the correlation between intelligibility and age at testing was just as high as for those two subgroups (r = +.86, p < .001). In contrast, the same correlation for TC users in this group was very low (r = +.16) and not significant (p = .30). It is quite apparent that, in contrast to the scores from the other two groups, intelligibility scores for children in the HA100-110 group follow a bimodal distribution. Indeed, scores obtained at ages greater than 9 years old (that is, to the right of the vertical dashed line in Fig. 3), are either greater than 35% or lower than 10%.

Reinforcing this result, Figure 4 shows that there is a trend toward a bimodal distribution for the whole sample of children tested in this study, not just for those in the $HA_{100\text{-}110}$ group. The figure shows the particular nature of the data set by joining with thick lines all the data points corresponding to a given subject over time. Some subjects, tested only once, are represented by only one symbol. Some of the children show clearly higher intelligibility scores at higher ages, achieving 50% to 100 % intelligibility by age 10. This group of more successful speakers includes all in the $HA_{90\text{-}100}$ group; most OC users and a few TC users in the $HA_{100\text{-}110}$ group, and no members of the HA_{110+} group. Other children show very poor intelligibility scores regardless of age and in spite of all the speech training that they receive. Their scores rarely exceed 30% and are typically lower than 20%. These children include all members of the HA_{110+} group; a few OC users and most TC users in the $HA_{100\text{-}110}$ group, and no members of the HA_{110+} group.

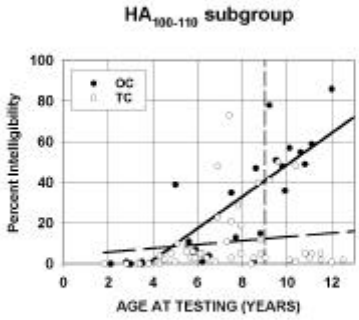


Figure 3. Percent intelligibility as a function of age at testing, for $HA_{100\text{-}110}$ subjects (100 < PTA110 dB HL). The black circles represent users of Oral Communication and the black line is the corresponding regression line. Total Communication users are represented by white circles and the corresponding regression line is dashed. The vertical dashed line at age 9 helps visualize the bimodal distribution observed at older ages in this subject group.

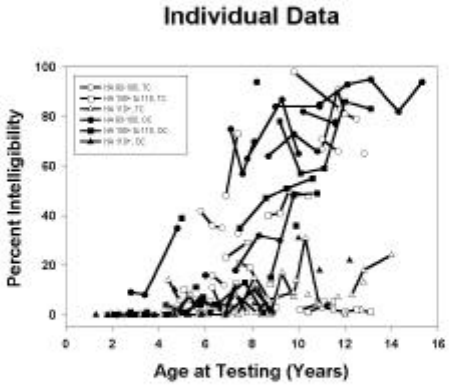


Figure 4. All the individual data points for OC and TC users within each subgroup. Successive data points for the same subject are connected by solid lines.

The careful reader will not be surprised to read that for both the OC and the TC groups, data were significantly better fit by three regression lines (one for each residual hearing subgroup) than by a single line[3]. In both groups, there was a significant trend toward higher intelligibility scores for those subgroups with greater amounts of residual hearing.

Discussion

The data show two strong overall trends, neither of which should be surprising: children with more residual hearing tend to be more intelligible, and some subgroups of OC users tend to be more intelligible than their TC-using counterparts. Although this second conclusion should be tempered due to the considerations discussed in the introduction (namely, that assignment to the OC or TC methods is not a random process), it may still be quite enlightening to ponder the differences between OC and TC users in each one of the subgroups. These differences were quite dramatic in the $HA_{100\text{-}110}$ subgroup, much less so in the $HA_{90\text{-}100}$ subgroup, and nonexistent in the HA_{110+} subgroup. It might be argued that a floor effect decreased the power of the comparison in this last subgroup, but it is still clear that the intelligibility levels of all children in the subgroup, OC and TC users alike, are functionally quite low and may not allow easy communication with naive listeners. Remember that a 20% intelligibility level means that a naive listener would understand one word out of every five uttered by the speaker. Even in a situation with high levels of contextual information, a listener would have a hard time communicating when only one word out of five is identified reliably. The low levels of intelligibility in this subgroup raise the question whether the many hours that these children have spent in speech training will pay off. It is important to remember that this consideration applies only to children with PTAs greater than 110 dB and who have not received cochlear implants, because cochlear implants typically increase the child's auditory skills to a level where he or she might derive much more benefit from oral rehabilitation than the typical child in the HA_{110+} subgroup who uses hearing aids.

The $HA_{90\text{-}100}$ subgroup is, in a way, the mirror image of the HA_{110+} subgroup. Again, the differences between OC and TC users are not overwhelming, but unlike in the HA_{110+} subgroup, this happens because both OC and TC users show much improved speech intelligibility at older ages. It is in the $HA_{100\text{-}110}$ subgroup that the average differences between OC and TC users are truly remarkable. Perhaps there is a particular range of hearing impairment where an intensive concentration on oral rehabilitation results in maximum payoff. The present results suggest that such a range may be in the vicinity of 100 to 110 dB HL, at least from the point of view of development of intelligible speech.

In contrast to the intricate pattern of differences between OC and TC users found in the present study, Geers and Moog (1992) found a clear advantage for OC users over TC users for each subgroup of children, regardless of the amount of residual hearing. However, there are a number of differences between the two studies that make a direct comparison of the results difficult. Firstly, Geers and Moog used the SPINE test instead of the write down procedure employed in this study. One consequence of using SPINE is that the listeners in that study were aware of the child's identity and communication mode (i.e., the listeners were not "blind" to the independent variable of communication mode). Secondly, The Geers and Moog study included speakers who were 16 or 17 years old, much older than the subjects in this study. The age difference might explain some of the differences between the two studies. For example, it's possible that the OC users in the HA_{110+} group in our study might outperform the speech intelligibility levels of their TC counterparts in future years. However, other differences between the two studies are more difficult to explain. In particular, most of the older TC subjects in our $HA_{90\text{-}100}$ subgroup

[3] The coefficients that were statistically significant in the OC regression were: d ($p < .001$), e ($p = .031$) and f ($p = .004$). The coefficient a was marginally nonsignificant at $p = .057$. In the TC regression, the significant coefficients were: a ($p = .010$), c ($p < .001$), and e ($p < .001$).

(those who were 9 to 12 years old) scored higher than Geers and Moog's TC subjects who had the same amount of residual hearing. This happened even though the Geers and Moog subjects were older and the listeners in that study had an easier task than in ours, namely, they only had to select one word out of four possible ones instead of writing down what the speaker said. One possibility is that the TC subjects in both studies were not similar. For example, our TC users may have received more oral training than those in the Geers and Moog study.

Finally, and even though the study of children with cochlear implants is not a focus of the present study, it must be said that children who have received implants before the age of six and who have used state-of-the-art devices and stimulation strategies since initial stimulation seem to perform at least as well as their hearing aid using peers in the $HA_{90\text{-}100}$ group, from the viewpoints of speech perception (Svirsky & Meyer, 1999; Meyer & Svirsky, in press), speech production (Svirsky et al., in press a) and language development (Svirsky, in press; Svirsky et al. in press b). Thus, the prognosis for oral communication for children who use cochlear implant technology may be quite more optimistic than those for any of the subgroups that are analyzed in this study. It is our hope that the present study will serve as a benchmark to refine our comparisons of speech intelligibility by cochlear implant and hearing aid users. Such comparisons, together with parallel studies of speech perception and language development, will continue to strengthen the knowledge base that we draw upon when we need to make clinical decisions about pediatric cochlear implantation.

References

Becker, R., Schildhammer, A., & Ruoß, M. (1994). Sprachliche Leistungen von gehörlosen Kindern und Jugendlichen beim mündlichen und schriftlichen Erzählen einer Bildergeschichte. *Zeitschrift für experimentelle und angewandte Psychologie*, 61, 349-377.

Brannon, J. B. (1964). *Visual feedback of glossal motions and its influence upon the speech of deaf children.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Caccamise, F., Hatfield, N., & Brewer, L. (1978) Manual/simultaneous communication (M/SC) research: results and implications. *American Annals of the Deaf* (Silver Spring, MD). 123(7):803-23.

Cowan, R.S.C., Brown, C.J., Whitford, L.A., et al. (1995). Speech perception in children using the advanced SPEAK speech-processing strategy. *Annals of Otology, Rhinology, & Laryngology, 104*, 318-321.

Fairbanks, G. (1960). *Voice and articulation drillbook*. (Second ed.). New York: Harper & Row.

Geers, A. E., & Moog, J. S. (1992). Speech perception and production skills of students with impaired hearing from oral and total communication education settings. *Journal of Speech and Hearing Research*, 35, 1384-1393.

Gold, T. (1980). Speech production in hearing-impaired children. *Journal of Communication Disorders*, 13, 397-418.

Guilder, R. P., & Hopkins, L. A. (1936). The importance of auditory functions studies in the educational program for the auditorily handicapped child. *The Volta Review*, 38, 69-74, 149-155.

Hippocrates. (1853). *Œuvres complètes d'Hippocrate* (Vol. 8) (E. Littré, Ed. and Trans.). Paris: J. B. Baillière.

Hirsch, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders*, 17, 321-337.

Hudgins, C. V., & Numbers, F. C. (1942). An investigation of the intelligibility of the speech of the deaf. *Genetic Psychology Monographs*, 25, 289-392.

John, J. E. J., & Howarth, J. N. (1965). The effect of time distortions on the intelligibility of deaf children's speech. *Language and Speech*, 8, 127-134.

Maasen, B., & Povel, D.-J. (1985). The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. *Journal of the Acoustical Society of America*, 78, 877-886.

Markides, A. (1970). The speech of deaf and partially-hearing children with special reference to factors affecting intelligibility. *British Journal of Disorders of Communication*, 5, 126-140.

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

Metz, D. E., Schiavetti, N., & Sitler, R. W. (1980). Toward an objective description of the dependent and independent variables associated with intelligibility assessments of hearing-impaired adults. In J. D. Subtelny (Ed.), *Speech assessment and speech improvement for the hearing impaired* (pp. 72-81). Washington, DC: Alexander Graham Bell Association for the Deaf.

Meyer, T.A. & Svirsky, M.A. (In press). Speech perception by children with the Clarion (CIS) or Nucleus (SPEAK) cochlear implant or hearing aids. *Annals of Otology, Rhinology, and Laryngology*.

Miyamoto, R.T., Kirk, K.I., Robbins, A.M., Todd, S.L. & Riley, A.I. (1996). Speech perception and speech production skills of children with multichannel cochlear implants. *Acta Oto-Laryngologica, 116*, 240-243.

Monsen, R. B. (1978). Toward measuring how well hearing-impaired children can speak. *Journal of Speech and Hearing Research*, 21, 197-219.

Monsen, R. B. (1981). A usable test for the speech intelligibility of deaf talkers. *American Annals of the Deaf*, 126, 845-852.

Monsen, R. B. (1983). The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech and Hearing Disorders*, 48, 286-296.

Osberger, M. J., & Levitt, H. (1979). The effect of timing errors on the intelligibility of deaf children's speech. *Journal of the Acoustical Society of America*, 66, 1316-1324.

Osberger, M. J., Maso, M., & Sam, L. K. (1993). Speech intelligibility of children with cochlear implants, tactile aids, or hearing aids. *Journal of Speech and Hearing Research*, 36, 186-203.

Osberger, M. J., Robbins, A. M., Todd, S. L,, & Riley, A. I. (1994). Speech intelligibility of children with cochlear implants. *The Volta Review*, 96, 169-180.

Osberger, M.J., Robbins, A.M. Todd, S.L., Riley, A.I., & Miyamoto, R.T. (1994). Speech production skills of children with multichannel cochlear implants. In I.J. Hockmair-Desoyer & E.S. Hochmair (Eds.), *Advances in cochlear implants* (pp. 503-508). Vienna: Manz.

Robbins, A.M., Kirk, K.I., Osberger, M.J. & Ertmer, D.J. (1995). Speech intelligibility of implanted children. *Annals of Otology, Rhinology, and Laryngology, 104*, 399-401.

Samar, V. J., & Metz, D. E. (1988). Criterion validity of speech intelligibility rating-scale procedures for the hearing-impaired population. *Journal of Speech and Hearing Research*, 31, 307-316.

Skinner, M.W., Clark, G.M., Whitford, L.A. et al., (1994). Evaluation of a new spectral peak (SPEAK) strategy for the Nucleus 22 channel cochlear implant system. *American Journal of Otology, 15*, 15-27.

Smith, C. R. (1972). *Residual hearing and speech production in deaf children*. Unpublished doctoral dissertation, City University of New York, New York, NY.

Smith, C.R. (1973). Residual hearing and speech production in deaf children. *CUNY Communications Science Laboratory Research Report, Volume 4*.

Smith, C. R. (1975). Residual hearing and speech production in deaf children. *Journal of Speech and Hearing Research*, 18, 795-811.

Smith, C. R. (1982). Differences between read and spontaneous speech of deaf children. *Journal of the Acoustical Society of America*, 72, 1304-1306.

Subtelny, J. D. (1977). Assessment of speech with implications for training. In F. H. Bess (Ed.), *Childhood deafness: Causation, assessment, and management* (pp. 183-194). New York: Grune & Stratton.

Svirsky, M.A., & Meyer, T.A.. (1999). A comparison of speech perception for pediatric Clarion and hearing aid users. *Annals of Otology, Rhinology, and Laryngology*, 108 (Suppl 177), 104-110.

Svirsky, M.A., Sloan, R.B., Caldwell, M., & Miyamoto, R.T. (In press a). Speech intelligibility of prelingually deaf children with multichannel cochlear implants. *Annals of Otology, Rhinology, and Laryngology*.

Svirsky, M.A., Robbins, A.M., Kirk, K.I., Pisoni, D.B., & Miyamoto, R.T. (In press b). Language development in profoundly deaf children with cochlear implants. *Psychological Science*.

Svirsky, M.A. (In press). Language development in children with profound and prelingual hearing loss, without cochlear implants. *Annals of Otology, Rhinology, and Laryngology*.

Wilbur, R.B. (1979). *American Sign Language and sign systems*. Baltimore, MD: University Park Press.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Eliciting Speech Reduction in the Laboratory II: Calibrating Cognitive Loads for Individual Talkers[1]

James D. Harnsberger and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH-NIDCD Training Grant DC00012 and NIH-NIDCD Research Grant R01-DC00111 to Indiana University. We would like to thank Jim Brink for his technical assistance and Richard Wright for his help and advice.

Eliciting Speech Reduction in the Laboratory II: Calibrating the Cognitive Load for Individual Talkers

Abstract. This experiment extended work done previously in our laboratory to develop a method to elicit from talkers three different speaking styles, reduced, citation, and hyperarticulated, using controlled materials in a laboratory setting. In the initial experiment, the reduced style was elicited by having subjects read a sentence while carrying out a distractor task that involved recalling a fixed number of digits from short-term memory. The original experiment was clearly limited in its success at eliciting a reduced style of speech: Only one of the six talkers showed significant differences between reduced and citation speech based on an acoustic analysis of the sentences. In this study, we chose to calibrate the distractor memory task to an individual's short-term memory span as measured by a simple digit span task. That is, the number of digits to be recalled after reading aloud a test sentence was determined by each individual's digit span. Twelve talkers were recorded in this experiment. The results showed that six of the twelve talkers produced a reduced style of speech for the test sentences in the distractor task relative to the same sentences in the citation style condition, as determined by a phonetically-trained judge. This initial evaluation was confirmed in a perceptual test using a pairwise comparison task in which normal-hearing, untrained listeners were presented with two sentences varying in speech style and were asked to choose the most carefully pronounced sentence. The results showed that 71% - 88% of the sentence sets tested were correctly differentiated by speech style, indicating that the individual calibration method was a substantial improvement over the elicitation method of the original experiment.

Introduction

Traditionally in studies of speech production and perception that use natural speech, utterances are recorded under highly controlled conditions in a laboratory setting. Control over the recording conditions and the nature of the materials recorded (particular syllables, words, sentences) serves to limit sources of error in the data collection process, or to avoid particular confounds that might render the results uninterpretable. Control over the quality and structure of the materials also insures that an experiment can be replicated in other laboratories, a key aspect of any experiment. However, it has long been recognized that the style of speaking elicited from talkers reading linguistic material aloud in a laboratory setting differs systematically from more reduced styles of speech that can be observed in unmonitored conversations outside of the laboratory (Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988; Picheny, Durlach, & Braida, 1989; Byrd, 1994). These differences can include the duration of the utterance and its constituent words, pausing, and the degree of centralization in the quality of vowels, to name a few. Such differences pose a problem for theories of speech perception and spoken word recognition, most of which were formulated from studies using controlled speech materials: To what extent do these findings generalize to the speech styles that people produce and perceive in a natural setting? The perception of variability that exists among speech styles has not been studied in detail, no doubt due to the problem of eliciting naturalistic speech in the decidedly unnatural manner and setting of reading aloud in a laboratory. Other types of “nonlinguistic” variability have proven to have an effect on speech perception and spoken word recognition, including talker, rate, and stimulus variability (Mullennix & Pisoni, 1990; Nygaard, Sommers, & Pisoni, 1995; Bradlow, Nygaard, & Pisoni, 1999). These studies suggest that listeners encode in long-term memory significant details and properties of speech signals that they encounter, and that these details

influence the subsequent perception and recognition of speech. If listeners are sensitive to detailed, episodic properties of speech, then variation in those properties due to speaking style differences may also play an important role in speech processing, one that has thus far been neglected in speech perception and spoken word recognition research.

One factor limiting the study of the perception of speech styles has been a methodological one: how does a researcher elicit different speech styles, including more reduced and naturalistic speech, while controlling for the particular syllables, words, or sentences to be studied? Recording natural conversation, or guided conversation on a particular topic, has been used in the study of sociolinguistic variability in speech production, namely, in the elicitation of stigmatized, or less prestigious, sounds, words, or syntactic structures of a dialect (Labov, 1972; Milroy, 1987). In other methods, subjects have been asked to participate in and narrate a task (Hirschberg & Nakatani, 1996). Such procedures have been useful in eliciting particular intonational forms and in studying particular aspects of discourse structure (Swerts & Collier, 1992; Speer, Sokol, & Schafer, 1999). However, none of these methods can guarantee the elicitation of specific sentences.

In our laboratory, we have attempted to develop methods for eliciting sentences in different speech styles in the laboratory while controlling for the particular sentence materials used. The range of speech styles being studied includes a reduced, or hypoarticulated style, that should more closely resemble the speech style employed in natural settings in conversation than laboratory read speech. The first version of this method was described by Brink, Wright, and Pisoni (1998). Brink et al. (1998) attempted to elicit three speaking styles, namely, reduced, or hypoarticulated speech; citation, or read speech, a style that is normally used in reading controlled materials in a laboratory setting; and hyperarticulated speech. Each style was elicited in a separate condition of the experiment.

Brink et al. (1998) attempted to elicit reduced speech by having subjects read a sentence while engaging in a concurrent task, specifically, remembering a digit sequence of five to seven digits in length that was presented immediately prior to the sentence. After reading the sentence, subjects were asked to recall the digit sequence in the same order in which it was presented. The digit span task was a distractor task, chosen to place the subject under a cognitive load while reading a sentence. The digit span task was chosen as the distractor task because, in piloting, it was successful in producing the desired speech style while minimizing talker disfluencies. Citation speech was elicited by simply having listeners read single sentences presented on a computer screen. Hyperarticulated speech was elicited in an experimental condition quite similar to the Citation speech condition. Subjects were asked to read single sentences presented on a computer screen. Over the course of the condition, they were prompted in a subset of trials to repeat the sentence “more clearly.” After responding to that prompt, subjects were given the same prompt a second time, and the second reading was chosen to represent hyperarticulated speech. This procedure had been used successfully in an earlier study by Johnson, Flemming, and Wright (1993).

Brink et al. (1998) tested this method with six talkers, all native speakers of English, and evaluated its success in a detailed acoustic analysis, examining properties of the sentence, as well as key words in the sentence, in terms of duration, f0 range, absolute RMS energy, energy range, degree of vowel centralization, and degree of vowel dispersion. The results of the acoustic analysis showed that the method was successful in eliciting a hyperarticulated speech style that was highly distinct from the citation style, a result that was found for all six talkers. The duration, vowel centralization, and vowel dispersion measures showed the most consistent differences across talkers in speech production. However, the method failed to elicit significant differences between the reduced and citation sentences for five of the six talkers. Only one talker, MD, produced reduced speech that was distinguishable from citation speech in the acoustic

analysis.[2] Interestingly, MD also had the highest error rate for correctly recalling digit sequences during the Reduced condition, indicating that the fixed load of five to seven digits was sufficiently challenging for MD. For the other five talkers, digit sequences of a length of five to seven may not have been sufficiently demanding for them, particularly given that the average immediate memory span for digits is 7.7 (Cavanagh, 1972). Individual differences in digit span may complicate the use of a fixed range of digit sequence lengths since, for a given individual, the degree of cognitive load that a subject is placed under may either be too great or little, depending on the subject's own digit span. If the load is too great, then the possibility exists that a subject will simply ignore the distractor task, making the Reduced condition in essence the same as the Citation condition. If the distractor task is too easy, a subject may be able to allocate sufficient attentional resources to the task of reading the sentence, a result that also effectively eliminates any differences between reduced and citation speech. One solution to the problem of using fixed sequence lengths would be to measure an individual's digit span and use that value to "calibrate" the Reduced condition, making the task difficult enough to draw attentional resources away from reading, but not so difficult so as to be ignored or induce disfluencies.

The goal of this study was to test an individually-calibrated cognitive load method of eliciting reduced speech. In the experiments reported here, the cognitive load task was calibrated to the individual talker's memory span in a forward digit span task administered prior to the Reduced condition. A talker's digit span, as measured in this task, was the value used to determine the range in the length of digit sequences in the Reduced condition of this study. The Citation and Hyperarticulation conditions of Brink et al. (1998) remained unchanged in this study. With the addition of individual calibration, we predicted that all of the talkers recorded would produce a reduced speech style that was perceptually distinct from the citation and hyperarticulated speech styles. Twelve talkers were recorded in this revised procedure for the Reduced condition, and in the original Citation and Hyperarticulation conditions. The elicited sentences were then evaluated by a phonetically-trained judge and by 25 untrained listeners to determine if the individual calibration method was a success.

EXPERIMENT 1

Methods

Participants

Twelve native speakers of American English, seven females and five males ranging in age between 18 and 30, participated in this study. Participants received $15 total compensation for participating in two one-hour sessions. None of the subjects reported any history of speech or hearing disorders at the time of testing.

Stimulus Materials

The participants read 34 sentences from the 200 sentences comprising the SPIN set (Kalikow, Stevens, & Elliot, 1977). The SPIN sentences are short sentences, five to eight words in length, ending in a high frequency monosyllabic noun. The 34 SPIN sentences selected for this study are listed in Appendix A. The recording took place in a sound-attenuated chamber (IAC Audiometric Testing Room, Model 402)

[2] Speaker MD's reduced sentences were also perceptually distinguishable from his/her citation sentences in a pilot Paired Comparison task with three native speakers of English. These native speakers successfully picked the citation sentences as "more carefully pronounced" in reduced-citation sentence pairs, on 89% of test trials. For a detailed description of the Paired Comparison task, see Experiment 2 for a study using the same methodology.

using a head-mounted Shure (SM98) microphone positioned one inch away from the subject's chin. The recordings were digitized at 22.05 kHz (16 bit sampling) using a Tucker-Davis Technologies System II and stored on an IBM-PC 486 computer.

Procedures

The participants carried out four tasks over two test sessions. In the first session, participants were administered a simple forward digit span task (see Digit Span Task) and were recorded reading sentences in the Reduced condition. In the second session, which took place within seven days of the first session, participants were recorded reading sentences in the Citation and Hyperarticulation conditions.

Digit Span Task. In the digit span task, participants were presented with a sequence of single digits (0 - 9) on a computer screen inside of the sound-attenuated chamber, and asked to recall the sequence correctly in the order in which it was presented. The participants' responses were digitized and played via headphones to the experimenter, who sat outside of the booth and scored the responses. The responses themselves were not stored to disk as sound files. The length of the digit sequence that was presented started at four, and then increased or decreased via an adaptive staircase algorithm (Levitt, 1971). The algorithm increased the sequence length by one digit for every two sequences at a given length that were successfully recalled by the participant. Whenever the participant responded to a sequence incorrectly, the sequence length was reduced by one digit on the following trial. Over the course of the 25 trials of the task, the sequence length for individual participants increased until the sequence length began eliciting errors. Thus, by the end of the task, participants were “oscillating” between the sequence length that they could consistently recall, and a longer sequence that induced errors. The longest sequence length that was consistently recalled was taken to be the participant's digit span. This value was then used to calibrate the cognitive load in the Reduced condition.

Reduced Condition. The Reduced condition was similar to the Reduced condition described by Brink et al. (1998), and consisted of 136 trials, four trials for each of the 34 SPIN sentences, with a 1 s inter-trial interval (ITI). The order of the blocks of four trials varied randomly for each participant. Each trial consisted of four parts: initially, participants were presented with a digit sequence, which remained on the screen for 2 s; then, after a 2.5 s interval, a sentence was displayed on the computer screen for the participant to read; next, the participant's response was recorded over a 6 s window; finally, participants were prompted to recall the digit sequence in the correct order. The length of the digit sequence was based on the participant's digit span as measured in the Digit Span Task. The length of the digit sequence in a given trial was either the same as the span score, or plus/minus one digit. For example, if a participant had a span of seven in the digit span task, he/she would be presented with digit sequences ranging in length from six to eight. The same sentence, embedded in the digit span task, was presented four times, with the fourth reading taken as the reduced sentence for subsequent analysis. Before the recording began for the Reduced condition, participants were told that they would be participating in a short-term memory experiment. Participants were instructed to focus on the digit span task in the Reduced condition, in the hope that they would be less careful in monitoring their production of the test sentences.

Citation and Hyperarticulation Conditions. The Citation and Hyperarticulation conditions were identical to those described earlier by Brink et al. (1998). In the Citation condition, participants were prompted to read aloud a sentence that appeared on the computer screen. Each sentence was presented once, for a total of 34 trials, with a 1 s ITI. The order in which the sentences were presented was randomized for each participant.

The Hyperarticulation condition was similar to the Citation condition, and consisted of two types of trials. The first trial type, the "citation cycle," was identical to a Citation condition trial. In the second trial type, the "hyperarticulation cycle," participants were also prompted to read aloud a sentence appearing on the computer screen. After reading this sentence, participants were then prompted to "Please read the sentence more clearly." After responding, they were asked again to read the sentence more clearly. Thus, for the hyperarticulation cycle, the same sentence was read three times, with the third reading taken to be the example of the "hyperarticulated" reading of the sentence for subsequent analysis. The 34 sentences each appeared in three citation cycles and one hyperarticulation cycle. The program controlling the experiment was designed to insure that the Hyperarticulation condition began with a citation cycle, and that hyperarticulation cycles were separated by at least two citation cycles.

Results and Discussion

In an earlier test of this method of eliciting different speaking styles (Brink et al., 1998), the elicited sentences were initially evaluated using a detailed acoustic analysis of those cues that have been commonly cited in prior work as important ones in differentiating speech styles. The cues measured by Brink et al. included the duration, RMS energy, and energy range of each sentence and of three "key" words in each sentence (usually the subject, verb, and object of each sentence); the sentence f0 range; the degree of vowel centralization of vowels in key words, and the degree of vowel dispersion of vowels in key words. The results of this lengthy acoustic analysis revealed no significant differences between the reduced and citation sentences of five of the six talkers recorded. Given this failure and the fact that the present experiment represents an "exploratory" phase in the development of a method of eliciting speech styles in the laboratory, a less time-consuming method of evaluating the results was chosen, namely, an impressionistic evaluation of the sentences by a single phonetically-trained judge. The results of this evaluation appear in Table 1. Each percentage in the three sentence pair columns represents the percentage of sentence pairs judged to be qualitatively different in terms of speaking style[3].

Subject	Reduced-Citation	Reduced-Hyperarticulated	Citation-Hyperarticulated
1	67%	100%	100%
2	88%	100%	100%
3	91%	100%	100%
4	6%	100%	100%
5	24%	100%	100%
6	41%	100%	100%
7	18%	100%	100%
8	41%	100%	100%
9	76%	100%	100%
10	41%	100%	100%
11	50%	100%	100%
12	82%	100%	100%

Table 1: The percentage of sentence pairs judged to be qualitatively different in speech style.

[3] The number of sentence pairs that each percentage represented varied slightly due to the fact that individual subjects occasionally produced disfluencies in their readings of a particular sentence. Sentence pairs were excluded from evaluation in cases in which one of the sentences involved a disfluency or disfluencies.

An examination of the impressionistic results shows that the sentence pairs that included hyperarticulated sentences were clearly differentiable, just as they were in the acoustic analysis of Brink et al. (1998). The critical sentence pairs for this study were the Reduced-Citation pairs, given the failure of the earlier method of Brink et al. to elicit measurable differences between these two styles. In the individual calibration method of this study, only half of the participants (1, 2, 3, 9, 11, and 12) produced qualitative differences in 50% or more of their hypoarticulated and citation sentence pairs. The percentage of pairs judged to be different varied widely by individual talker, from as low as 6% to as high as 91%. Thus, the method of individually calibrating the cognitive load of the Reduced condition was effective for a subset of the talkers tested. This result is clearly an improvement over the method of Brink et al. that successfully elicited reduced sentences from only one participant out of six.

EXPERIMENT 2

Methods

Participants

Twenty-five native speakers of American English, seventeen females and eight males ranging in age between 18 and 21, participated in this study. For participating in a single one-hour session, the participants received either $7.50 or one credit towards their research requirement if they were enrolled in an undergraduate psychology class. None of the subjects reported any history of speech or hearing disorders at the time of testing.

Stimulus Materials

The stimulus materials consisted of 26 to 34[4] hypoarticulated, citation, and hyperarticulated sentences from the four talkers, namely subjects 2, 3, 9, and 12 from Experiment 1, whose reduced-citation sentence pairs were most frequently judged to be qualitatively different in Experiment 1.

Procedures

A trial in the Paired Comparison Task consisted of two different readings of each sentence from each talker. The two readings were presented in pairs, with a 1 s interstimulus interval. Participants were asked to choose which sentence was read more carefully by using a mouse to press on one of two buttons on a computer screen, denoting the first or the last sentence of the pair. The sentence pairs differed only in terms of the speaking style in which they were produced, resulting in three types of pairs: reduced-citation, reduced-hyperarticulated, and citation-hyperarticulated. The sentence pairs always involved the same sentence produced by the same talker. An example of a Paired Comparison trial would be a reduced “The farmer harvested the crop,” produced by Talker 2, paired with a hyperarticulated “The farmer harvested the crop,” also produced by Talker 2. The 25 participants were divided into four groups of three to eight participants each. Each group listened to the sentence pairs of a single talker. The sentence pairs were presented in both orders (i.e., citation-hyperarticulated and hyperarticulated-citation). Thus, each participant responded to 156 to 204 trials, depending on which talker they were randomly assigned to.

[4] The full set of 34 sentences were not used for each talker because, in a limited number of cases, some talkers produced sentences with disfluencies.

Results

The results of the Paired Comparison Task appear in Table 2, which lists the percentage of sentences judged correctly for the three types of sentence pairs for each talker. In this table, "Sentences" refers to the number of different SPIN sentences from each talker, while "Listeners" refers to the number of participants that judged the sentence pairs of a given talker. As expected, listeners correctly chose the hyperarticulated sentence as the one that was read "more carefully" in a high percentage of sentence pairs, 85% - 100% of trials, indicating that the method was successful in eliciting citation and hyperarticulated sentences. In the critical test pairs, the Reduced-Citation pairs, percent correct scores were the same or slightly lower than those of the phonetically-trained judge in Experiment 1, ranging between 71% and 88%. Overall, the percent correct scores in the Paired Comparison Task showed a significant correlation with the corresponding percentages in Table 1 ($r = 0.77$, $p \leq 0.05$), confirming the judgments of the phonetically-trained listener.

Talker	Sentences	Listeners	Reduced-Citation	Reduced-Hyperarticulated	Citation-Hyperarticulated
2	31	8	88%	100%	99%
3	34	7	78%	93%	85%
9	30	7	71%	86%	96%
12	26	3	73%	96%	98%

Table 2: The percentage of sentence pairs judged correctly by the untrained listeners.

General Discussion and Conclusions

The results of Experiments 1 and 2 showed that the individual calibration method of eliciting reduced speech, in conjunction with the Citation and Hyperarticulated conditions, was successful in producing three distinct speech styles from half of the subjects, in a large majority of those subjects' sentences (71 - 88%). However, the results did not match the predicted success rate because half of the subjects did not produce reduced sentences that were perceptually distinct from their citation sentences. One possible reason for this failure may be the restrictive criteria used for selecting reduced and citation sentences. The individual participants' success rates were based on comparisons of one example, out of two to four elicited, of each sentence in each speech style. For instance, the fourth sentence in a block of four in the Reduced condition was taken to be the "reduced" example of that sentence. Each sentence also has three citation readings, one from the Citation condition and two from the Hyperarticulated condition. This strategy of sentence selection may have unduly biased the outcome by restricting the capacity of the method to elicit different speech styles. The other readings of each sentence in each style condition may have been equally good, or better, representatives of each style. For instance, instead of comparing only the fourth reading of a sentence in the Reduced condition to the first reading of a sentence in the Citation condition, the third Reduced sentence could be compared to the second Citation sentence, to see if discernible differences exist between the two. A preliminary examination of all of the readings of each style of each sentence by the phonetically-trained judge from Experiment 1 indicates that in the case of one talker, Participant 6, the percentage of his/her perceptually distinct Reduced-Citation pairs increased to the range of talkers 1, 2, 3, 9, 11, and 12 when other readings are considered (see Table 1). Thus, relaxing the

criteria of which sentence readings are chosen to represent reduced- and citation-style sentences may give a more accurate analysis of the relative success of the method.

Even using relaxed criteria, however, one third of the subjects still failed to produce a Reduced-Citation distinction. Moreover, the success rate across sentences for participants that were judged as producers of the distinction was less than perfect. Apparently, other factors are at work in determining whether or not a sufficient cognitive load will induce reduction in the sentences of a given participant. One possible factor may be differences among individuals in terms of their performance on the Digit Span Task versus their ability to correctly recall digits in the Reduced condition. Individuals may differ in their capacity to recall digits while engaging in a distractor task that would not be reflected in their performance on a simple digit span task. Such individual differences would be reflected in the percentage of digit sequences correctly recalled in the Reduced condition. Table 3 lists these percentages for the twelve participants, who are grouped in terms of their digit span.

Participant	% Correct	Span
10	62	5
12	38	6
11	60	6
1	68	6
2	75	6
5	78	6
6	21	7
7	41	7
3	42	7
8	53	7
9	26	8
4	32	8

Table 3. The percentage of trials in which digit sequences were correctly recalled in the reading portion of the Reduced condition.

As Table 3 shows, participants who had performed the same on the simple digit span task were not equivalent in their ability to recall digit sequences in the Reduced condition. Participants with a span of six ranged from 38% - 78% correct, averaging 67% correct. Participants with spans of 7 also varied widely, and participants with spans of 7 and 8 generally did worse in recalling digit sequences with a distractor sentence than participants with spans of 5 or 6. The implications of these differences lie in the issue of cognitive load and its capacity to induce reduction in sentences. The calibration of the cognitive load is crucial to the method's success. If the cognitive load is too great, participants may ignore the load task and simply read the sentence in citation style. If the load is not sufficiently demanding of attentional resources, then participants may find it too easy to both read the distractor sentence and recall a digit sequence, rendering the distractor sentence a citation, rather than a reduced, sentence. In this experiment, some participants may have been highly successful in recalling digit sequences in the simple digit span task. However, when asked to recall digit sequences in the context of an intervening reading task, their effective span may have been much lower than that indicated in the simple digit span task, making the calibrated load too difficult to induce reduction. Thus, participants may differ in terms of their capacity to cope with a distractor task in recalling digit sequences. If this is indeed a possible source of error in the experimental

method, one solution may be to redesign the Digit Span Task to match the Reduced condition. In future work, we intend to test such a revised method, with a greater number of subjects representing each span length (i.e., 5, 6, 7, and 8) to attempt to calculate the necessary correlations to support this hypothesis.

Of course, it is possible that small changes in the Reduced condition will still fail to produce the desired reduced-citation style difference for all talkers. This method may not represent the optimal solution to the problem of eliciting reduced speech in a controlled experiment. Ultimately, the success of the method outlined here, or any variant of it, must be judged not only in simple perceptual experiments, such as the Paired Comparison task, which simply indicates that different speech styles were elicited on a "carefulness" continuum. The acoustic properties of the elicited sentences must also be measured, and their differences correlated with those differences measured in studies of speech styles in natural speech. Only if the reduced sentences elicited by the method described here display the properties of naturally-occurring reduced speech can the method be judged a success. Thus, acoustic analysis, following successful perception tests, should form the basis of method evaluation in our subsequent work.

In summary, our study successfully tested a revised method to elicit three different speaking styles, reduced, citation, and hyperarticulated, by calibrating the task for eliciting reduced speech to the individual participant's short-term memory span. Relative to the original experiment, the calibration method was a substantial improvement, eliciting reduced sentences in 71 – 88% of test trials from half of the talkers recorded as determined by a phonetically-trained judge as well as a group of naïve listeners. Using more relaxed criteria for selecting "reduced" sentences out of the set of four recorded from each talker, the success rate of the revised method increased to seven of the twelve talkers tested. Of course, this new scoring method still resulted in five talkers from whom reduced sentences were not successfully elicited. To improve on the current method, a new individual calibration technique was proposed, involving a digit sequence distractor task that is continuously calibrated over the course of the Reduced condition.

References

Bradlow, A.R., Nygaard, L.C., & Pisoni, D.B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics, 61,* 206-219.

Brink, J., Wright, R., & Pisoni, D.B. (1998). Eliciting speech reduction in the laboratory: Assessment of a new experimental method. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 396-420). Bloomington, IN: Speech Research Laboratory, Indiana University.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15,* 39-54.

Cavanagh, J.B. (1972). Relation between the immediate memory span and the memory search rate. *Psychological Review, 79,* 525-530.

Hirschberg, J. & Nakatani, C.H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. ACL-96.

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.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language, 69*, 505-528.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*, 467-477.

Milroy, L. (1987). *Observing and analyzing natural language*. Oxford: Basil Blackwell.

Mullennix, J.W. & Pisoni, D.B. (1990). Stimulus variability and processing dependenices in speech perception. *Perception and Psychophysics, 61*, 206-219.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception and Psychophysics, 57*, 989-1001.

Picheny, M.A., Durlach, N.I., & Braida, L.D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research, 32*, 600-603.

Speer, S.R., Sokol, S.B., & Schafer, A.J. (1999). Prosodic disambiguation of syntactic ambiguity in discourse context. *Journal of the Acoustical Society of America, 106*, 2275.

Swerts, M. & Collier, R. (1992). On the controlled elicitation of spontaneous speech. *Speech Communication, 11*, 463-468.

Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., & Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America, 84*, 917-928.

Appendix 1: Stimulus Sentences

The farmer harvested his crop.
His boss made him work like a slave.
He caught the fish in his net.
Close the window to stop the draft.
The beer drinkers raised their mugs.
I made the phone call from a booth.
The cut on his knee formed a scab.
The railroad train ran off the track.
They drank a whole bottle of gin.
The airplane dropped a bomb.
I gave her a kiss and a hug.
The soup was served in a bowl.
The cookies were kept in a jar.

How did your car get that dent?
The baby slept in his crib.
The cop wore a bullet-proof vest.
No one was injured in the crash.
The hockey player scored a goal.
How long can you hold your breath?
At breakfast he drank some juice.
The king wore a golden crown.
He got drunk in the local bar.
The doctor prescribed the drug.
The landlord raised the rent.
Playing checkers can be fun.
Throw out all this useless junk.

Her entry should win first prize.
The stale bread was covered with mold.
I ate a piece of chocolate fudge.
The story had a clever plot.

He's employed by a large firm.
The mouse was caught in the trap.
I've got a cold and a sore throat.
The judge is sitting on the bench.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Effects of Multimodal Presentation and Lexical Density on Immediate Memory Span for Spoken Words[1]

Lorin Lachs, Winston D. Goh,[2] and David B. Pisoni[3]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by a grant from the NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington. Special thanks go to Luis Hernàndez, Tyler Emley and Patrick Kelley for their invaluable assistance during the completion of this study.
[2] Also, Department of Social Work & Psychology, National University of Singapore.
[3] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Effects of Multimodal Presentation and Lexical Density on Immediate Memory Span for Spoken Words

Abstract. Working memory span for spoken words was measured using a non-repeated sampling procedure. Stimuli were presented in audio-only (AO) or audiovisual (AV) conditions. In the AO condition, participants were presented with only the audio track of a talker's voice. In the AV condition, participants simultaneously heard and saw a talker speaking a word. The stimulus words differed in their neighborhood density, neighborhood frequency and word frequency. The results show that audiovisual presentation enhances memory span for words from sparse lexical neighborhoods, but has no effect on words from dense lexical neighborhoods. The findings are discussed with respect to the effects of perceptual distinctiveness on working memory.

Visual information about the articulation of spoken words has been shown to have large effects on speech perception. Conflicting information about speech in the visual and auditory modalities can lead to the illusory perception of speech sounds not included in either modality alone (commonly referred to as the "McGurk effect;" McGurk & MacDonald, 1976). The McGurk illusion is commonly elicited by presenting the syllable "ba" in the audio track and the syllable "ga" in the video track of a cross-dubbed movie clip. In the overwhelming majority of cases, participants report that the syllable presented was "da". The very existence of this phenomenon demonstrates that some sort of integration of information in the auditory and visual domains occurs during the process of speech perception. The precise nature of this integration, however, remains in question.

The McGurk effect has generated a great deal of interest in the idea of audiovisual integration. Much of the research on the integration issue has concerned the effects of conflicting audio and visual cues on segmental identification. This work has produced a large body of knowledge concerning the ways in which acoustic and visual information influence each other. For example, it has been shown that the vowel context in which a segment is presented affects the degree to which visual information influences the McGurk illusion (Green, 1996; Green, Kuhl, & Meltzoff, 1988), that inverted faces reduce the effects of McGurk integration (Massaro & Cohen, 1996; Jordan & Bevan, 1997) and that temporal asynchrony between the audio and visual tracks has no effect on McGurk integration (Massaro, Cohen, & Smeele, 1996; Munhall, Gribble, Sacco, & Ward, 1996; Smeele, Sittig, & van Heuven, 1992). Even separating the spatial location of the auditory and visual aspects of the stimulus makes little difference on the extent of the illusion (Bertelson, Vroomen, Wiegeraad, & de Gelder, 1993; Fisher & Pylyshyn, 1994; Jones & Munhall, 1997).

Other studies have shown that the auditory and visual information in a bimodal stimulus is integrated during processing, such that the information in each channel is evaluated relative to the information present in the other channel. For example, Green and Kuhl (1989) showed that the perceived boundary in voice onset time along an /ibi/ to /ipi/ continuum was dependent on whether or not there was concurrent visual information. Presentation of the visual portion of the word caused the VOT boundary to shift as though it were along an /idi/ to /iti/ continuum, precisely the continuum specified by the McGurk effect. This finding demonstrated that low-level auditory cues are evaluated in the context of the *combined* audiovisual stimulus, suggesting that integration of the information in the various modalities happens before classification. A related study showed interference in the processing of one sensory modality based on variation in the other (Green & Kuhl, 1991), providing further support for the notion that the information in both channels is processed in tandem.

Most of this knowledge about audiovisual integration, however, has been accumulated using tasks that measure segmental identification. With the exception of Dekle, Fowler and Funnel's (1992) study of audiovisual integration in real words, the stimulus materials used in all of the McGurk effect experiments have been nonsense syllables. In order to fully understand how the phenomena associated with audiovisual speech stimuli are related to speech perception in general, it is necessary to expand investigations into more naturalistic settings. The work of Sumby and Pollack (1954) showed that audiovisual speech effects are not confined to illusory percepts and segmental contexts. On the contrary, their study demonstrated that the intelligibility of words can be enhanced by as much as +15 dB in noisy environments, a gain that surpasses even the best hearing aid devices.

The use of audiovisual information in spoken word recognition is a relatively unexplored area, but some work has already been carried out. In spoken word recognition, it is hypothesized that bottom-up perceptual processes interact with multiple, higher sources of information during processing (e.g. Luce & Pisoni, 1998; McClelland & Elman, 1986). An idea central to the concept of spoken word recognition is the mental lexicon, an entity in long term memory that stores information about all the words ever encountered by a specific individual (see Lively, Pisoni, & Goldinger, 1994). The structure of lexical representations in long-term memory has profound effects on the process of spoken word recognition. For example, the frequency of occurrence in the language of a target word is directly related to its intelligibility, while the number of words similar to the target word ("neighborhood density") is inversely related to intelligibility (Luce & Pisoni, 1998). Thus, words are recognized in the context of other phonetically similar words in the lexicon.

Some evidence exists to support the notion that long-term memory representations for spoken words exist in a multimodal form. Lachs and Pisoni (submitted) found that the repetition of studied, dynamic visual information of a talker's face during test facilitated performance on a recognition memory task. The results were taken to imply that information about the dynamic aspects of visual articulation is stored in lexical memory for spoken words.

If audiovisual information is represented in the long-term memory representations of spoken words, and the structure of those representations is important during the process of spoken word recognition, then lexical structure should have an effect on audiovisual speech perception. Indeed, Brancazio (1999) found that varying the lexical properties of the auditory and visual parts of McGurk stimuli had effects on the extent to which those stimuli were susceptible to illusory percepts. Auer and Bernstein (1997) have investigated the computational properties of the mental lexicon when words are represented using visemes. Because visemes are sets of phonetic segments considered indistinguishable in visual-only environments, transcribing the lexicon in this way is presumed to be equivalent to collapsing across perceptual dimensions that are irrelevant to the process of spoken word recognition while speechreading. The structure left in these viseme-transcribed lexicons remains useful because many of the words in the lexicon remain unique, even when the number of segments is reduced by 75%. Although it has not yet been explored experimentally, this structure could be a source of additional information in audiovisual situations.

Cognitive process that may interact with speech perception processes is working memory. It is well known that serial recall of verbal material appears to be affected by the phonological properties of the material to be remembered, as well as other experimental manipulations on the acoustic environment of the experiment (cf. Baddeley, 1998). Phenomena such as the *phonological similarity effect*, (Baddeley, 1966; Conrad, 1964), the *word length effect* (Baddeley, Thomson, & Buchanan, 1975), the *unattended speech effect* (Salame & Baddeley, 1982), and the *articulatory suppression effect* (Baddeley et al., 1975) suggest that the representation of verbal material in working memory is in a phonological or speech-based code. Baddeley and Hitch's (1974) well-known model of working memory specifically posits a

phonological loop to handle the encoding and rehearsal of verbal material. The phonological loop comprises 2 subsystems, a phonological store and an articulatory rehearsal loop. The passive phonological store is a temporary, limited capacity buffer that can hold auditory memory traces. These traces begin to decay after about 2 seconds if they are not rehearsed by the articulatory loop, which serves to maintain the information in the store. However, the precise mechanisms underlying the phonological loop and the nature of the relationship to the phenomena mentioned above is still much debated (e.g., Bavelier & Potter, 1992; Cowan, Wood, Nugent, & Treisman, 1997; Nairne, Neath, & Serra, 1997).

Recent investigations have also indicated that the structure of LTM representation affects working memory (e.g., Bourassa & Besner, 1994; Goh & Pisoni, 1998; Hulme, Maughan, & Brown, 1991). These studies have begun to describe in detail the interaction between long-term properties of the lexicon and short-term memory processes (e.g., Gathercole, Frankish, Pickering, & Peaker, 1999; Hulme, Roodenrys, Schweickert, Brown, Martin, & Stuart, 1997; Schweickert, 1993). In general, this research has demonstrated that the information stored in the mental lexicon can facilitate the recall of verbal information in working memory. Memory span for nonwords or words of an unfamiliar language is lower than memory span for familiar words (Hulme et al., 1991).

Memory span is also affected by the semantic attributes of words (Bourassa & Besner, 1994), phonotactic properties (Gathercole et al., 1999), and word frequency (Hulme et al., 1997). Lexical status, phonotactic information, and word frequency are all assumed to be stored with the memory trace of the word in the mental lexicon. Using the framework of Schweickert's (1993) multinomial processing tree model of immediate recall, one can argue that such properties of the long-term traces of words can be used to aid in the redintegration of decayed traces in working memory, and thus facilitate their subsequent retrieval and recall. For example, items that do not have a lexical representation, such as nonwords, would not have the additional advantage of permanent traces to aid in the recall process, which thus translates to a lower memory span.

Goh and Pisoni (1998) recently investigated the effects of phonological neighborhoods on immediate memory span. One way to represent and quantify the organization of phonological information in the mental lexicon is to consider the notion of lexical neighborhoods defined by their phonological similarity (Landuaer & Streeter, 1973; Treisman, 1878; Luce, 1986; Luce & Pisoni, 1998). A similarity neighborhood is defined by the number of words that can be obtained by a single substitution, addition, or deletion of a phoneme. Using this metric, the neighbors of *cat* would include *hat*, *cut*, *cap*, *scat* and *at*, among others. Two factors have been used to characterize the structure of a lexical neighborhood (Luce & Pisoni, 1998). *Neighborhood density* refers to the number of words in a similarity neighborhood, whereas *neighborhood frequency* refers to the average frequency of occurrence of the words in a given similarity neighborhood. A third property, *word frequency*, refers to an individual word's frequency of occurrence in the language.

Based on these three variables, words can be classified into two dichotomous categories based on ease of recognition. *Easy words* are words that are higher in frequency relative to their neighbors, and reside in low density and low frequency neighborhoods. *Hard words* are words that are lower in frequency relative to their neighbors, and reside in high density and high frequency neighborhoods. Previous work has shown that these lexical properties influence spoken word recognition. Lexically easy words are recognized faster and more accurately than lexically hard words (Luce, 1986; Luce & Pisoni, 1998).

Goldinger, Pisoni, and Logan (1991) found lexical effects in serial recall of 10-word lists–easy words were recalled better than hard words. Goh and Pisoni (1998) extended the Goldinger et al. (1991) findings by replicating the lexical effects using an immediate memory span task. Furthermore, they

showed that the difference in immediate recall between easy and hard word spans was not related to participants' working memory capacity, as measured by the traditional digit-span task. This finding indicated that the locus of the effect was most likely from the long-term properties of the lexicon, and not from individual differences in working memory capacity. More interestingly, Goh and Pisoni found that the lexical effect emerges only when a non-repeated sampling procedure was used, i.e., each word was used only once on a particular list and trial, and never repeated in subsequent lists or trials. No lexical effect was observed when a repeated sampling procedure was used, in which words were resampled again across trials. Again, this pattern suggested a contribution from long-term storage, because repeated sampling would keep the working memory traces active and thus severely attenuate the effects from the permanent store. Collectively, these findings suggest that phonological neighborhood properties and their ramifications on the representational distinctiveness of a word's long-term memory trace affect the serial recall performance of these words. Lexically easy words, which are perceptually more distinctive, will have a greater probability of redintegration from partially decayed traces in working memory compared to lexically hard words, which are less distinctive (cf. Schweickert, 1993).

If the structure of spoken word representations in memory has an effect on immediate recall, and long-term memory representations for spoken words contain multimodal information, then audiovisual presentation of stimuli should also produce effects on working memory. The present study manipulated the lexical characteristics of the stimulus words used in the memory task, and examined memory capacity in auditory-only (AO) and audiovisual (AV) settings. Because lexical similarity is defined across inherently multimodal dimensions (phonemes), 'extra' visual information will not significantly add to the distinctiveness of target candidates in memory, if those targets come from an already dense neighborhood (hard words). If targets are easy, however, then the extra visual information will make differences between target candidates even *more* distinct, thereby improving memory capacity for AV presentation over AO presentation.

Method

Design

The present experiment measured memory span in a 2 x 2 repeated-measures factorial design. The two levels of the "Lexical Class" factor used lexically "easy" and lexically "hard" words. The two levels of the "Presentation Mode" factor presented stimuli in "Audiovisual (AV)" and "Audio-Only (AO)" conditions. All factors were administered within subjects. The order in which conditions were presented to participants was counterbalanced using a balanced Latin-square design.

In order to assess and control for individual differences in short-term memory, a control condition was administered first to all participants. This condition tested memory span using spoken digits as the items for study. This was presented using auditory information only.

Participants

Participants were 34 undergraduate students enrolled in an introductory psychology course who received partial credit for their participation. One participant's data was eliminated from the final analysis for failure to follow the instructions provided. All of the participants were native speakers of English who reported no hearing or speech disorders at the time of testing. In addition, all participants reported having normal or corrected-to-normal vision.

Materials

Two Apple Macintosh computers, each equipped with a 17" Sony Trinitron Monitor (0.26 dot pitch) and its own video processing board were used to present the stimuli to subjects. The video processing boards were each capable of handling clips digitized at 30fps with a size of 640 x 480 and 24-bit resolution. The auditory portion of each stimulus was presented over Beyer Dynamic DT100 headphones.

The word stimuli were a subset of the tokens contained in the Hoosier Audiovisual Multitalker Database [HAVMD] (Lachs & Hernàndez, 1998; Sheffert, Lachs, & Hernàndez, 1997). The video portion of the HAVMD tokens is digitized at 30 fps with 24-bit resolution with 640 x 480 pixel size. The audio portion of the HAVMD tokens is digitized at 22 kHz with 16-bit resolution. The digitized movie clips of one talker (F1) uttering 264 isolated words were used during this study. The talker F1 was chosen because previous intelligibility studies on the HAVMD showed that she is the most intelligible talker in the database in audiovisual (AV) and audio-only (AO) conditions. Half of the words chosen were classified as lexically "Easy" words, and the other half were classified as lexically "Hard" words.

The spoken digits (0 to 9) were tokens obtained from the Texas Instruments 46-Word (TI46) Speaker-Dependent Isolated Word Corpus (Texas Instruments, 1991). The original tokens on the CD-ROM were in 11,025 Hz 16-bit PCM-Motorola formatted files. These files were edited using a digital waveform editor to remove the silent portions from either side of the token and saved as 12,500 Hz, 16-bit, mono WAV files. The overall RMS (root-mean-square) amplitude levels for each digit token were digitally equated with the word tokens to ensure equal presentation levels. The tokens recorded by a female speaker were used for the digit-span task. This speaker was chosen because her mean intelligibility was 100%, as determined by a token identification task with ten volunteer participants who were not a part of the current study.

Procedure

In order to measure memory span, lists of increasing length were presented to participants for immediate recall. Two lists of each length were presented in each condition. The shortest length used was three items, while the longest length used was eight items. All participants received all the list lengths, in ascending order, in all the conditions. In all, 66 stimulus words were presented in each condition.

One response booklet for each condition was provided so that participants could write down the items in each list. Participants were instructed not to start writing down their answers until the end of each list. The response sheet for each list contained eight blank spaces. The response for each list was recorded on separate pages in the response booklet. Participants were told that they had to recall the items in the order they were presented. If they were unable to recall a particular word, they were instructed to leave a blank space in the position where it occurred. The next list was not presented until the participants indicated they were ready.

Prior to the presentation of the first stimulus, participants were instructed in the procedure and encouraged to ask questions if they were unclear on their task. Stimuli were randomly presented by computer.

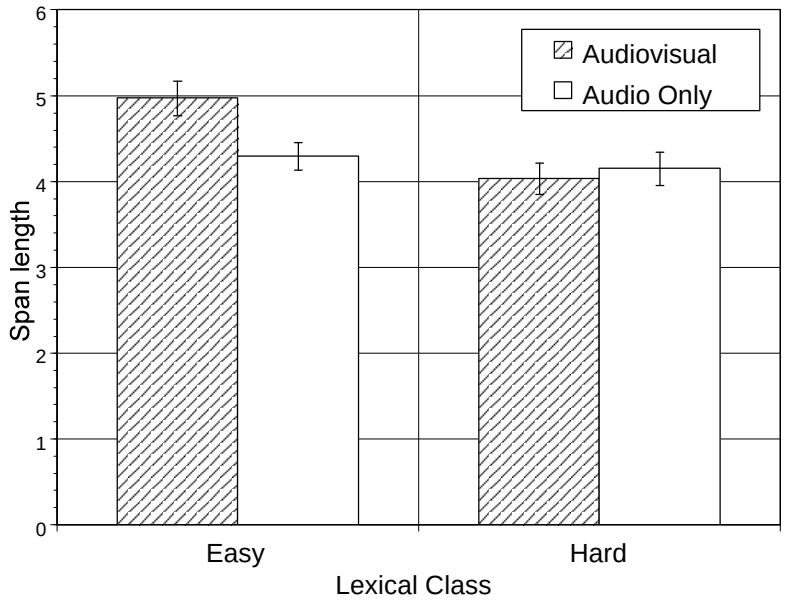


Figure 1. Average H/L score as a function of Lexical Class and Presentation Mode. Error bars denote standard errors.

Scoring

Each response list was scored independently. An item was scored as correct if, and only if, the item and serial order were both reproduced correctly. Phonetic transcriptions, homophones and obvious spelling errors (e.g., "cheif" for the target "chief") were counted as correct.

Two measures of memory span were computed. The "Highest/Longest" (H/L) score was the longest length at which all items were scored correct *on both lists of that length*. The "Strict" (S) score was computed by taking one less than the minimum list length and adding 0.5 for every list that was completely correct. These two scoring methods are referred to as "length-based" measures, since they score in terms of list-length correct. These measures have been traditionally used for indexing working memory capacity (see La Pointe & Engle, 1990).

Results

Figure 1 shows the "Highest/Longest" (H/L) scores in each condition, averaged across participants. The left panel displays performance on lists that used "Easy" words, and the right set of columns displays performance on lists that used "Hard" words. Within each set of columns, the white bar represents performance on lists presented in the AO condition, while the striped bar represents performance on lists presented in the AV condition.

A 2 (Lexical Class) x 2 (Presentation Mode) repeated measures analysis of variance (ANOVA) on the H scores revealed a main effect of Lexical Class, $F(1, 33) = 15.32$, $MSE = 0.66$, $p < .001$, $\eta^2 = 0.32$. Memory span for lexically easy words ($M = 4.63$, $SD = 0.15$) was higher than memory span for lexically hard words ($M = 4.09$, $SD = 0.14$). There was no main effect of Presentation Mode.

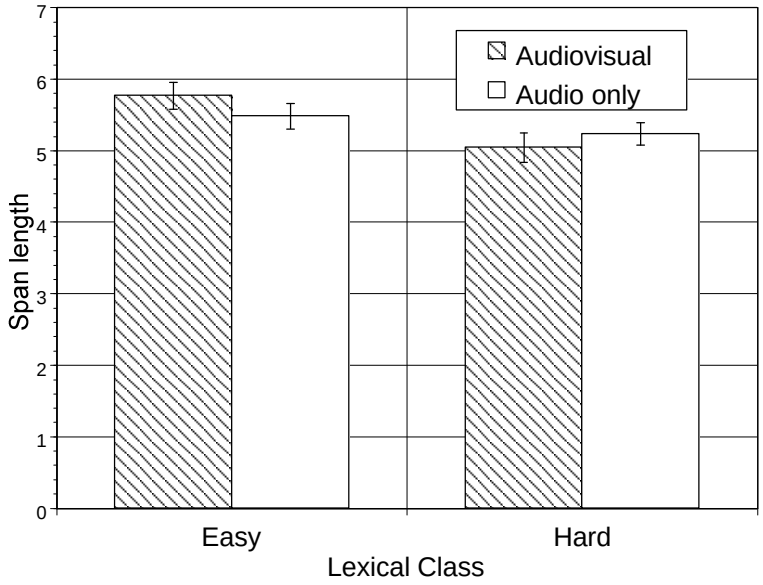


Figure 2. Average Strict (S) score as a function of Lexical Class and Presentation Mode. Error bars denote standard errors.

We also observed a significant interaction between Presentation Mode and Lexical Class, $F(1,33) = 5.30$, $MSE = 1.01$, $p < .05$, $h^2 = 0.14$. Simple effects analyses revealed that this interaction was supported by an effect of Presentation Mode within the "Easy" level of the Lexical Class factor, $F(1,33) = 9.98$, $MSE = 0.78$, $p < .01$. Performance in the Easy-AV condition was clearly better than performance in any of the other conditions. For lists made up of lexically easy words, memory span for lists presented audiovisually was better than memory span for lists presented audio-only. However, there were no differences in memory span as a function of Presentation Mode within the "Hard" level of Lexical Class, $F < 1$.

Figure 2 shows the "Strict" (S) scores for all conditions, averaged across participants. Clearly, the S scores show the same pattern of results evidenced by the H scores. A 2 (Lexical Class) x 2 (Presentation Mode) repeated measures ANOVA on the S scores again showed the main effect of Lexical Class, $F(1, 33) = 15.22$, $MSE = 0.53$, $p < .001$, $h^2 = 0.32$, with the memory span for lexically easy lists ($M = 5.63$, $SD = 0.16$) higher than those for hard lists ($M = 5.14$, $SD = 0.16$). Again, there was no main effect of Presentation Mode.

As with the H scores, we also observed a significant interaction between Presentation Mode and Lexical Class, $F(1,33) = 5.18$, $MSE = 0.36$, $p < .05$, $h^2 = 0.14$. Additional analyses revealed that this interaction was supported by a marginally significant effect of Presentation Mode within the Easy level of the Lexical Class factor, $F(1,33) = 2.86$, $MSE = 0.53$, $p = 0.10$. Again, audiovisual presentation facilitated memory for easy lists, relative to audio-only presentation. There was no effect of Presentation Mode for the Hard level of Lexical Class $F(1,33) = 1.11$, $MSE = 0.46$, n.s.). This means that for lists made up of lexically hard words, there was no difference in memory span due to differences in presentation mode.

Because it was important to determine whether our results replicated those of Goh and Pisoni (1998), the simple effect of Lexical Class within each level of Presentation Mode was analyzed for both the H and S scores. Our experiment would constitute a replication if we found an effect of Lexical Class in the Audio Only level of Lexical Class. The simple effect of Lexical Class within the Audio Only

condition is represented in both figures by comparing the white bar in the Easy panel with the white bar in the Hard panel. However, the simple effects analysis for H scores revealed that there was no effect of Lexical Class within the AO condition, $F < 1$. The analysis for S scores only revealed a marginal effect of Lexical Class within the AO condition, F (1,33) = 3.57, MSE = 0.30, $p < .07$. Thus, the present study did not replicate Goh and Pisoni (1998).

However, there were significant effects of Lexical Class in the AV condition (comparing the striped bars in each panel) for both sets of scores; H scores: F(1,33) = 19.92, MSE = 0.76, $p < .001$; S scores: F(1,33) = 14.9, MSE = 0.59, $p < .001$.

Discussion

While it is problematic that we did not replicate the results of Goh and Pisoni (1998), the marginal effect found using S scores indicates that running more subjects may reveal a difference in Lexical Class in the Audio-only condition. In fact, there are a couple of reasons to support the notion that increasing the N will actually lead to a difference. First, Goh and Pisoni used an N of 40 in their investigation. The present experiment only used an N of 33. Second, Goh and Pisoni only required participants to participate in two conditions. It may be that the additional two conditions in the current experiment served to lower memory spans in general, thereby reducing the size of differences between conditions.

Assuming that the lack of replication denotes a lack of power and not a fundamental flaw in the experiment, the results show that audiovisual presentation of spoken words improves memory span, but only when the words are lexically Easy. We propose that this advantage is due to differences in the ability to maintain the distinction between various list items in memory. Because Easy words are inherently more distinct from their phonological neighbors than Hard words, it is easier to maintain long lists of them in short-term memory (Goh & Pisoni, 1998). We speculate that this advantage may be due to less competition from similar sounding words. Retrieval cues for lists of easy words may be less affected by proactive interference effects from previous lists. Rehearsal processes may be more efficient because less neighbors would be activated during each rehearsal cycle compared to hard words.

Audiovisual presentation, however, interacts with the effects of lexical distinctiveness. Because a hard word has more neighbors, the chances of those neighbors possessing similar audio *and* visual qualities are very high. Thus, for hard words, “added” visual information does not add substantively to the distinctiveness of a target from its neighbors. However, for Easy words, the chance that the additional visual information will distinguish a target item from it's neighbors is greater, since there is less of a chance that those neighbors will share *both* auditory and visual characteristics with the target item.

Previous research has shown an effect of audiovisual presentation on working memory. Pichora-Fuller (1996) showed that presenting stimuli at high signal-to-noise ratios actually decreases working memory capacity. However, presenting the words audiovisually counteracted this pattern. Memory spans for audiovisual items presented at high signal-to-noise ratios were roughly equivalent to memory spans for audio-only words presented at lower signal-to-noise ratios. Pichora-Fuller's findings were interpreted as evidence for the reallocation of resources from working memory to the lexically based processes involved in spoken word recognition. When AO items were presented in noise, resources are deallocated from working memory processes and sent to the processes involved in lexical access. AV presentation, it was reasoned, allows for more facile access to lexical representations, and thus, the deallocation of resources from working memory is unnecessary.

Because the task used by Pichora-Fuller was very different from the one used in this study, it is not possible to draw comparisons between her findings and the current ones. Similarly, it is not possible to distinguish between the interpretation of that study's findings with the interpretation offered here. Pichora-Fuller's findings are based on perceptual, "external" noise's effects on memory span, while ours are based on the "internal" noise generated by lexical competitors. It is entirely possible that these two sources of noise have completely different interactions with working memory processes. Our "internal" noise explanation relies on the fact that words from dense neighborhoods will be less likely to become distinct from their neighbors when visual information is included in the signal. In contrast, "external" noise may obscure the acoustic dimensions of spoken stimuli to such a degree that visual information becomes the only information in the signal that actually *can* make competitors distinct. An extension of the current study is currently planned which examines the differences between these two sources of noise and the ways in which they interact with working memory.

Other pilot experiments run in our own lab have also shown an effect of presentation mode on short-term memory tasks. Pisoni, Saldaña and Sheffert (1995) showed that audiovisual presentation of letters leads to a decrease in the number of items correctly recalled in lists of length 6 and 7. Again, this task and the stimuli used were extremely dissimilar from the ones presented in the current study, and so it is hard to draw direct comparisons between their results and our own. Still, the fact that multiple studies have found some effect of presentation mode on short-term memory capacity provides support for the proposal that subtle aspects of stimulus presentation have effects at higher levels of processing, especially on immediate memory span.

The present investigation indicates that visual information about a talker's utterances does interact with higher sources of information in short-term memory for spoken words. Further investigation into the nature of this interaction and its effects on spoken word recognition are necessary to explore more fully the ramifications of such a finding. However, these findings hint that the benefits of audiovisual speech may extend beyond the realm of perception, propagating up the processing system to affect encoding, rehearsal, and immediate memory span.

References

Baddeley, A.D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, *18*, 363-365.

Baddeley, A.D. (1998). *Human memory: Theory and practice* (revised ed.). Boston: Allyn & Bacon.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *Recent advances in the psychology of learning and motivation* (Vol. VII, pp. 47-89). New York: Academic Press.

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.

Bavelier, D., & Potter, M.C. (1990). Visual and phonological codes in repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 134-147.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. *Proceedings of 1994 International Conference on Spoken Language Processing, 13,* 559-562.

Bourassa, D.C., & Besner, D. (1994). Beyond the articulatory loop: A semantic contribution to serial order recall of subspan lists. *Psychonomic Bulletin & Review, 1*, 122-125.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55*, 75-84.

Cowan, N., Wood, N.L., Nugent, L.D., & Treisman, M. (1997). There are two word-length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science, 8*, 290-295.

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics, 51,* 355-362.

Fisher, B. D., & Pylyshyn, Z. W., (1994). The cognitive architecture of bimodal event perception: a commentary and addendum to Radeau (1994). *Current Psychology of Cognition, 13,* 92-96.

Gathercole, S.E., Frankish, C.R., Pickering, S.J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 84-95.

Goh, W.D., & Pisoni, D.B. (1998). Effects of lexical neighborhoods on immediate memory span for spoken words: A first report. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 195-213). Bloomington, IN: Speech Research Laboratory, Indiana University.

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.

Green, K. P. (1996). The use of auditory and visual information in phonetic perception. In D. Stork & M.E. Hennecke, (Eds.), *Speechreading by humans and machines* (pp. 55-77). Springer-Verlag: Berlin.

Green, K. P., Kuhl, K. P., & Meltzoff, N. A. (1988). Factors affecting the integration of auditory and visual information in speech: the effect of vowel environment. Paper presented at the meeting of the Acoustical Society of America, Honolulu.

Hulme, C., Maughan, S., & Brown, G.D.A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language, 30*, 685-701.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G.D.A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1217-1232.

Jones, J. A., & Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics, 25*, 13-19.

Jordan, T. R., & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology : Human Perception and Performance, 23*, 388-403.

Lachs, L., & Hernàndez, 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. (1999). *Effects of multi-modal speech cues on recognition memory for spoken words*. Manuscript submitted for review.

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.

La Pointe, L.B., & Engle, R.W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1118-1133.

Lively, S. E., Pisoni, D. B., & Goldinger, S. E. (1994). Spoken word recognition: Research and theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. New York: Academic Press.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon *Research on Speech Perception Technical Report No. 6*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1-36.

Massaro, D. W. & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics, 58*, 1047-1065.

Massaro, D. W., Cohen, M. M. & Smeele, P. M. T. (1995). Cross-linguistic comparisons in the integration of visual and auditory speeh. *Memory & Cognition, 23*, 113-131.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1-86.

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

Munhall, K. G., Gribble, P., Sacco, L. & Ward, M. (1995). Temporal constraints on the perception of the McGurk effect. *Perception & Psychophysics, 58*, 351-362.

Nairne, J.S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin & Review, 4*, 541-545.

Pichora-Fuller, M. K. (1996). Working memory and speechreading. In D. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 257 - 274). Springer-Verlag: Berlin.

Salame, P., & Baddeley, A.D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior, 21*, 150-164.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition, 21*, 168-175.

Sheffert, S., Lachs, L. & Hernandez, L. R. (1996-1997). 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.

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

Smeele, P. M. T., Sittig, A. C., & Van Heuven, V. J. (1994). Temporal organization of bimodal speech information. In *International Conference on Spoken Language Processing: Vol. 3* (pp. 1431-1434).

Texas Instruments. (1991). TI 46-word speaker-dependent isolated word corpus (CD-ROM). Gaithersburg: NIST.

Treisman, M. (1978). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior, 17*, 37-59.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

The Influence of Lexical Neighborhoods and Stimulus Sampling Procedures on Children's Immediate Memory Span for Spoken Words: A Report of Work in Progress[1]

Miranda Cleary, Winston D. Goh,[2] Jaime Brumfield, and David B. Pisoni[3]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH Research Grant DC00111 and Training Grant DC00012 to Indiana University.
[2] Also Department of Social Work & Psychology, National University of Singapore.
[3] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

The Influence of Lexical Neighborhoods and Stimulus Sampling Procedures on Children's Immediate Memory Span for Spoken Words: A Report of Work in Progress

Abstract. Twenty-four normal-hearing children completed a memory span task requiring immediate recall of lists of spoken words. Memory span for lists of lexically "easy" words was compared to memory span for lists of lexically "hard" words, using items selected from the children's Lexical Neighborhood Test (LNT; Kirk, Pisoni, & Osberger, 1995). In addition, two different "sampling" conditions were compared. Lists were created either by sampling repeatedly from a small closed set of words such that in each successive list, some words from previous lists were again heard, or by "non-repeated" sampling, such that on each list, every word was uniquely encountered. A main effect of sampling procedure was obtained. Overall, the children's performance was better when the test items used were repeated within the set of lists in a given condition than when novel items were used for each test list. No significant main effect of the lexical discrimination manipulation was observed, although the means fell in the predicted direction; lists of lexically "easy" words were recalled somewhat better than lists composed of lexically "hard" words. No significant interaction was found. A revised version of this study is proposed.

Introduction

Recently, there has been renewed interest in the relationship between factors that influence the identification of spoken words and the mechanisms of phonological short-term memory. It has long been observed that short-term memory can hold only a remarkably small number of items (Miller, 1956) and that without the use of rehearsal or other maintenance processes, the information in short term storage is often lost within a matter of minutes (Atkinson & Shiffrin, 1971). The current paper is concerned with how young children's representations of spoken words in long-term memory may interact with the short-term activation involved in performing an auditory short-term/working memory span task.

Users of a spoken language acquire a repertoire of words that they are able to recognize. Various authors have referred to these long-term memory representations as a language user's "mental lexicon." The structure of these mental representations has been argued to reflect certain key aspects of language experience. For example, some words in the language are encountered very often and are therefore robustly coded as high frequency lexical entries. Other words are less commonly encountered and thus may not be as firmly established in long-term memory. As a consequence, a low-frequency word may then be perceived less reliably than a high-frequency word. For example, the word "meek," if heard out of context through noise, might be interpreted as "meat," one of its *lexical neighbors*. Two words can be operationally defined as lexical neighbors of each other on a segmental level of analysis if a single phoneme can be substituted, subtracted or deleted to transform one word into the other (Landauer & Streeter, 1973; Luce, 1986). When a word has many such lexical neighbors, it is considered to be high in neighborhood density (Luce, Pisoni, & Goldinger, 1990). Word frequency and lexical neighborhood characteristics can combine to make some words particularly easy to identify and other words much harder to recognize. A word can operationally be defined as "hard" to identify if it is a low frequency item having many high frequency lexical neighbors: "competitive" activation of high-frequency neighbors hinders correct identification of the intended word. In contrast, a lexically "easy" word can be defined as being of high frequency and having few lexical neighbors, with its few existing neighbors being low in frequency. It has been shown that individuals

recognize lexically "easy" words more accurately than lexically "hard" words (Cluff & Luce, 1990; Kirk, Pisoni & Osberger, 1995; Luce, 1986; Luce, Pisoni, & Goldinger, 1990).

Word frequency and lexical neighborhood characteristics clearly depend critically on the language experience of the language user. An adult's mental lexicon contains a large vocabulary of lexical items with very wide range in experienced frequency among the items. For an adult, with increased vocabulary size comes a greater likelihood of "densely populated" lexical neighborhoods. A young child's lexicon, in contrast, is fairly limited. Vocabulary size is smaller and experienced frequencies are much lower. By virtue of a smaller vocabulary size, the number of neighbors that a child could potentially confuse a known word with is reduced (Charles-Luce & Luce, 1990; Charles-Luce & Luce, 1995).

The perceptual confusability of words within a child's lexicon as predicted by word frequency and lexical neighborhood characteristics were investigated by Logan (1992) using children's speech samples from the CHILDES database (MacWhinney & Snow, 1985). Logan found systematic relationships between the lexicons of preschool-age children and those of adults. In particular, neighborhood density increased as a function of age, suggesting that basic underlying principles of lexical organization might be similar for children and adults (see Charles-Luce & Luce (1990) however, for a discussion of key differences).

As an extension of Logan's computational analyses, Kirk, Pisoni and Osberger (1995) developed the Lexical Neighborhood Test (LNT) to assess spoken word recognition in hearing-impaired children. One hundred monosyllabic words were chosen from the speech corpus previously analyzed by Logan. Multi-syllabic words, proper nouns, possessives, contractions, plurals and inflected forms of words were excluded from the LNT word lists. The words chosen consisted of 50 lexically "easy" words and 50 lexically "hard" words. Pediatric cochlear implant users, that is, hearing-impaired children that use a prosthetic device to receive partial auditory information via direct electrical stimulation of their auditory nerve, were tested on their ability to recognize lexically "easy" vs. lexically "hard" words. As predicted, Kirk, et al. found that these children displayed better open set word recognition scores for the LNT "easy" words than the LNT "hard" words. Whether or not these lexical influences continued to be demonstrated in auditory *memory* tasks, rather than just simple word identification tasks, was however, not tested at that time.

The original LNT word lists cannot be used to test for lexical influences in adults since these lists were specifically designed for preschool-age children. A database consisting of lexically "easy" vs. lexically "hard" words as defined as per a typical adult vocabulary was, however, recently constructed in our lab (see Torretta, 1995). In order to examine possible lexical influences on auditory memory span, Goh and Pisoni (1998) used this database to test adults' immediate recall for lists of lexically "easy" vs. lexically "hard" words. They were also was interested in whether the effects of prior exposure on immediate recall would differ in magnitude for "easy" vs. "hard" word lists. A sampling variable was therefore also manipulated. Some subjects were required to recall lists of words in which individual items were repeated from list to list, though never within a single list ("repeated sampling"). Another group of participants was required to recall lists in which a given word was never repeated either within or across lists ("non-repeated sampling"). After hearing the target list, the participants were required to write down the words in the correct serial order. Goh and Pisoni found that in the non-repeated sampling condition, memory span for lists of lexically "easy" words was significantly better than memory span for lists of lexically "hard" words. However, their results also showed that in the *repeated* sampling condition, neighborhood characteristics had little effect on memory span. The absence of a lexical effect in the repeated sampling condition was accounted for by proposing that the repeated use of the same lexical items from list to list only required repeated access to a small, and already activated subset of the adult's mental

lexicon. Once the representations of the repeated words were successfully activated, it was suggested that on repeated exposure, neighborhood-competition effects had little influence on the probability of subsequent identification and interference with recall. In the non-repeated sampling conditions, however, there was uncertainty in identification not just due to the novelty of every word in each list, but in the "lexically hard" condition, also as result of each target word's many lexical neighbors/competitors being activated during the identification process.

In the present study we attempted to replicate the results of Goh and Pisoni (1998) using the children's LNT word lists already described. Like Goh and Pisoni, we manipulated the sampling variables (repeated vs. non-repeated) and the lexical characteristics of the test items used ("easy" vs. "hard"). Effects analogous to those obtained by Goh and Pisoni (1998) were predicted: i. e., evidence of a lexical effect (recall of "easy" words would be better than recall of "hard" words) in the non-repeated sampling condition, but not the repeated sampling condition. Obtaining this pattern of results would provide support for the hypothesis that lexical neighborhood characteristics influence the encoding of and memory for spoken words only when the items have not already been recently activated within the mental lexicon.

Method

Participants

Thirty normal-hearing children participated in the study. Six children were unable to complete the task. The data from twenty-four children were retained in the analysis. The children ranged in age from 53 months (4;5) to 89 months (7;5). In reporting the results below, the four- and five-year-old children between the ages of 53 to 70 months (N=15) are referred to as "the young group" while the remaining nine children, ranging in age from 76-89 months are referred to as "the older group."

The wide range in ages represented in this sample arose for the following reason: Initially, we had planned to use this task with children for whom the LNT statistics were appropriate--that is, since the lexical characteristics of the LNT were designed to match the lexicons of three-to-five-year-old normally developing children, it would be most appropriate to see if children within this same age range would demonstrate a sensitivity to the neighborhood characteristics manipulation. However, after recruiting a number of young participants, it became clear that most of the four-year olds and some of the five-year olds simply could not complete the immediate recall task as required for inclusion in the data analysis. We then tried to recruit some older children to see how old the child had to be to complete the task as designed. We have since redesigned the procedure so that it can be run with young five-year-olds with reasonable levels of performance. This short report-in-progress, however, documents our early experimentation with the procedure. Since all variables were manipulated within subject, we can at least begin to look for possible effects in this sample, despite the wide age range.

Materials and Procedure

List Items. The stimulus materials were digitally recorded tokens taken from the LNT word lists (Kirk, Pisoni, & Osberger, 1995; Kirk, 1999). The LNT consists of one hundred monosyllabic words known to be familiar to children of age three years and up. This word list was specifically designed to contain words having very high vs. less high probability of accurate word recognition through the simultaneous manipulation of several variables. Half the words, the lexically "easy" words, were selected to be relatively more frequent in the child's usage, to have relatively few lexical neighbors, and to have their lexical neighbors be high in frequency. The remaining words, the lexically "hard" words, were less

frequent (though still highly familiar) in children's speech and had many relatively high frequency lexical neighbors. Individually edited speech files were created by Kirk and her colleagues for a past study. Details of the recording process can be found in Kirk, Eisenberg, Martinez, and Hay-McCutcheon (1998). All LNT recordings were from a single young adult male talker. During the testing session, auditory materials were presented via a table-top loudspeaker (Advent Model AV280, 10 Watts amplifier output power, THD < 1%, frequency response 70 Hz-20 kHz) at approximately 70 dB SPL as determined via a hand-held sound level meter (Triplett Model 370) held at approximately the level of the child's head.

The lexical statistics for the words used in each of the four conditions in this study are provided in the Appendix. Statistics for each word were provided by authors of the LNT and are based on the corpus analyzed by Logan as previously mentioned. Forty-two words were required to create the word lists. Eight words from each of the LNT "easy" and "hard" lists were therefore eliminated (see note to Appendix). Six "easy" words and six "hard" words were selected from their respective LNT lists to be used in the "repeated" conditions. Both sets of six words were selected to roughly be representative of the larger list from which they were chosen, in terms of their lexical characteristics. The remaining thirty-six words from each list were used in the "non-repeated" conditions.

Lists. Presentation of the stimuli was controlled by a computer program specially created for this purpose, running on a PC computer. The child's responses were recorded on audio-tape using a omni-directional microphone (Electro-Voice DO54 Dynamic Omni Directional Microphone) mounted on a table-top stand positioned approximately a foot away from the child's mouth, and a Sony audio cassette tape-recorder (Sony TC-D5M).

The lists were generated pseudo-randomly by the computer program. In the "non-repeated" condition, all words were used exactly once in a given test session. In the "repeated" sampling condition, the words in the six-word set were repeated from list to list, but not within any one list. A 500 ms delay was inserted between each list item. A computer-generated "bell-like" sound was played after the last item in each list as a signal for the child to begin recall.[4]

Each experimental condition contained eight word lists. The child heard two lists of three items apiece, two at length four, two at five, and two lists at length six. There were thirty-two lists total. A short break was instituted between each of the four conditions. The four conditions were counterbalanced across subjects (each of the twenty-four permutations of the four conditions was used exactly once).

Participants were tested at the Speech Research Laboratory at Indiana University by graduate and undergraduate research assistants. All children were tested individually in a quiet room. A practice trial was given to insure that the child understood what was being of asked of them. The child was told to "repeat the list of words" back to the experimenter. The experimenter provided no explicit feedback regarding the accuracy of the child's responses. Verbal encouragement was given only to keep the child on task during the procedure. Because the procedure appeared to tax the attention span of most of the children we observed, coaxing was often required to get the child to finish the task. Each condition of the memory span task took approximately five minutes to complete, for a total length of approximately twenty to twenty-five minutes for all four conditions.

[4] We had found that several pilot subjects began repeating the lists before presentation had finished. Although we are aware of possible "suffix" effects as reported in the literature on immediate recall (e.g., Conrad, 1960), this procedure was in place for all children and all conditions, and thus any resulting decrement in scores should be reflected in the means for every condition equally. Note that a "non-speech" suffix was used.

Scoring. All scoring procedures were identical to those used in the earlier study by Goh and Pisoni (1998). Goh and Pisoni used four different scoring criteria, of which we report only two—"total span" and "absolute span." We used only these measures for analysis because the other scoring methods were far more conservative in awarding credit and yielded scores of zero for many participants. The Total Span score was obtained by counting how many words were recalled correctly in the correct list position regardless of whether the entire list was correct. The Absolute Span score was obtained by counting only those words that were recalled in the context of a list that was perfectly recalled. For both Total and Absolute Span scoring methods, the child's set of responses was aligned such that the cardinal position of the response had to match that in the presented list to be marked correct. For example, if the target list was "toe, dry, game" and the child said, "dry, game," no points were awarded for either Total or Absolute Span. This scoring procedure, which requires correct recall of both item and order information, was adopted to match the paper-and-pencil scoring methods used in Goh and Pisoni.

WISC Digit Span. In order to obtain a widely used and accepted measure of memory span against which to compare the word span results, all of the children also completed the WISC Digit Span Supplementary Verbal Sub-test of the Wechsler Intelligence Scale for Children, Third Edition (WISC-III; Wechsler, 1991). A recorded version of this test that was created for another study (Kirk, 1999) was used. This memory span task requires the child to repeat back a list of digits as spoken by a recorded female talker at a rate of approximately one digit per second (WISC-III Manual; Wechsler, 1991). In the "digits-forward" section of the task, the child is required to simply repeat back the list as heard. In the "digits-backwards" section of the task, the child is told to "say the list backward." An example of reversing a list length of two items is provided and a practice trial is given for the backwards task. In both parts of the WISC task, the lists begin with two items, and are subsequently increased in length upon successful repetition until a child gets two lists incorrect at a given length, at which point testing stops. Points are awarded for each list correctly repeated with no partial credit. The WISC-III testing manual provides two lists for use at each list length, with the possible list length ranging from two to nine items in the forward condition, and from two to eight items in the backwards condition. Digits are not repeated consecutively within any list and each list is unique.

Results and Discussion

An alpha level of .05 was adopted for all statistical tests reported below. The means for Total Span scores in each of the four conditions are shown in Figure 1. The average spans for the "repeated" conditions are shown in the two bars to the right, while the average spans for the "non-repeated" sampling conditions are shown on the left. Clear bars indicate the means for the "easy" word list conditions and the hatched bars the means for the "hard" word list conditions. As can be seen from the figure, performance was better in the "repeated" sampling conditions than in the "non-repeated" sampling conditions. Contrary to expectation, however, little to no advantage was observed for the "easy" word list conditions over the "hard" word list conditions. A 2 x 2 repeated measures ANOVA on the Total Span scores showed a significant main effect of "repeated" vs. "non-repeated" sampling, F (1, 23) = 16, p = .001, MS_e = 9.53. No significant main effect was obtained for the "easy" vs. "hard" manipulation and there was no statistical evidence of an interaction (all Fs < 1). Analogous results were obtained using the Absolute Span scores, once again showing a significant main effect of sampling, F (1, 23) = 14.24, p < .01, MS_e = 14.34, and no evidence of either a main effect of lexical neighborhood or an interaction.

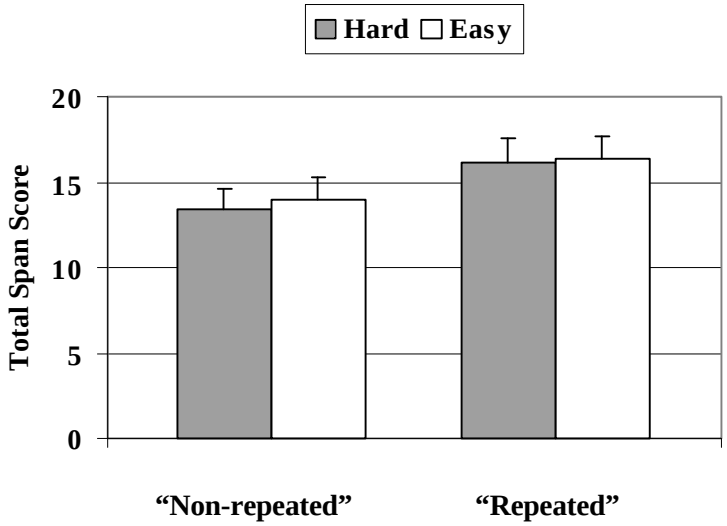


Figure 1. Mean spans (+SE) for "Non-Repeated" [left] vs. "Repeated" [right] conditions. Hatched bars indicate mean span in "Hard" word conditions, clear bars indicate mean span for "Easy Word" conditions.

The data in Figure 2 indicate that for this younger subgroup of children, repeated sampling clearly resulted in significantly higher Total Span scores (F (1, 14) = 10.67, $p < .01$, MS_e = 11.83), and that lexically "easy" word lists tended to be better recalled than lexically "hard" word lists F (1, 14) = 3.78, p = .07, MS_e = 7.41, although this difference did not reach statistical significance. Note also that while the overall "easy" vs. "hard" effect fell in the predicted direction, this difference was slightly larger in the "repeated" sampling conditions as opposed to the "non-repeated" sampling conditions, contradicting our prediction that the lexical manipulation would have an effect only in the "non-repeated" conditions. The interaction between sampling and lexical neighborhood was not significant, p = .33. Analogous analyses on the Absolute Span scores gave a very similar pattern of results: a significant main effect of sampling, F (1, 14) = 5.20, $p < .05$, MS_e = 16.61, weak evidence of a main effect of lexical neighborhood, F (1, 14) = 2.90, p = .11, MS_e = 7.46, and no interaction between the two variables.

At the time of this writing, the data set from the fifteen four and five year olds is not completely counterbalanced. Additional subjects need to be recruited to complete this data set. The remaining nine children (age 6 and 7 years), demonstrated the sampling effect, but showed little influence of the "easy" vs. "hard" word manipulation in the "non-repeated" condition (like the younger children). These older children also showed a non-significant difference in the "repeated" condition in the direction opposite to that predicted, thus contributing to failure to find a main effect of neighborhood in the larger group of twenty-four children.

The correlations between children's WISC verbal digit spans and their word span scores in each of the four experimental conditions (scored as "total span") are shown below in Table I. As expected, these two sets of measures were, in general, strongly positively correlated. We had, however, expected that the

forward digit span measure would correlate more highly with the repeated-sampling word-span conditions than the non-repeated word span conditions, since digit span involves repeated sampling from a small set of possible items. As can be seen in Table I, the data did not provide support for this conjecture.

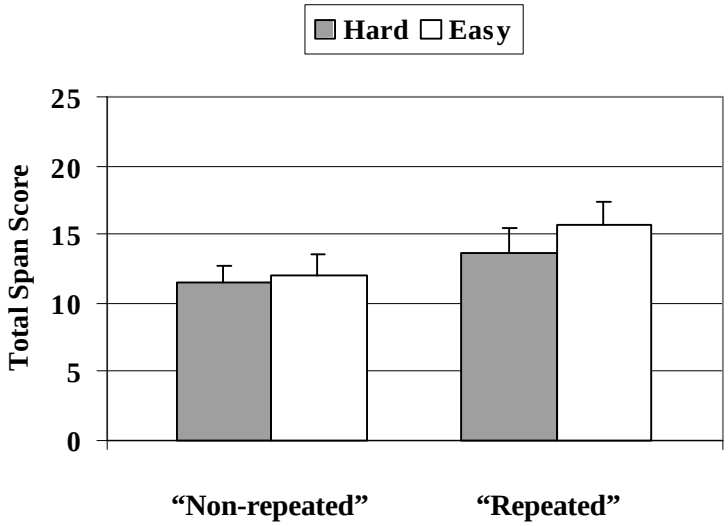


Figure 2. Mean spans (+SE) for only the four- and five-year old children (N=15) in the "Non-Repeated" [left] vs. "Repeated" [right] conditions. Hatched bars indicate mean span in "Hard" word conditions, clear bars indicate mean span for "Easy Word" conditions.

Although we still have to collect additional data, several trends are already clear. Both groups of children are clearly sensitive to the sampling manipulation, thus showing that when spoken words have already been identified and activated from long-term memory, subsequent recall using these same words is better than when novel words must be accessed for every list. The lexical neighborhood manipulation, however, is difficult to interpret due to methodological problems that we are currently in the process of rectifying. The memory task needs to be made easier, and the lexical statistics must be appropriate to the age group tested. It would also probably be preferable to rotate the six words chosen for the repeated conditions across the items in the "easy" and "hard" word lists for different subjects rather than using the same set of six words as we did in this version of the experiment. (A fixed set of eight words was also used for all subjects in Goh and Pisoni [1998]). This manipulation might help to insure that the sampling effect and any neighborhood effects that might be obtained are not due specifically to the word set used but to the independent variables being manipulated.

	WISC Forward Digit Span	**WISC Backward Digit Span**
Non-repeated Hard	.71**	.31
Non-repeated Easy	.41	.53*
Repeated Hard	.42*	.54**
Repeated Easy	.60**	.29

* $p < .05$

** $p < .01$

Table I. Correlations between WISC digit span measures and LNT word span conditions. Age in months, although not a strong co-variate, has been statistically partialled out.

It would also be advantageous to be able to separate out the independent effects of word frequency, neighborhood density, and neighborhood frequency in contributing to any lexical neighborhood effects found. To do this, new word lists specifically designed to test these factors will need to be constructed. A revised study is underway to address some of these issues.

In summary, the present study replicated the main effect of the stimulus sampling procedure as reported by Goh and Pisoni (1998) but failed to obtain evidence in young children for the interesting lexical interaction reported in that paper--namely, that lexical neighborhood effects influence the encoding of and memory for spoken words only when the items have not already been recently (and repeatedly) activated within the mental lexicon.

References

Atkinson, R. C., & Shiffrin, R.M. (1971). The control of short-term memory. *Scientific American*, 225, 82-90.

Baddeley, A. D. (1986). *Working Memory*. London: Oxford University Press.

Baddeley, A. D. (1992). Working memory. *Science*, 255, 556-559.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language*, 17, 205-215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighborhoods in young children's receptive vocabularies. *Journal of Child Language*, 22, 727-735.

Cluff, M. S., & Luce, P.A. (1990). Retroactive effects on multiple activation. Journal of Experimental Psychology: *Human Perception and Performance*, 16, 551-563.

Conrad, R. (1960). Very brief delay of immediate recall. Quarterly Journal of Experimental Psychology, 12, 45-47.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test, Third Edition*. Circle Pines, Minnesota: American Guidance Service.

Goh, W.D., & Pisoni D.B. (1998). Effects of lexical neighborhoods on immediate memory span for spoken words. *In Research on Spoken Language Processing Report No. 22* (pp. 195-213). Bloomington, IN: Speech Research Laboratory, Indiana University.

Kirk, K. I. (1999). Assessing speech perception in listeners with cochlear implants: The development of the lexical neighborhood tests. *The Volta Review*, 100, 63-85.

Kirk, K., Eisenberg, L., Martinez, A. & Hay-McCutcheon, M. (1998). The lexical neighborhood test: test-retest reliability and inter-list equivalency. *Research on Spoken language Processing: Progress Report 22*. (pp. 170-190). Bloomington, IN: Speech Research Laboratory, Indiana University.

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.

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. (1986). *Neighbors of words in the mental lexicon* (Research on Spoken Language Processing Technical Report No. 8). Bloomington, IN: Indiana University.

Luce, P. A., Pisoni, D. B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), Cognitive models of speech processing: *Psycholinguistic and computational perspectives*. Cambridge, MA: MIT Press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, 12, 271-296.

Miller, G.A (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63, 81-97.

Pisoni, D. B., Nusbaum, H. C., Luce, P.A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon, *Speech Communication*, 4, 75-95.

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

Appendix

Eliminated LNT words: (Easy: down, old, orange, six, three, want, watch. Hard: cut, cut, ear, mine, nine, ten, zoo.) Rationale: not clearly a monosyllabic word ("orange"), number names since digit span was used as a separate task ("six," "three," "nine", "ten"), extreme on the frequency distribution ("want," "down," "watch," "put"), incomplete statistics available ("zoo"), function word with low intelligibility ("mine"), appears twice on LNT ("cut," "cut"), began with a vowel ("old," "ear"), potentially confusing ("wrong").

Lexical Statistics for Words Used in the Easy Repeated Condition

	Orthography	Pronunciation	Word Frequency	Mean Neighborhood Density	Mean Neighborhood Frequency
1	live	lIv	4	3	4
2	jump	J^mp	5	1	2
3	cold	kold	5	3	16.33
4	need	nid	16	2	3
5	boy	bO	16	4	8
6	truck	tr^k	4	1	2
		MEAN	8.33	2.33	5.89
		SD	5.96	1.21	5.58

Lexical Statistics for Words Used in the Hard Repeated Condition

	Orthography	Pronunciation	Word Frequency	Mean Neighborhood Density	Mean Neighborhood Frequency
1	fat	f@t	1	8	14
2	cake	kek	3	8	6
3	toe	to	2	11	52.36
4	use	yuz	2	4	55
5	dry	drY	1	4	4
6	game	gem	1	4	2
		MEAN	1.67	6.50	22.23
		SD	0.82	2.95	24.72

Lexical Statistics for Words Used in the Easy Non-Repeated Condition

	Orthography	Pronunciation	Word Frequency	Mean Neighborhood Density	Mean Neighborhood Frequency
1	farm	farm	15	1	2
2	house	hWs	14	2	24
3	white	hwYt	11	2	41.5
4	gray	gre	11	1	2
5	brown	brWn	11	1	1
6	drive	drYv	10	2	2
7	green	grin	10	0	0
8	break	brek	9	1	9
9	snake	snek	9	1	1
10	food	fud	9	0	0
11	push	pUS	8	2	30
12	count	kWnt	8	2	15
13	friend	frEnd	8	1	3
14	girl	gRl	8	1	1
15	good	gUd	12	4	3
16	more	mor	7	3	21
17	stop	stap	7	3	3
18	school	skul	7	1	8
19	give	gIv	7	1	4
20	cow	kW	6	4	34
21	home	hom	6	4	4
22	sit	sIt	5	4	52
23	hold	hold	5	4	12
24	hard	hard	8	2	1
25	bird	bRd	5	3	2.67
26	fish	fIS	8	0	0
27	mouth	mWT	5	0	0
28	foot	fUt	4	3	17
29	catch	k@C	4	3	16
30	hurt	hRt	5	3	3
31	swim	swIm	4	3	2
32	dance	d@ns	4	2	9.5
33	time	tYm	10	4	10
34	juice	Jus	4	1	1
35	stand	st@nd	4	0	0
36	pig	pIg	14	4	8
		MEAN	7.83	2.03	9.52
		SD	3.08	1.36	12.70

Lexical Statistics for Words Used in the Hard Non-Repeated Condition

	Orthography	Pronunciation	Word Frequency	Mean Neighborhood Density	Mean Neighborhood Frequency
1	pie	pY	4	11	60
2	wet	wEt	4	8	12.5
3	sing	sIG	4	8	3
4	hand	h@nd	4	5	40
5	read	rid	4	5	7
6	hi	hY	3	16	56
7	tea	ti	3	14	48
8	hot	hat	3	11	14
9	call	kal	3	8	11
10	fun	f^n	3	6	18
11	ball	bal	3	6	10
12	book	bUk	3	6	7
13	work	wRk	3	5	1
14	cut	k^t	3	4	12
15	dad	d@d	2	8	17
16	bed	bEd	2	7	5
17	cook	kUk	2	6	6
18	pink	pIGk	2	5	3
19	grow	gro	2	4	18
20	lost	last	2	4	3
21	fight	fYt	1	10	4
22	goat	got	1	8	31
23	toy	tO	1	7	43
24	gum	g^m	1	7	18
25	bone	bon	1	7	11
26	sun	s^n	1	6	24.83
27	thumb	T^m	1	6	21
28	seat	sit	1	6	20.17
29	run	r^n	1	6	17
30	cap	k@p	1	6	11
31	kick	kIk	1	6	2
32	meat	mit	1	5	14
33	bag	b@g	1	5	10
34	ride	rYd	1	5	8
35	song	saG	1	5	4.4
36	bath	b@t	1	5	4
		MEAN	2.08	6.86	16.53
		SD	1.13	2.66	15.30

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

Audio-Visual Integrative Abilities of Prelingually Deafened Children with Cochlear Implants: A First Report[1]

Lorin Lachs, Karen I. Kirk,[2] and David B. Pisoni[2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was funded, in part, by NIH-NIDCD Grants R01 DC00111, T32 DC00012, K08 DC00126, and K08 DC00083.
[2] DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Audiovisual Integrative Abilities of Prelingually Deafened Children with Cochlear Implants: A First Report

Abstract. Although there has been a great deal of recent empirical work and new theoretical interest in audio-visual perception of speech in both normal-hearing and hearing-impaired adults, relatively little is known about the development of these abilities and skills in deaf children with cochlear implants (CIs). This study examined the way in which children integrate visual speech cues available in the talker's face with auditory speech cues provided by their CIs to enhance spoken language comprehension. A measure of audiovisual integration ability, R, developed by Sumby and Pollack (1954), was computed using sentence comprehension scores obtained from the Common Phrases Test that was administered using auditory and audiovisual conditions. R represents the amount of gain provided in the AV condition relative to the possible gain in performance in the A condition. Because R is normalized relative to absolute performance in the A condition, R measures the ability of a perceiver to utilize and integrate combined auditory and visual information. The results of our analyses indicated that children who were better at recognizing isolated words through listening alone were also better at combining and integrating the complementary information about articulation that is available under multimodal presentation. In addition, we found that children who were better integrators also displayed higher speech intelligibility scores. Treatment programs that aim to increase receptive or productive ability in children who have CIs, therefore, may wish to focus and emphasize the inherent cross-correlations between auditory and visual information in speech.

Introduction

With continued broadening of candidacy criteria and technological advances in cochlear implant (CI) signal processing, more children than ever before have the potential to develop spoken word recognition and language processing skills through a CI. However, there are enormous individual differences in pediatric CI outcomes (Fryauf-Bertschy, Tyler, Kelsay, & Gantz, 1992; Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997; Miyamoto et al., 1989; Osberger et al., 1991; Staller, Beiter, Brimacombe, Mecklenberg, & Arndt, 1991; Tyler et al., 1997a; Tyler, Fryauf-Bertschy, Gantz, Kelsay, & Woodworth, 1997b; Tyler et al., 1997c; Zimmerman-Phillips, Osberger, & Robbins, 1997). Some children can communicate extremely well using the auditory/oral modality and acquire age-appropriate language skills, whereas other children may display only minimal spoken word recognition skills and/or display very delayed language abilities (Bollard, Chute, Popp, & Parisier, 1999; Pisoni, Svirsky, Kirk, & Miyamoto, submitted; Robbins, Bollard, & Green, 1999; Robbins, Svirsky, & Kirk, 1997; Svirsky, Sloan, Caldwell, & Miyamoto, 1998; Tyler, Tomblin, Spencer, Kelsay, & Fryauf-Bertschy, in press). These differences currently cannot be predicted prior to implantation but emerge only after electrical stimulation has commenced. Although demographic and audiologic factors play an important role in pediatric CI outcomes, these variables alone are not sufficient to account for the large individual differences (Fryauf-Bertschy et al., 1992; Fryauf-Bertschy et al., 1997). Rather, variability in performance may be related to perceptual, cognitive and linguistic processes involved in the acquisition and processing of spoken language by the central nervous system (Pisoni & Geers, 1998).

One important perceptual process that may provide some insight into the basis of individual differences in multimodal sensory integration. It is well known that normal-hearing adults can make use of both auditory and visual information to enhance speech intelligibility by up to +15 dB (Massaro & Cohen, 1995; Rosenblum, Johnson, & Saldaña, 1996; Sumby & Pollack, 1954; Summerfield, 1987). Relatively little, however, is known about the integrative abilities of prelingually deaf children with CIs.

Cross-modal integration is thought to reflect the ability to use different sources of information in order to perceive the underlying articulations of the talker's vocal tract (Fowler, 1989; Fowler & Dekle, 1991; Massaro & Cohen, 1995; Remez, Fellowes, Pisoni, Goh, & Rubin, 1999; Rosenblum, 1994; Rosenblum & Saldaña, 1996; Vatikiotis-Bateson, Munhall, Kasahara, Garcia, & Yehia, 1996). The seminal work in this area was conducted by Sumby and Pollack (1954) who showed that the intelligibility of spoken words in noise could be increased by as much as +15 dB when normal-hearing participants were able to both hear and see the speaker uttering test items. Because this gain is so large, perceptual integration has the potential to be extremely useful for the study and treatment of speech perception in individuals with hearing impairments. The findings on cross-modal integration may also provide new information about the basis of individual differences among children with cochlear implants.

Some analysis of the integrative abilities of individuals with hearing impairments has been carried out (see Massaro & Cohen, 1999 for a meta-analysis and modeling of some of these studies). For example, Erber (1972) tested the speech perception abilities of children with normal hearing, severely impaired hearing, and profoundly impaired hearing under auditory-alone, visual-alone and auditory-visual conditions. The stimuli used in this experiment were consonants embedded in an /aCa/ environment. Like the participants in Sumby and Pollack's (1954) study, Erber found that the hearing impaired children were able to capitalize on the complementary information provided by the visual modality in the auditory-visual condition, increasing their speech perception performance relative to their performance in the auditory-alone condition. In another study, consonant confusions made by two multiple-channel cochlear implant users were explored when stimuli were presented in auditory-alone, visual-alone, and auditory-visual conditions (Dowell et al., 1982). They found that speech intelligibility was enhanced when patients were able to combine auditory information in the form of electrical stimulation through the CI and visual information. Taken together, these studies demonstrate the nearly universal finding that audiovisual information facilitates speech perception performance relative to performance with auditory-only or visual-only information.

Substantial variation, however, exists in the extent to which an individual's scores are facilitated. Grant, Walden and Seitz (1998) compared observed audiovisual facilitation to that predicted by several models of integration. These models predicted performance based on the unimodal recognition scores for segmental identification. From these unimodal scores, measures of optimal audiovisual integration can be calculated. Grant et al. found that not all individuals integrate auditory and visual information optimally, especially when speech perception accuracy was measured using higher-order units of speech like words and sentences. While the predictions based on segmental accuracy could account for a large amount of the variance observed (e.g., 50% of sentence-level integration variability), the authors conclude that much more work must be carried out before an adequate model of word- and sentence-level integration can be formed. They speculate that such a model will incorporate both lexical factors and semantic contextual information.

Lexical factors have been shown to play a role in the variability observed in the speech perception of cochlear implant users. The Neighborhood Activation Model of spoken word recognition (Luce, 1986; Luce & Pisoni, 1998) proposes that all spoken words are perceived and recognized within the context of similar words contained in the mental lexicon, the mental storehouse of words known by a listener. In this model, three factors are important for the recognition of words. First, the frequency of occurrence of a word in the language acts as a bias for word recognition, such that more frequent words will be recognized more easily than less frequent words. Second, the model assumes that all words with similar acoustic/phonetic form (called the “neighbors”) compete with each other and the target word for activation during the word recognition process. Accordingly, words from dense neighborhoods are less easily recognized than words from sparse neighborhoods. Finally, because the frequency bias acts on the neighbors as well as the target word, the average frequency of occurrence in the neighborhood plays a

role in accuracy of recognition, as well. If a target word is of low frequency relative to the frequency in the neighborhood, then it will be harder to recognize than if it were of high frequency relative to the rest of the neighborhood.

These factors have been incorporated into two tests of open set spoken language processing ability. The Lexical Neighborhood Test (LNT) and multisyllabic Lexical Neighborhood Test (MLNT, Kirk, 1999; Kirk, Eisenberger, Martinez, & Hay-McCutcheon, 1999; Kirk, Pisoni, & Osberger, 1995) incorporate words from two extreme combinations of the lexical factors mentioned above. Lexically "easy" words are words from sparse similarity neighborhoods with relatively high frequencies of occurrence. Lexically "hard" words are words from dense similarity neighborhoods with relatively low frequencies. These lexical factors were calculated relative to children's productive vocabulary (Kirk, 1999; Logan, 1992). Initial analysis of performance on these two tests by children with cochlear implants showed that word recognition accuracy on lists of easy words was better than accuracy on lists of hard words, indicating that children with cochlear implants recognize words in the context of the other words contained in their lexicons (Kirk et al., 1995).

Given the fact that both the variation in audiovisual integrative ability and the variation in cochlear implant users' auditory-alone speech perception ability seem to relate to an underlying level of lexical representation, the present study was carried out to investigate the relationship between auditory-alone measures of spoken language comprehension and audiovisual integrative abilities in hearing-impaired children who use cochlear implants.

Method

Participants

Participants were 27 children with prelingual deafness who had used a multichannel cochlear implant for two years. The average age of onset of deafness was 0.51 years. Their average age at implantation was 4.52 years. The mean unaided auditory threshold (as measured with pure tones) was 112.20 dB HL. Table 1 shows the specific demographic data for each of the participants involved.

Procedures and Measures

Three tests designed to measure spoken word recognition performance in audio-alone conditions were administered live-voice to participants by an audiologist or speech-language pathologist. The Lexical Neighborhood Test (LNT) and Multisyllabic Lexical Neighborhood Test (MLNT, Kirk, 1999; Kirk et al., 1999; Kirk et al., 1995) are new open-set tests of word recognition for children that assess the effects of word frequency and lexical similarity on spoken word recognition in children. Lexical similarity is measured by calculating the number of words that differ from the target word by only one phoneme. Words from "dense" neighborhoods (i.e., words with many neighbors) tend to be more difficult to identify than words that come from "sparse" neighborhoods. Additionally, high frequency words tend to be recognized more easily than low frequency words. Using these lexical factors, it is possible to classify the items on the LNT and MLNT tests into "Easy" and "Hard". Easy words are high frequency words that reside in relatively low frequency, sparse neighborhoods. Hard words are low frequency words that reside in relatively high frequency, dense neighborhoods. Differences between easy and hard words will be maintained throughout the rest of this report because they provide important information about how lexical competition operates during word recognition.

Participants	Etiology	Age at profound loss (yr.)	Unaided PTA (dB HL)	Age CI fit (yr.)	Processor	Strategy	# Active Electrodes	Age at testing (yr.)
1	unknown	Congenital	116.73	3.50	MSP	MPEAK	18.00	5.50
2	unknown	Congenital	111.67	5.50	MSP	MPEAK	22.00	7.70
3	unknown	1.00	120.07	4.90	MSP	MPEAK	22.00	6.80
4	meningitis	.40	111.67	3.80	SPECTRA	SPEAK	18.00	5.90
5	unknown	Congenital	101.67	5.80	MSP	MPEAK	13.00	7.80
6	unknown	Congenital	116.73	4.10	MSP	MPEAK	19.00	6.10
7	genetic	Congenital	103.33	5.20	MSP	MPEAK	22.00	6.90
8	unknown	Congenital	118.43	4.90	WSP	F...	13.00	6.90
9	genetic	3.00	113.37	5.20	MSP	MPEAK	19.00	7.30
10	unknown	Congenital	110.00	3.70	SPECTRA	SPEAK	18.00	5.90
11	meningitis	1.90	118.40	5.40	MSP	F...	8.00	8.00
12	unknown	Congenital	108.37	4.40	SPECTRA	SPEAK	18.00	6.30
13	unknown	Congenital	108.37	4.60	MSP	MPEAK	18.00	6.60
14	genetic	Congenital	111.70	5.30	SPECTRA	SPEAK	19.00	7.20
15	unknown	Congenital	105.00	5.00	MSP	MPEAK	19.00	7.00
16	unknown	Congenital	118.40	5.30	SPECTRA	SPEAK	20.00	7.30
17	unknown	Congenital	100.00	5.20	SPECTRA	SPEAK	22.00	7.40
18	unknown	Congenital	111.67	4.30	MSP	MPEAK	22.00	6.30
19	meningitis	1.80	116.73	4.30	WSP	F...	15.00	6.30
20	meningitis	1.80	116.73	5.30	MSP	MPEAK	16.00	7.10
21	unknown	Congenital	98.33	4.30	SPECTRA	SPEAK	20.00	6.30
22	meningitis	1.40	113.40	2.20	SPECTRA	SPEAK	18.00	4.20
23	unknown	Congenital	113.37	2.90	MSP	F...	21.00	4.90
24	meningitis	1.20	115.07	4.60	MSP	MPEAK	19.00	6.60
25	genetic	Congenital	118.40	4.20	SPECTRA	SPEAK	20.00	6.30
26	meningitis	1.30	113.40	3.10	MSP	MPEAK	19.00	5.30
27	unknown	Congenital	118.43	5.00	MSP	MPEAK	19.00	7.10
Means		0.51	112.20	4.52			18.41	6.56

Table 1. The demographic characteristics of the participants in the present study. For all participants, the length of device use was 2 years.

In addition to the LNT and MLNT tests, the Phonetically Balanced Kindergarten word lists (PB-K, Haskins, 1949) were administered via live voice to test speech perception in an audio-only setting. This test is an open-set measure whose items are balanced for phonetic content. The PB-K is a widely used measure of speech perception ability in children who use cochlear implants (Kirk, Diefendorf, Pisoni, & Robbins, 1997; Meyer & Pisoni, 1999). The three measures referred to above (LNT, MLNT and PB-K) will herein be referred to collectively as the "auditory measures."

A fourth test, the Common Phrases (CP) Test, was administered under three conditions, auditory-alone (A), visual-alone (V) and audiovisual (AV). In this report, we only focus on data for the A and AV conditions because we were interested in the possible gain in performance from visual integration. This test measures the ability to understand phrases used in everyday situations, like "It is cold outside". Performance in each condition is scored by the percentage of phrases correctly repeated by the child. The scores in each of these conditions can be combined to measure R, the relative gain in auditory intelligibility due to the availability of visual information about articulation (Sumby & Pollack, 1954).

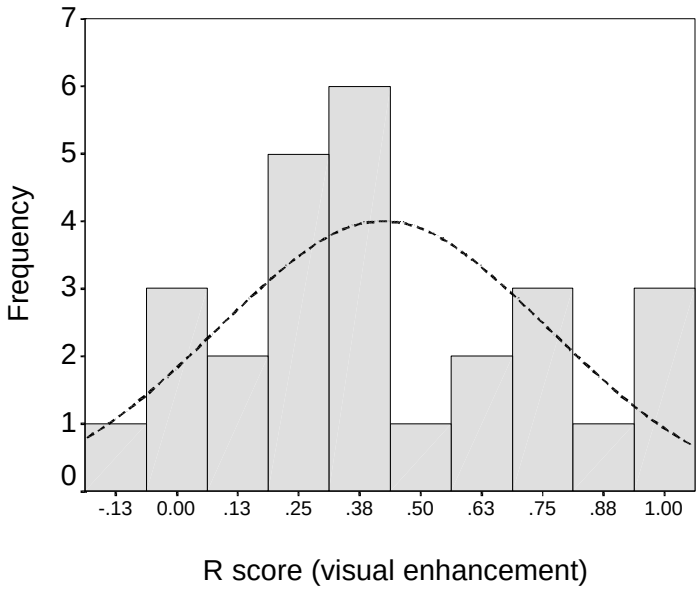


Figure 1. Frequency distribution of R (visual enhancement) scores for the sample. Higher R scores denote higher gains in accuracy in the audiovisual condition relative to accuracy in the audio-alone condition.

R is computed using the following formula:

$$R = {AV - A}/{100 - A}$$

where AV and A represent the accuracy scores obtained in the audiovisual and auditory-alone conditions. From the formula, one can see that R describes the gain in accuracy in the AV condition relative to the accuracy in the A condition, normalized relative to the amount by which intelligibility could have possibly improved. We take R to operationally define the ability to integrate information from disparate sensory modalities.

Finally, a test of spoken language production was administered for each child (Miyamoto et al., 1997). In this test, ten sentences produced by the child are recorded and played back to 3 listeners for transcription. Intelligibility is measured by calculating the average number of words correct for the three talkers.

Results

Audiovisual Integration Scores

Figure 1 shows the frequency distribution of R scores for children in the sample under study in the present experiment. The normal curve superimposed over the data is based on the mean and standard deviation (M = 0.42, SD = 0.34) of the sample. As shown in the figure, our sample was relatively normally distributed with respect to audiovisual integrative ability, with a slight positive skew. It is also interesting to note that these scores vary over a very large range. Some children were able to maximally

Participant	Modality of Common Phrases			
	Auditory-alone	Visual-alone	Audiovisual	R score
1	.00	20.00	30.00	.30
2	20.00	50.00	50.00	.38
3	.00	.00	.00	.00
4	80.00	50.00	100.00	1.00
5	80.00	80.00	100.00	1.00
6	.00	30.00	40.00	.40
7	.00	40.00	20.00	.20
8	10.00	.	80.00	.78
9	60.00	.	70.00	.25
10	80.00	20.00	90.00	.50
11	.00	80.00	90.00	.90
12	70.00	50.00	90.00	.67
13	.00	20.00	10.00	.10
14	90.00	50.00	100.00	1.00
15	10.00	.00	30.00	.22
16	20.00	10.00	20.00	.00
17	30.00	40.00	60.00	.43
18	20.00	.00	20.00	.00
19	40.00	.	60.00	.33
20	.00	50.00	60.00	.60
21	70.00	60.00	80.00	.33
22	40.00	.00	30.00	-.17
23	.00	.	40.00	.40
24	10.00	.	80.00	.78
25	20.00	.	30.00	.13
26	.00	.00	20.00	.20
27	.00	.	70.00	.70
Mean	27.7778	32.5000	54.4444	.4231

Table 2. Performance for each participant on each of the subtests of the Common Phrases test, along with the R score calculated from those measures. A "." indicates that the participant was not tested under that condition.

capitalize on the additional visual information in the AV condition, indicated by R scores of 1.0, while others received little if any benefit, indicated by the zero and negative values. Table 2 shows the specific scores obtained by each child in the auditory-alone, visual-alone, and audiovisual conditions of the Common Phrases test, along with the corresponding R scores. Clearly, children with cochlear implants exhibit a wide range of multimodal integrative abilities in this task.

Correlations with Word Recognition Scores

We were also interested in whether measures of spoken word recognition were related to cross-modal integration. As a first pass at answering this question, the median score for each auditory measure was calculated, and participants were categorized as being in either the low or high group of the split. Using this method, it was entirely possible for a particular patient to be classified in the low group for the split based on one measure and be classified in the high group for the split based on another measure. However, when this happened for a particular child, it only affected classification based on one test. Because not all of the children participated in each of the auditory-alone measures, the total Ns for the various splits differed. The median score, along with the N, for each auditory measure is displayed in Table 3. The maximum score obtainable on each of these measures was 100.

	Auditory Speech Perception Test				
	LNT Easy	**LNT Hard**	**MLNT Easy**	**MLNT Hard**	**PB-K**
Median	28	20	47	53	8
N	17	13	15	11	24

Table 3. Median values and Ns for the five measures of auditory perception.

Figure 2 shows the R scores for the low and high median split groups. Each panel shows the results based on a different median split. The shaded bar in each panel shows the average R score for children in the low group of the median split, and the white bar shows the average R score for children in the high group of the median split. Overall, there is a consistent numerical trend for children in the high median split group to also have higher R scores. However, for some median splits, this trend is more evident than it is for others.

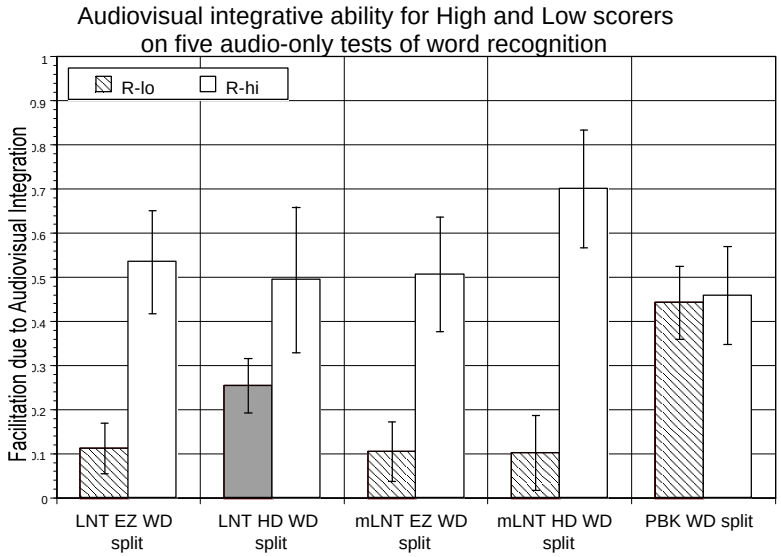


Figure 2. Audiovisual integrative ability (R) for the High and Low groups of the median splits for five measures of auditory-alone word recognition. Error bars are standard errors.

Two-tailed, independent sample t-tests were conducted (using an α level of 0.05) to determine if there were differences in the R scores for children in the high scoring vs. low scoring groups for each auditory measure. Significant differences were found when the children were split by median scores on the LNT Easy test, $t(15) = 3.36$, $p = 0.004$, the MLNT Easy test, $t(13) = 2.62$, $p = 0.021$, and the MLNT Hard test, $t(9) = 3.927$, $p = 0.003$. However, the statistics failed to show a significant difference for splits based on the LNT Hard test, $t(11) = 1.275$, *ns*, and the PB-K test, $t(22) = 0.116$, *ns*. As shown below, these were the two most difficult measures. The data from these t-tests indicates that children who were good performers on auditory-alone measures of spoken word recognition were also good integrators.

In order to more fully characterize the relationship between audiovisual integrative ability and our measures of speech perception, bivariate correlations were calculated between R scores and each of the three auditory measures. Table 4 shows the correlations between each of the measures and R. The correlations show the same pattern of results as the t-tests using median splits. There was a significant relationship between the R score and performance on the LNT Easy, MLNT Easy, and MLNT Hard tests. However, there was no significant relationship between R scores and performance on the LNT Hard and PB-K tests.

	Auditory Speech Perception Test				
	LNT Easy	**LNT Hard**	**MLNT Easy**	**MLNT Hard**	**PB-K**
R score	0.78**	0.28	0.57*	0.68*	0.28

Table 4. The correlation (r) between R scores and the various measures of speech perception. ** indicates significance using an α-level of 0.01; * indicates significance using an α-level of 0.05.

Figure 3 shows the scatterplots for each of these conditions. In each figure, the Y-axis represents the R score and the X-axis represents the score on one of the spoken word recognition tests. Each triangle represents one participant. Shaded triangles are for participants who scored below the median for the test represented on the X-axis, open triangles are for participants who scored above the median. These scatterplots also show the regression line for each relationship. Not surprisingly, the graphs show a strong relationship between R score and the LNT Easy test and both subsections of the MLNT test.

An examination of the scatterplots shows that the lack of a significant relationship between R and the LNT Hard score may be due to an unusually low outlier (the shaded triangle with a negative R score and an LNT Hard score of around 33). This participant is below the median score for all the other tests, except the PB-K, which also failed to show a significant relationship with R. In both the LNT Hard test and the PB-K test, the range of scores is reduced relative to the other tests. Consequently, the median score is relatively low. Table 5 shows the skew values associated with the distribution of scores for each of the auditory measures. Examination of the table shows that the distribution of scores for the LNT Hard and PB-K tests are more positively skewed than the distributions for the other measures used in the present study. These distributional measures confirm that the LNT Hard and PB-K tests were extremely difficult for most of the children involved in our study. We suspect that the lack of significant correlations between these tests and R are due to a floor effect.

	Auditory Speech Perception Test				
	LNT Easy	**LNT Hard**	**MLNT Easy**	**MLNT Hard**	**PB-K**
Skewness	.377	.890	0.322	-0.655	1.167

Table 5. Measures of skewness for the five auditory measures of word recognition. Positive values denote rightward skews and negative values denote leftward skews.

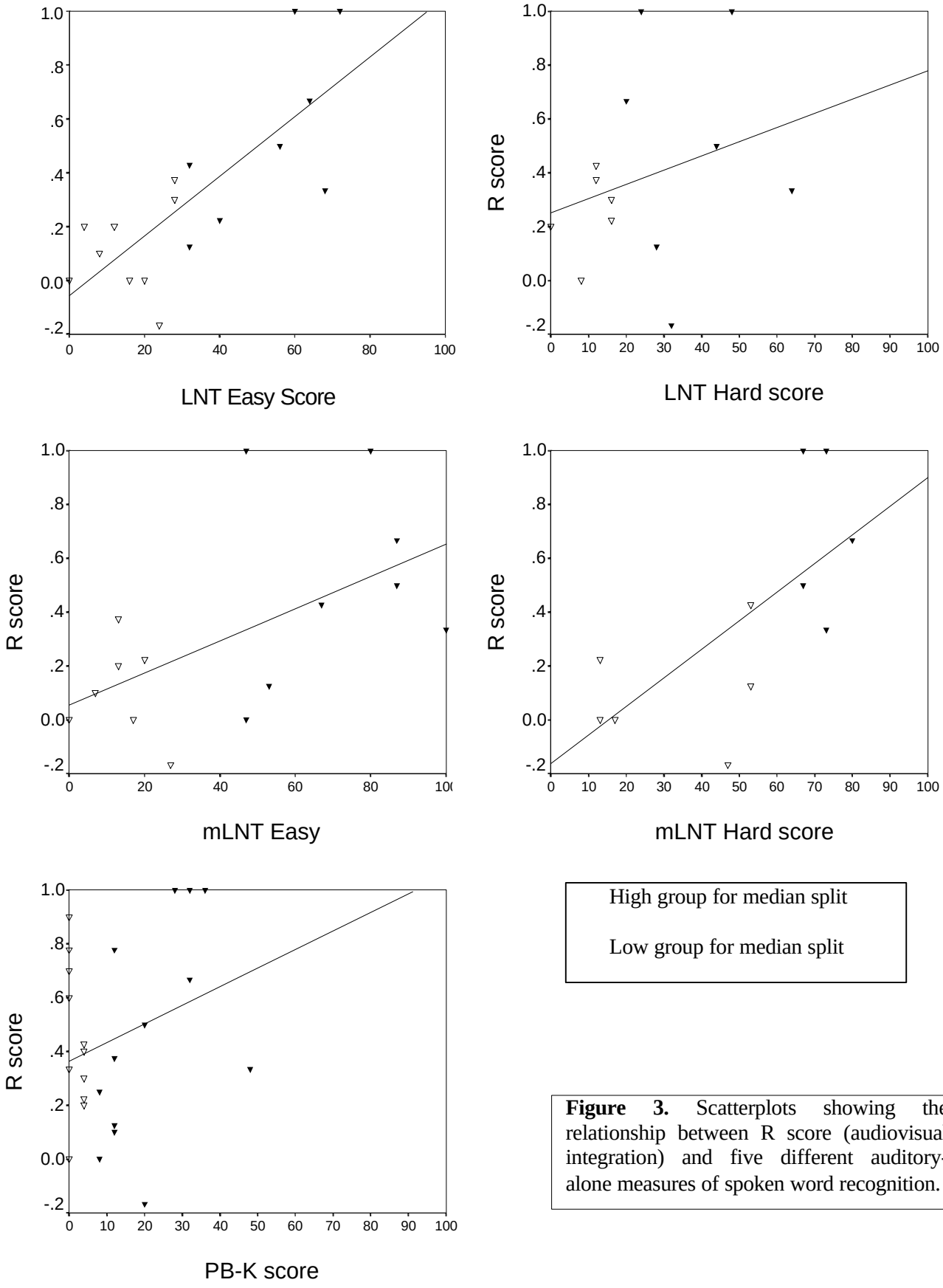


Figure 3. Scatterplots showing the relationship between R score (audiovisual integration) and five different auditory-alone measures of spoken word recognition.

Speech Intelligibility

In addition to the word recognition scores, 23 of the 27 participants also provided scores on a test of speech intelligibility. The mean intelligibility score was 17.13 with a standard deviation of 12.47. The range of intelligibility scores spanned from 2% to 45%. To assess the relations between R and speech intelligibility, we carried out a correlation. The results showed that there was also a significant correlation between intelligibility score and R score, $r(23) = +0.421$, $p = 0.046$. This relationship denotes that children with more intelligible speech are also better audiovisual integrators.

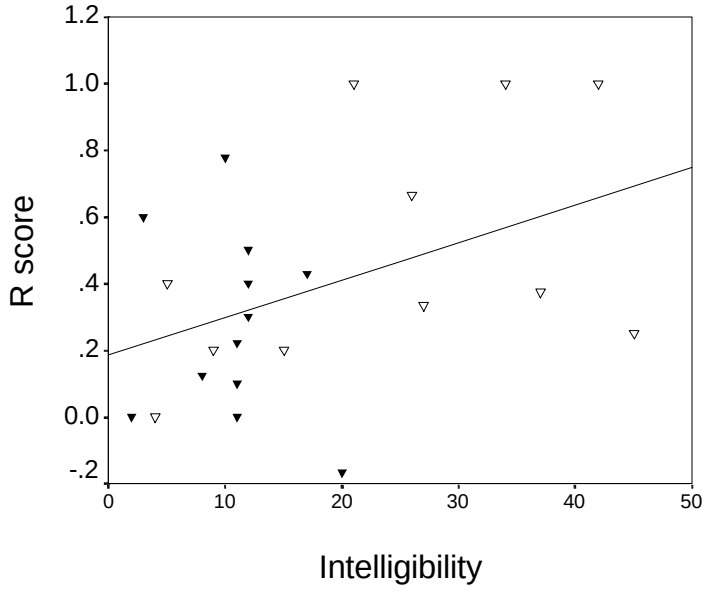


Figure 4. The relationship between integrative ability and intelligibility split by communication mode. OC children are represented by open triangles; TC children are represented by filled triangles.

We also examined the effects of communication mode on R. Eleven of the children who were tested using our measure of speech production were being educated in the Oral Communication (OC) method, while 12 of the children were being educated using the Total Communication (TC) method. Children educated in an oral program are taught to use speaking and listening skills (including lipreading) for communication. Children in this study educated in a TC approach use a combination of spoken and manually-coded English for communication. Figure 4 shows a scatterplot displaying the relationship between intelligibility and visual enhancement (R) for these two groups of children. The diagonal line through the scatterplot is the regression line predicting R from speech intelligibility. For this scatterplot, the data are split by communication mode, with filled triangles indicating children who use TC and open triangles representing children who use OC. As indicated by the positive correlation noted above, higher intelligibility scores were associated with higher R scores. In addition, it appears that most of the high scorers on the intelligibility scale (and consequently the R scale) used OC. In fact, 72.7% of the OC children in this figure were above the median intelligibility score of 12%. In contrast, only 41.7% of the TC children achieved speech intelligibility scores higher than the median.

Discussion

The present study examined the ability of prelingually deafened children with cochlear implants to integrate information about speech from multiple sensory modalities. The results of our analyses demonstrate that these multimodal integration skills are not isolated or independent but are related to more general, unimodal speech perception and production abilities. We observed a strong relationship between R score, a measure of visual enhancement, and the LNT Easy, MLNT Easy, and MLNT Hard tests. Better audiovisual integrators were also better performers on auditory-alone measures of spoken word recognition. In addition, we found a positive correlation between R and speech intelligibility: children who were better AV integrators produced more intelligible speech.

Audiovisual integration reflects the ability of perceivers to combine and use information from multiple sensory modalities to recognize spoken words, syllables, and phonemes (Braida, 1991; Fowler & Dekle, 1991; Green & Gerdeman, 1995; Green & Kuhl, 1991; Kuhl & Meltzoff, 1984; Massaro & Cohen, 1995; Remez et al., 1999; Rosenblum & Saldaña, 1996; Summerfield, 1987; Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzepoulos, 1997). Recent theoretical accounts of AV integration assume that auditory and visual information about speech are integrated by the perceptual system because both carry information relevant to the dynamic behavior of articulating vocal tracts (Rosenblum & Saldaña, 1996). Visual access to the action of the lips, the tongue tip and even the jaw can provide useful information about the behavior of the vocal tract. Auditory access to the action of more internal articulators, such as the tongue blade, the tongue body, and the velum provides information about the same articulatory events (Summerfield, 1987). It is precisely this time-varying articulatory behavior of the vocal tract that has been shown to be of primary importance in the perception of speech (Liberman & Mattingly, 1985; Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez, Rubin, Pisoni, & Carrell, 1981).

Viewed with this theoretical context, the information relevant for speech perception is said to be *modality-neutral*, since it can be carried by more than one sensory modality. In addition, because the acoustic and optic specifications of speech information are produced by the same, underlying articulatory events, they are lawfully related to each other and to the underlying events that produce them (Vatikiotis-Bateson et al., 1996). Consequently, as long as information about the articulations of the vocal tract can be perceived, some degree of speech perception will be possible. Indeed, this fact is clearly demonstrated by the remarkable speechreading abilities of some people with hearing impairments (Rönnberg et al., 1999), and even by those with normal hearing (Bernstein, Demorest, & Tucker, in press). Even more impressive is the finding that information obtained via the *tactile* modality can be used and integrated across modalities during speech perception (Fowler & Dekle, 1991), albeit with limited utility.

However, while the *information* necessary for speech perception may be amodal, the internal *representation* of speech must be based on an individual's experience with the sensory world. Clearly, awareness of the intermodal relationships between auditory and visual information must be contingent on experience with more than one sensory modality. Over time, processes of perceptual learning will associate lawful co-occurences in disparate modalities, until a rich, multimodal representation of speech is obtained.

For prelingually deafened children with cochlear implants, however, these perceptual learning processes only begin to develop after they receive their implant and begin to experience the lawful relationships between sights and sounds in their environment. We assume that the degree to which an individual can integrate information across modalities reflects the degree to which they have internalized the inherent, one-to-one relationships between auditory and visual information about speech. The correlations between audiovisual integrative ability and performance on auditory-alone measures of speech perception for children with cochlear implants indicate that the ability to integrate multimodal sources of information reflects a generalized ability to utilize speech information in *any* of its forms,

including in unimodal auditory specifications. This conclusion is further supported by the finding that the *productive* capabilities of children with cochlear implants, as measured by speech intelligibility, is significantly correlated to their ability to *perceptually* integrate multimodal sources of information.

We suggest that the generalized ability to utilize articulatory information during speech perception and production arises from rich, more fully specified, multimodal internal representations of the articulatory behavior of the vocal tract. At present, we can only speculate as to how these desirable representations are obtained, and why there might be individual differences in the ability to form them. However, the distribution of TC and OC children in the scatterplot of the relationship between intelligibility and integration ability suggests that early experience that focuses on the vocal communication of language may help to strengthen and solidify speech representations in memory. Teaching hearing-impaired children about language *in general* does not appear to be sufficient for building the kinds of representations most advantageous for vocal speech perception. Rather, a focus on the articulatory events that produce speech, by training orally and aurally, leads to improved ability in the perception and production of spoken language. The relationships observed in the present study between audiovisual integrative ability on the one hand and performance measures of spoken word recognition and speech intelligibility on the other hand suggests that treatment programs should concentrate on teaching children with CIs about the inherent cross-correlations between auditory and visual information in speech. In this way, more robust mental representations of the kinds of events that produce speech will be formed, ultimately leading to more robust performance.

General Discussion and Conclusions

The findings from the present investigation suggest that larger gains in spoken word recognition and language comprehension performance might be obtained if deaf children with CIs were forced to use all sources of sensory information during early stages of perceptual learning after implantation. Our results also suggest that these gains might be more readily acquired if intervention programs, like some oral/aural programs, emphasized the robust multimodal nature of speech events. At the present time, most assessments of performance are often done using “auditory-only” presentations of test materials. For some children who are good lip-readers and who display large gains in visual enhancement, these tests may not provide an accurate assessment of their “true” perceptual skills because an important component and source of sensory information has been arbitrarily removed for purposes of assessment.

The finding of a strong relationship between audiovisual integration performance in perception (as measured by R) and speech intelligibility in some way suggests that the underlying factors that differentiate good CI users from poor CI users must be related to a common underlying set of phonological processing skills. These skills include perceptual, cognitive, and linguistic processes that are involved in the initial encoding, storage, rehearsal and manipulation of phonological representations of spoken words and the construction of sensory-motor programs for speech production and articulation. It is difficult to imagine any theoretical account of the present findings that is framed entirely in terms of peripheral sensory factors related to audibility without the additional assumption of some kind of phonological and/or lexical representation, used in conjunction with some type of linguistic mediation. To account for the present findings, it is necessary to assume the presence of some *underlying linguistic structure and process* which mediates between speech perception and speech production. Without a common underlying linguistic system - a grammar - these separate perceptual and productive abilities would not be so closely coordinated and mutually interdependent. It is well known and clearly documented in the literature on language development that reciprocal links exist between speech perception, production and a whole range of language-related abilities and skills. These links reflect the child's developing linguistic knowledge of phonology, morphology and syntax and his/her attempts to use this knowledge productively in a variety of expressive language tasks.

It needs to be emphasized strongly here that these children have been deprived of auditory experience for a substantial portion of their early lives before they received a cochlear implant. Some of these children are able to quickly learn to perceive and understand spoken language via electrical stimulation from their cochlear implant. Although they are delayed developmentally relative to their normal-hearing peers, these children appear to be making large gains in language development relative to other deaf children who have not received a CI. However, other children with CIs are not so fortunate and they seem to have much more difficulty making use of the sensory information provided by their implant. We believe that the study of individual differences in outcome and effectiveness of CIs is one of the most important research problems to investigate and understand over the next few years. If we had a better understanding of the reasons for these individual differences, we would be in a much better position to recommend changes in the child's language learning environment that were based on some theoretical motivation and set of operating principles. At the present time, decisions are made about intervention and therapy without a firm understanding of exactly why some children do well with their implant and why other children do more poorly.

The findings from the present analysis suggest that the gains in visual enhancement and the excellent audiovisual integration abilities observed in some of these children reflect the development and operation of their underlying linguistic systems. In this connection, several recent investigations have pointed to a close relationship between language and working memory. Pisoni and Geers (1998) reported that differences in working memory may be the “locus” of the wide range of variation observed in children with cochlear implants. Among other findings, they observed a strong positive correlation between auditory digit span and measures of performance on the Chive, a test of audiovisual integration that is similar to the Common Phrases Test used in the present analysis. In an analysis of data collected from a group of 43 eight- and nine-year old prelingually deafened children who had used their CI at least five years, Pisoni and Geers (1998) found that the WISC forward digit span was correlated with Chive V, a measure of lip-reading ability (r = +0.52) and Chive VE, a measure of visual enhancement abilities (r = +0.66). These findings were interpreted as support for the proposal that the large individual differences observed among children with CIs may be related to fundamental information processing operations in working memory, specifically, the operation of phonological working memory and the “phonological loop” which has been hypothesized as the primary rehearsal mechanism used to code and maintain the phonological representations of spoken words (Baddeley, Gathercole, & Papagno, 1998; Gathercole, Hitch, Service, & Martin, 1997).

Pisoni and Geers' (1998) findings on the role of working memory suggest that an important source of variance in outcome performance in these children has to do with the processing operations in working memory that a child uses to encode the sensory information he/she receives through the cochlear implant. The presence of a correlation between WISC auditory digit spans and visual enhancement due to lip-reading suggests that the locus of these effects may reside in a common representation format that is independent of the specific input modality. In some sense, it doesn’t matter if the sensory information is visual or auditory. What matters to the perceiver is whether the sensory input helps to specify the source of the underlying perceptual event, facilitating the recovery of the talker's articulatory gestures.

The results from Pisoni and Geers (1998) on auditory digit span and the present findings on AV integration are consistent with current theoretical conceptualizations of memory. Many theorists view short-term memory as simply that portion of long-term memory that is currently active (Anderson & Bower, 1973; Atkinson & Shiffrin, 1971; Engle, 1996) and closely related to attentional processes. As a consequence, long-term memory representations of spoken words and lexical knowledge play an important role in tasks that involve working memory. We suggest that the relationship observed in the current study between audiovisual integration and auditory-alone measures of speech perception reflects the contribution of fully specified, multimodal articulatory representations of speech in long-term

memory and knowledge of the cross-correlations between auditory and visual information about the same underlying articulatory events.

We speculate that the relationship between representational specificity and memory span arises because working memory plays a role in the formation of rich, multimodal representations in memory. Attention has been shown to play a large role in the implicit learning of artificial grammars (Nissen & Bullemer, 1987) and new verbal associations (Hartman, Knopman, & Nissen, 1989). Indeed, it is thought that the underlying statistical structure of patterns in the world facilitates the learning of those patterns (Reber, 1993; Stadler, 1993). If working memory is involved in the acquisition and internalization of statistical form in the world, then the relationships we have observed in this and other studies have a coherent explanation. Becoming a good integrator very likely entails perceptual learning and the internalization and representation of natural, lawful co-occurrences between disparate sensory information channels. Further work on this issue is needed and is currently being planned for our lab.

In summary, the present study uncovered relationships between the ability of children with cochlear implants to integrate cross-modal information about speech and their performance on auditory-alone measures of spoken word recognition. In addition, we found a close relationship between a child's audiovisual integrative performance and speech intelligibility. We conclude that these relationships point to the importance of building rich, multimodal long-term memory representations of speech that emphasize the amodal nature of speech information. Because these representations are necessary for a variety of speech tasks, performance on those tasks improves with complementary sensory inputs. We suspect that the most fruitful treatment programs for the speech perception and production of children with cochlear implants will involve emphasis on the vocal source of spoken language.

References

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Washington, DC: Winston.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225*, 82 - 90.

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

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (in press). Speech perception without hearing. *Perception & Psychophysics*.

Bollard, P. M., Chute, P. M., Popp, A., & Parisier, S. C. (1999). Specific language growth in young children using the Clarion cochlear implant. *Annals of Otology, Rhinology, & Laryngology, 108*(Suppl. 117), 119 - 123.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology, 43A*(3), 647 - 677.

Dowell, R. C., Martin, L. F. A., Tong, Y. C., Clark, G. M., Seligman, P. M., & Patrick, J. E. (1982). A 12-consonant confusion study on a multiple-channel cochlear implant patient. *Journal of Speech & Hearing Research, 25*, 509 - 516.

Engle, R. W. (1996). Working memory and retrieval: An inhibition-resource approach. In J. T. E. Richardson (Ed.), *Working memory and human cognition* (pp. 89 - 119). Oxford: Oxford University Press.

Erber, N. P. (1972). Auditory, visual and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech, Language, and Hearing Research, 15*, 413 - 422.

Fowler, C. A. (1989). Real objects of speech perception: A commentary on Diehl and Kluender. *Ecological Psychology, 1*(2), 145 - 160.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 17*(3), 816 - 828.

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M., & Gantz, B. J. (1992). Performance over time of congenitally deaf and postlingually deafened children using a multichannel cochlear implant. *Journal of Speech & Hearing Research, 35*, 913 - 920.

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 & Hearing Research, 40*, 183 - 199.

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.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*(5), 2677 - 2690.

Green, K. P., & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance, 21*(6), 1409 -1426.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 278 - 288.

Hartman, M., Knopman, D. S., & Nissen, M. J. (1989). Implicit learning of new verbal associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(6), 1070 - 1082.

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

Kirk, K. I. (1999). Assessing speech perception in listeners with conchlear implants: The development of the lexical neighborhood tests. *The Volta Review, 100*(2), 63 - 85.

Kirk, K. I., Diefendorf, A. O., Pisoni, D. B., & Robbins, A. M. (1997). Assessing speech perception in children. In L. L. Mendel & J. L. Danhauer (Eds.), *Audiologic evaluation and management and speech perception assessment* (pp. 101 - 132). San Diego: Singular Publishing.

Kirk, K. I., Eisenberger, L. S., Martinez, A. S., & Hay-McCutcheon, M. (1999). The Lexical Neighborhood Tests: Test-retest reliability and interlist equivalency. *Journal of the American Academy of Audiology, 10*, 113 - 123.

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.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant behavior and development, 7*, 361 - 381.

Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition, 21*, 1 - 36.

Logan, J. S. (1992). *A computational analysis of young children's lexicons* (Research on Speech Perception Technical Report No. 8). Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon, *Research on Speech Perception Technical Report No. 6* . Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear & Hearing, 19*, 1 - 36.

Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science, 4*(4), 104-109.

Massaro, D. W., & Cohen, M. M. (1999). Speech perception in perceivers with hearing loss: Synergy of multiple modalities. *Journal of Speech, Language, and Hearing Research, 42*, 21 - 41.

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 & Hearing, 20*, 363 - 371.

Miyamoto, R. T., Osberger, M. J., Robbins, A. M., Renshaw, J. J., Myres, W. A., Kessler, K., & Pope, M. L. (1989). Comparison of sensory aids in deaf children. *Annals of Otology, Rhinology, & Laryngology, 98*, 2 - 7.

Miyamoto, R. T., Svirsky, M. A., Kirk, K. I., Robbins, A. M., Todd, S., & Riley, A. (1997). Speech intelligibility of children with multichannel cochlear implants. *Annals of Otology, Rhinology, & Laryngology, 106*, 35 - 36.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*, 1 - 32.

Osberger, M. J., Miyamoto, R. T., Zimmerman-Phillips, S., Kernink, J. K., Stroer, B. S., Firzst, J. B., & Novak, M. A. (1991). Independent evaluation of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear & Hearing, 12*(Suppl.), 66S -80S.

Pisoni, D. B., & Geers, A. (1998). Working memory in deaf children with cochlear implants: Correlations between digit span and measures of spoken language processing, *Research on Spoken Language Processing Progress Report No. 22* (pp. 335 - 343). Blooomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D. B., Svirsky, M. A., Kirk, K. I., & Miyamoto, R. T. (submitted). Looking at the "stars": A first report on the intercorrelations among measures of speech perception, intelligibility and language in pediatric cochlear implant users. *Journal of Speech, Language, and Hearing Research*.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious.* New York, NY: Oxford University Press.

Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D., & Rubin, P. E. (1999). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication, 26*(1-2), 65 - 73.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101*(1), 129-156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947 - 950.

Robbins, A. M., Bollard, P. M., & Green, J. (1999). Language development in children implanted with the Clarion cochlear implant. *Annals of Otology, Rhinology, & Laryngology, 108*(Suppl. 117), 113 - 118.

Robbins, A. M., Svirsky, M. A., & Kirk, K. I. (1997). Children with implants can speak, but can they communicate? *Otolaryngology-Head and Neck Surgery, 117*, 155 - 160.

Rönnberg, J., Andersson, J., Samuelsson, S., Söderfeldt, B., Lyxell, B., & Risberg, J. (1999). A speechreading expert: The case of MM. *Journal of Speech, Language, and Hearing Research, 42*, 5 - 20.

Rosenblum, L. D. (1994). How special is audiovisual speech integration? *Current Psychology of Cognition, 13*(1), 110 - 116.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research, 39*, 1159 - 1170.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 22*(2), 318 - 331.

Stadler, M. A. (1993). Implicit serial learning: Questions inspired by Hebb (1961). *Memory and Cognition, 21*(6), 819 - 827.

Staller, S., Beiter, A. L., Brimacombe, J. A., Mecklenberg, D., & Arndt, P. (1991). Pediatric performance with the Nucleus 22-channel implant system. *American Journal of Otology, 12*, 126 - 136.

Sumby, W. H., & Pollack, I. (1954). Visual contribution of 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, NJ: Lawrence Erlbaum Associates.

Svirsky, M. A., Sloan, R. B., Caldwell, M., & Miyamoto, R. T. (1998). *Speech intelligibility of prelingually deaf children with multichannel cochlear implants.* Paper presented at the Seventh Symposuim on Cochlear Implants in Children, Iowa City, IA.

Tyler, R. S., Fryauf-Bertschy, H., Gantz, B. J., Kelsay, D. M. R., Tyler, R. S., Woodworth, G. G., & Parkinson, A. (1997a). Speech perception by prelingually deaf children using cochlear implants. *Otolaryngology-Head and Neck Surgery, 117*, 180 -187.

Tyler, R. S., Fryauf-Bertschy, H., Gantz, B. J., Kelsay, D. M. R., & Woodworth, G. G. (1997b). Speech perception in prelingually implanted children after four years. In I. Honjo & H. Takahashi (Eds.), *Cochlear implant and related science update (Advances in Otolaryngology 52)* (pp. 187 - 192). Basel: Karger.

Tyler, R. S., Parkinson, A. J., Fryauf-Bertschy, H., Lowder, M. W., Parkinson, W. S., Gantz, B. J., & Kelsay, D. M. R. (1997c). Speech perception by prelingually deaf children and postlingually deaf adults with cochlear implants. *Scandinavian Journal of Audiology, 26*(Suppl.46), 65 - 71.

Tyler, R. S., Tomblin, J. B., Spencer, L. J., Kelsay, D. M. R., & Fryauf-Bertschy, H. (in press). How speech perception through a cochlear implant affects language and education. *Otolaryngology-Head and Neck Surgery*.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzepoulos, D. (1997). The dynamics of audiovisual behavior in speech. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines* (pp. 221 - 232). Berlin: Springer-Verlag.

Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., & Yehia, H. (1996). *Characterizing audiovisual information during speech.* Paper presented at the International Conference on Spoken Language Processing, Philadelphia, PA.

Zimmerman-Phillips, S., Osberger, M. J., & Robbins, A. M. (1997). *Infant toddler meaningful auditory integration scale (IT-MAIS).* Sylmar, CA: Advanced Bionics Corporation.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

"Vowel Spaces" of Normal-Hearing and Hearing-Impaired Listeners with Cochlear Implants[1]

James D. Harnsberger, Mario A. Svirsky,[2] Adam R. Kaiser,[2]
Richard Wright,[3] and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

[1] This work was supported by NIH-NIDCD Training Grant DC00012, NIH-NIDCD Grant R01-DC00111, and NIH-NIDCD Grant R01-DC03937. We would like to thank Chris Quillet for his technical assistance.
[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, Indiana; Department of Biomedical Engineering, and Department of Electrical Engineering, Purdue University.
[3] Now at the University of Washington, Seattle, WA.

"Vowel Spaces" of Normal-Hearing and Hearing-Impaired Listeners with Cochlear Implants

Abstract. Cochlear implant users show substantial individual differences in their ability to understand speech in general, and vowels in particular. One possible reason for these differences lies in their widely different abilities to identify formant frequencies. Another possible reason is that cochlear implants present spectral information to cochlear locations that are more basal than normal. The latter explanation has been controversial. Some authors have proposed that the spectral mismatch introduced by cochlear implants may be completely overcome by cochlear implant users (Rosen et al., 1999), while others believe that spectral mismatch may result in important limitations to speech perception, no matter how much time is used by the cochlear implant users to adapt to the new percepts. In the present study, we designed a vowel perception test, using a Method-of-Adjustment (MOA) paradigm, to compare the vowel spaces of eight cochlear implant users to those obtained from 43 normal hearing listeners with the same dialect as the cochlear implant users. The MOA vowel test consisted of a set of 330 steady-state synthetic stimuli arranged in a two-dimensional grid, generated by varying the first and second formants of the vowels. Subjects were asked to label and rate on a seven-point scale those stimuli which matched ten visually-presented vowel stimuli corresponding to /i/, /ɪ/, /e/, /ɛ/, /æ/, /ʌ/, /ɑ/, /u/, /ʊ/, and /o/. Two-dimensional plots of subjects' responses for all ten target stimuli constituted the "vowel spaces" of the subjects. With one exception, no systematic shift was observed across all ten vowel categories in the vowel spaces of cochlear-implant users, suggesting that these subjects were able to adapt completely to the spectral shift introduced by the implant. However, the cochlear-implant users' spaces differed substantially from normal vowel spaces in terms of the relative size of the vowel categories, and their location in perceptual space.

Introduction

Although cochlear implants allow profoundly deaf people to hear, cochlear implant users show a very wide range of speech perception skills. The most successful cochlear implant users can easily hold a face-to-face conversation, and they can even communicate on the telephone, a difficult task because there are no visual cues available and because the signal itself is highly degraded. On the other hand, the least successful cochlear implant users have a difficult time communicating even in a face-to-face situation, and can barely perform above chance on auditory-alone speech perception tasks. It is important to remember that electrical hearing as provided by a cochlear implant is quite different from normal acoustic hearing. One important difference lies in the ability to discriminate formant frequencies. For example, Kewley-Port and Watson (1994) report difference limens between 12 and 17 Hz in the F1 frequency region for normal hearing listeners. In the F2 frequency region, they found a frequency resolution of approximately 1.5 percent. In cochlear implant users, discrimination of formant frequencies is dependent on two factors: the frequency-to-electrode map that is programmed in their speech processor, and the individual's ability to discriminate stimulation pulses delivered to different channels. It is not uncommon for some cochlear implant users to have formant frequency difference limens that are one order of magnitude higher than those of listeners with normal hearing, or even more. It is reasonable to hypothesize that cochlear implant users with such limited frequency discrimination skills will find it quite difficult to identify vowels accurately because formant frequencies are important cues to vowel identity. Another important difference between acoustic and electric hearing is related to the fact that cochlear implants do not stimulate the entire

neural population of the cochlea but only the most basal 25 mm at best. Therefore, cochlear implants stimulate cochlear locations that are more-basal-than-normal and thus elicit higher frequency percepts than normal acoustic stimuli. For example, when the input speech signal has a spectral peak at 300 Hz, the neurons stimulated in response to this signal may have characteristic frequencies of 1000 Hz or even higher. This represents a rather extreme modification of the peripheral frequency map, resulting in more-basal-than-normal stimulation. To the extent that CI users have enough plasticity in the auditory nervous system to successfully "re-map" the place frequency code in the cochlea, the more-basal-than-normal stimulation provided by a CI should not hinder speech perception. On the other hand, an inability to re-map the place frequency code may severely limit speech perception in CI users and may be an important source of individual differences in speech perception.

Several studies have addressed the issue of adaptation to changes in frequency-to-electrode assignments for cochlear implant users. Skinner et al. (1995) showed that users of the SPEAK stimulation strategy identified vowels better with a frequency-to-electrode table that mapped a more restricted acoustic range into the subject's electrodes than the default frequency-to-electrode table. The experimental table that resulted in better vowel perception represented a more extreme shift in spectral information than the default table, suggesting that listeners with cochlear implants can indeed adapt to such shifts, at least within certain limits. Another study that demonstrates the adaptation ability of human listeners in response to spectral shifts was conducted by Rosen et al. (1999), who used acoustic simulations of the information received by a cochlear implant user who has a more basal spectral shift of 6.5 mm in the basilar membrane (equivalent to 1.3-2.9 octaves, depending on frequency). Initially, the spectral shift reduced word identification (1% correct, as compared to 64% for the unshifted condition), but after only 3 hours of training, subjects' performance improved to 30% correct. This result raises the question of what may have been the maximum performance of a listener who had a better chance to reach asymptotic levels. Recently, Shannon et al. (1999) performed an experiment in which the frequency-to-electrode tables of three cochlear implant users were shifted one octave with respect to the table they had been using daily for at least three years. It is important to note that this one-octave shift was in addition to the original shift imposed by the cochlear implant. After three months of experience with the new table, it was apparent that adaptation was not complete because, on average, subjects did not reach the same levels of speech perception that they had achieved before the table change. Taken together, these studies show that auditory adaptation to a modified frequency map is possible but it may be limited, depending on the size of the spectral shift.

In the present study, we investigated the issue of spectral shift with a new paradigm, a method of adjustment (MOA) procedure. This procedure was used to map the perceptual vowel spaces of adult, postlingually deafened cochlear implant users. Similar tasks have been used with normal hearing listeners (Johnson et al., 1993). In this task, subjects select the region of the F1-F2 plane that sounds (to them) like a given vowel, and the procedure is repeated for ten English vowels. This task gives us the opportunity to simultaneously assess a cochlear implant user's plasticity, by comparing the locations of his/her selected regions to those selected by normal hearing listeners, and his/her frequency discrimination skills, by examining the spread of his/her selected regions. More specifically, a listener who was unable to adapt to the basalward spectral shift introduced by their cochlear implant would select regions that are systematically shifted to lower frequencies with respect to the regions of the vowel space selected by normal hearing listeners. In addition to the MOA task, the ability of cochlear implant users to identify synthetic vowels that differed only in F1 was measured, as well as their ability to identify natural vowels. One long-term goal of our research is to understand the mechanisms that underlie speech perception by cochlear implant (CI) users and, in so doing, gain an understanding of the individual differences in psychophysical characteristics which may explain individual differences in speech perception with a CI. With the combination of tasks employed in the present study we hope to be able to tease out the effect of

two kinds of limitations on vowel perception by cochlear implant users: frequency discrimination and auditory plasticity.

Methods

Participants

Forty-three normal-hearing Indiana University undergraduates and eight cochlear-implant (CI) users, all monolingual speakers of English, participated in this experiment. The normal-hearing participants consisted of 20 males and 23 females ranging in age between 18 and 28, none of who reported any history of a speech or hearing problem. The normal-hearing participants were recruited to represent the dialect of American English spoken in central Indiana with a common inventory of vowels. Only normal-hearing listeners who reported living their entire lives in central Indiana were allowed to participate in this experiment. Central Indiana was defined in terms of a 60 mile radius around Indianapolis, roughly covering the Midland dialect as described by Wolfram and Schilling-Estes (1999), and avoiding two other regional dialects found at the northern and southern extremes of the state. These latter two regional dialects are reported to differ from the Midland dialect in terms of vowel quality and degree and type of dipthongization (Labov, 1991). For participating in two 1 hour sessions, the participants received either $7.50/hour or two credits towards their research requirement if they were enrolled in an undergraduate psychology class.

The CI user participants were recruited from the population of patients served at the Department of Otolaryngology-Head and Neck Surgery at the Indiana University School of Medicine in Indianapolis. The demographics of the CI users are given in Table 1. All of the CI users had received implants at least one year prior to participating in this study. Five were users of the Nucleus-22 device with the SPEAK strategy, while three were users of the Clarion device with the CIS strategy. The SPEAK strategy (Skinner et al., 1994) filters the incoming speech signal into up to 20 frequency bands, which are associated with different intracochlear stimulation channels. Typically, six channels are *sequentially* stimulated in a cycle that is repeated 250 times per second. The channels to be stimulated each cycle are chosen based on the filters with the highest output amplitude. In contrast, the CIS strategy (Wilson et al., 1991) as implemented in the Clarion device filters the signal into eight bands, one for each stimulation channel. All channels are sequentially stimulated with pulses whose amplitudes are determined by the filters' outputs. The stimulation cycle is repeated at a fast rate of at least 833 times per second. The main differences with the SPEAK strategy are that CIS uses a higher stimulation rate, fewer stimulation channels, and that CIS stimulates all channels in a cycle rather than choosing the channels with more energy within the corresponding filter passbands. The CI users' age ranged from 37 to 67, averaging 59.

Subject	Age (Years)	Age at Onset of Deafness	Age at Implantation	Implant Use (Years)	Gender	Implant Type	Insertion Depth
CI1	67	43	61	6	F	Clarion 1.0	-
CI2	35	29	31	3	M	Clarion 1.2	-
CI3	37	34	36	1	M	Nucleus 24	Full (5)
CI4	74	27	71	2	F	Nucleus 22	Full (7)
CI5	63	56	57	5	M	Nucleus 22	Full (4)
CI6	70	*	66	3	M	Nucleus 22	Full (7)
CI7	68	*	65	2	F	Nucleus 22	Full (8)

CI8	58	*	52	6	M	Clarion 1.0	-

Table 1: Demographics of subjects with cochlear implants. Insertion depth and the number of stiffening rings not inserted are noted in parentheses. *CI4, CI8 Progressive; CI5, Progressive childhood.

Stimulus Materials

Method-of-Adjustment Task. The stimulus set consisted of 330 synthetic, steady-state vowels in isolation, generated by a Klatt synthesizer, that varied from one another in their first and second formants, in equal increments of 0.377 Bark (Klatt & Klatt, 1990). The Bark increment size was chosen as a close approximation of the just-noticeable-difference for vowel formants of Flanagan (1957). The F1 and F2 values for this stimulus set ranged between 2.63 Z (250 Hz) - 7.91 Z (900 Hz) and 7.25 Z (800 Hz) - 15.17 Z (2800 Hz). These ranges were chosen to represent the full range of possible values for speakers of American English, and were successfully used in an earlier method-of-adjustment study of vowel perception (Johnson et al., 1993), as well as in piloting. All of the other synthesis parameters for this stimulus set also followed Johnson et al. (1993). The values, or the formulas for calculating the values, of the most relevant synthetic parameters are summarized in Table 2. The f0 parameter was varied to generate two sets of the 330 stimuli, one representing a male voice and one representing a female voice. All of the synthetic sounds were calibrated to a 70 dB listening level.

Parameter	Value or Formula
Duration	250 ms
f0	Male: 120 Hz over the first half, dropping to 105 Hz at the end Female: 186 Hz over first half, dropping to 163 Hz at the end
F3	{(0.522*F1) + (1.197*F2) + 57} or {(0.7866*F1) - (0.365*F2) + 2341}[4]
F4	3500 Hz or (F3+300 Hz), whichever higher
B1	29.27 + (0.061*F1) - (0.027*F2) + (0.02*F3)
B2	-120.22 - (0.116*F1) + (0.107*F3)
B3	-432.1 + (0.053*F1) + (0.142*F2) + (0.151*F3)
B4	200 Hz

Table 2: The values or formulas used for an important subset of the parameters used in generating the synthetic stimulus sets.

Vowel Identification Task. The vowel identification task is a closed-set task that uses 9 vowels in an "h-vowel-d" format. The stimuli were digitized from the female vowel tokens of the Iowa laserdisc (Tyler et al., 1987). Only the steady state vowels (i.e., not the diphthongs) were used. There were three separate productions of each vowel. Listeners were administered three lists that consisted of five repetitions of each vowel and three practice tokens.

F1 Identification Task. The stimuli for this experiment were seven synthetic three-formant vowels, with an F2 value of 1500 Hz and an F3 of 2500 Hz. F1 varied linearly, from 250 Hz for stimulus 1 to 850 Hz for stimulus 7. Three-formant vowels rather than single-formant vowels were used in order to measure

[4] The first formula was used for the half of the grid with higher F2 values, while the second was used for the half of the grid with lower F2 values.

place-pitch discrimination with more realistic and speech-like stimuli. The stimuli were created using the Klatt 88 (Klatt & Klatt, 1990) speech synthesizer software. Voicing amplitude increased linearly in dB from zero to steady state over the first 10 ms, and back to zero over the last 10 ms of the stimulus. Steady-state amplitude was loudness balanced within 1 dB for the seven stimuli. Total stimulus duration was 1 second. The stimuli were digitally stored using a sampling rate of 11025 Hz at 16 bits of resolution and were presented from an Intel® based PC equipped with a SoundBlaster compatible sound card. Stimuli were presented at a level of at least 70 dB C weighted SPL over an Acoustic Research loudspeaker. Custom task specific software was used to present stimuli and record responses.

Procedures

Method-of-Adjustment Task. The procedures varied slightly for each participant group in the study. Normal-hearing participants were tested in a quiet room in two 1-hour sessions, with the second session taking place approximately one week after the first session. In the first session, normal-hearing participants completed the method of adjustment task with one of the synthetic stimulus sets, either the male or the female voice set. In the second session, participants completed the method of adjustment task with the remaining stimulus set. The experiment was balanced for the order in which to the stimulus sets were presented to the normal-hearing participants. In contrast, the CI users were tested in a quiet room or a sound-attenuated chamber, in a single test session, varying in length by individual CI user between 1 to 3 hours. Given the length of time CI users required to complete the MOA task and other demands on their time, they were presented with only one of the two stimulus sets, the male-voice set.

Each participant was presented with a two-dimensional (15 rows and 22 columns) visual grid centered in a computer screen. The grid consisted of the 330 synthetic stimuli described above. A single word appeared above this grid, constituting the target stimulus for a given trial. The visual target stimulus for a given trial was one of ten words, “heed,” “hid,” “aid,” “head,” “had,” “who'd,” “hood,” “owed,” “odd,” and “hut,” each of which contained one of the 10 vowels under study, /i/, /ɪ/, /e/, /ɛ/, /æ/, /ʌ/, /ɑ/, /u/, /ʊ/, and /o/. Subjects were be instructed to search the grid, playing out individual sounds until they found the region of the grid that played out synthetic sounds that matched the vowel in the visual target stimulus. After selecting one or more synthetic sounds that matched the target, subjects were asked to give each synthetic sound a rating on a 1 - 7 scale, grading how close a match the synthetic sound was to the target. The order of presentation of each target stimulus varied randomly from participant to participant, with a single repetition of a stimulus set presented to listeners.

The particular stimuli chosen and their respective ratings were used to calculate category “centers” and sizes for each vowel type for each listener group. Category centers were determined by averaging, in the F1 and F2 dimensions, across all selected stimuli, with their contribution to the average weighted by their rating. Category sizes in each formant dimension were based on the standard deviation of the formant mean. Normal-hearing listeners’ category centers were expected to appear in F1 X F2 space in a similar arrangement to that observed in vowel production studies with American English (i.e., Peterson & Barney, 1952). The category centers of CI patients were expected to deviate from those of normal-hearing listeners depending on the extent of their more basal stimulation, which would result in an overall space that is shifted lower in F1 and F2.

Vowel Identification Task. The vowel identification task was a closed-set speech perception task in which three separate productions of each of ten /hVd/ tokens were presented in random order, one at a time, and the CI users had to say which one of the ten stimuli they thought they heard by responding

verbally.[5] All subjects heard a total of at least 15 presentations of each vowel (except CI5 who heard 10), and were instructed to guess if they did not know which vowel was presented. The subjects' responses were tabulated and scored for total percentage of correct responses.

F1 Identification Task. The F1 identification task was an absolute identification task, with the stimuli labeled "1" through "7" in order of increasing F1. In this task, all stimuli were played in sequence several times, so the subjects could become familiar with the stimuli. Then, the stimuli were presented ten times each in random order and subjects were asked to identify the stimulus that was presented. The subject's response and the correct response were displayed on the computer monitor before moving on to the presentation of the next stimulus. After each block of 70 presentations (ten presentations of each of seven stimuli), the mean and standard deviations of each of the seven stimuli were calculated. The d' for each pair of successive stimuli was calculated as the difference of the two means divided by the average of the two standard deviations. These d' measurements were then cumulated to calculate a cumulative d' curve, which provided an overall measure of the subject's ability to discriminate and pitch rank the seven stimuli. To calculate the cumulative d' curves, we followed the assumption that the maximum possible value of d' was 3. The average JND (defined as the mean stimulus difference resulting in d' = 1) was calculated based on the cumulative d' curve. Given that the F1 range spanned by the seven stimuli was 600 Hz, the JND was defined as (600/cumulative d'). At least eight blocks of 70 presentations were carried out, as these were sufficient for all subjects to reach a plateau in performance, as measured by the cumulative d'. The cumulative d' reported here is the average of the best two blocks for each subject.

Results

Normal-Hearing Participants

The normal-hearing listener group was expected to select vowel categories centers with first and second formant values that corresponded to those typical in vowel production in American English. Figure 1 shows the mean vowel categories for all of the normal-hearing subjects, for the male-voice stimulus set. Each category center is shown along with error bars denoting two standard deviations from the category centers in both formant dimensions. In this figure, all of the ratings have been used to calculate the center and size of all ten vowel categories. Figure 2 shows the vowel spaces of normal-hearing subjects calculated using only ratings of four and above. The rating of four was chosen because it was the highest rating that still allowed for category sizes to be calculated for all ten vowels of all of the normal-hearing and CI participants. The center of each category was determined by averaging the Bark values of all synthetic stimuli that were chosen in each dimension, with the means weighted by the ratings given to individual stimuli.

The perceptual spaces shown in Figures 1 and 2 demonstrate that the method-of-adjustment technique for measuring vowel categories was successfully used to generate vowel spaces that bear the typical intervowel relationships that have been observed in F1 X F2 spaces generated from vowel production data. For instance, front vowel category centers have a higher F2 than back vowel centers; high vowel category centers have a lower F1 than low vowel category centers. While no vowel production studies have been published for English speakers from central Indiana,[6] Hillenbrand, Getty, Clark, and Wheeler (1995) examined the vowel production spaces of 45-48 men, women, and children who were

[5] If a subject's answer was unclear to the experimenter, subjects were asked to respond by pointing to the answer on a test sheet.

[6] However, work is in progress in our laboratory on a vowel production study of Central Indiana English, using as talkers the same normal-hearing listeners that participated in this study.

native speakers of American English, specifically the variety spoken in southern Michigan. Hillenbrand et al. (1995) was a replication and extension of the classic work on the acoustics of American English vowels carried out by Peterson and Barney (1952). Figures 3 and 4 are plots of the vowel production centers from these two studies, respectively, both including the MOA vowel category centers (calculated using ratings of 4 and above). While there are differences between the absolute locations of the MOA vowel centers and the vowel production centers of the two studies, all three vowel spaces show a common set of F1 and F2 distances between vowels. The first and second formants from the normal-hearing listeners were significantly correlated with their counterparts in both the Hillenbrand et al. (1995) set (r = .98, $p \leq 0.01$ for F1; r = .89, $p \leq 0.01$ for F2) and the Peterson and Barney (1995) set (r = 1.0, $p \leq 0.01$ for F1; r = .83, $p \leq 0.05$ for F2).

Cochlear Implant Patients

The results of the MOA, F1 identification, and vowel perception tasks for the eight CI users are shown in Figures 5 - 12. Each figure shows the labeled centers of the ten vowel categories, with the size of each category in each dimension indicated via error bars. The centers and sizes of the categories were computed via a mean in each dimension which was weighted by the rating given to each synthetic stimulus, using ratings of only four or above, just as in the normal-hearing listeners' vowel space in Figure 2. To the right of each CI user's vowel space, the results of his/her F1 identification (JND_{F1}) and vowel perception tests (VOWEL) are listed.

An examination of all eight vowel spaces reveals little evidence of any systematic shift due to a lack of adaptation to more-basal-than-normal stimulation, with the possible exception of CI1. Instead, we find individual CI user spaces that differ from the normal space in terms of the sizes of perceptual categories, their degree of overlap, and the region of perceptual space that particular categories occupy. These observations are supported by measures of the differences between CI and normal vowel categories. Table 3 lists the absolute differences, in Bark, between the category centers of both normal and individual CI users, in both F1 and F2. Positive differences indicate that the formant of the normal-hearing listeners for that vowel was greater than the formant of the CI patient. If a shift were observable with a particular CI user, one would expect to see positive differences in one or both of the formants of the vowels of that user. Of the eight CI users, seven showed positive and negative differences, depending on the vowel and formant in question. Only one subject, CI1, systematically lowered formants for his/her category centers, as would be predicted for a listener who has not adapted completely to the spectral shift introduced by the cochlear implant (see Figure 5). We can see that CI1's categories were shifted towards the lowest formant values in the upper right corner, resulting in a particularly compressed space. However, the magnitude of this shift was not the same for all ten vowels, or for both formants. The magnitude of the shift varied from 0.28 Z-3.53 Z, with on average a greater shift observed in F1 than F2.

Name	F	Vowel Category										Mean
		i	ɪ	e	ɛ	æ	ʌ	ɑ	u	ʊ	o	
CI1	F1	0.43	1.06	1.79	2.37	3.53	0.62	2.82	0.45	1.32	1.52	1.59
	F2	0.28	0.8	1.09	1.46	1.46	0.39	0.57	1.41	2.35	0.66	1.05
CI2	F1	0.05	-1.7	0.02	0.14	-0.86	-1.16	-0.93	-0.52	0.07	-0.92	-0.58
	F2	-0.13	0.31	-0.77	-1.13	0.52	-0.19	0.57	0.24	1.97	-4.1	-0.27
CI3	F1	-0.62	-0.57	-1.47	-1.65	-0.75	0.62	-0.15	0.2	0.09	1.15	-0.32
	F2	0.27	0.2	0.06	0.29	1.16	0.64	0.43	-1.29	1.54	0.21	0.35

CI4	F1	-0.47	0.39	-0.35	0.08	0.09	-0.44	1.04	-0.11	0.89	0.63	0.18
	F2	0.24	-1.3	-0.59	-0.81	-0.53	-0.57	-1.83	0.46	-0.6	0	-0.55
CI5	F1	-2.01	-2.91	-0.01	-0.89	-0.31	0.29	-0.1	-0.55	-0.69	0.12	-0.71
	F2	0.05	2.08	-0.1	0.13	4.31	-0.38	-0.26	-1.07	-0.14	-2.07	0.26
CI6	F1	-2.03	-1.23	-1.16	-1.2	0.44	0.51	0.6	-0.14	-0.65	0.86	-0.4
	F2	0.3	-0.45	0.38	-1.04	0.9	0.75	-1.08	0.1	0.6	-0.38	0.01
CI7	F1	-2.09	-1.74	-0.94	-0.27	1.97	0.67	1.49	-1.25	-0.62	-0.49	-0.33
	F2	2.57	0.15	1.33	1.35	2.73	-0.27	-2.09	-2.49	-1.01	-2.15	0.01
CI8	F1	-2.18	-0.6	-0.06	-0.05	0.65	-0.08	0.93	-2.22	-1.28	-1	-0.59
	F2	3.42	1.23	2.72	1.1	1.85	-1.89	-2.47	-2.71	-2.34	-2.93	-0.2

Table 3: The differences in Bark between normal and CI vowel categories, for each formant (F), for individual vowel categories, and the mean difference across all categories.

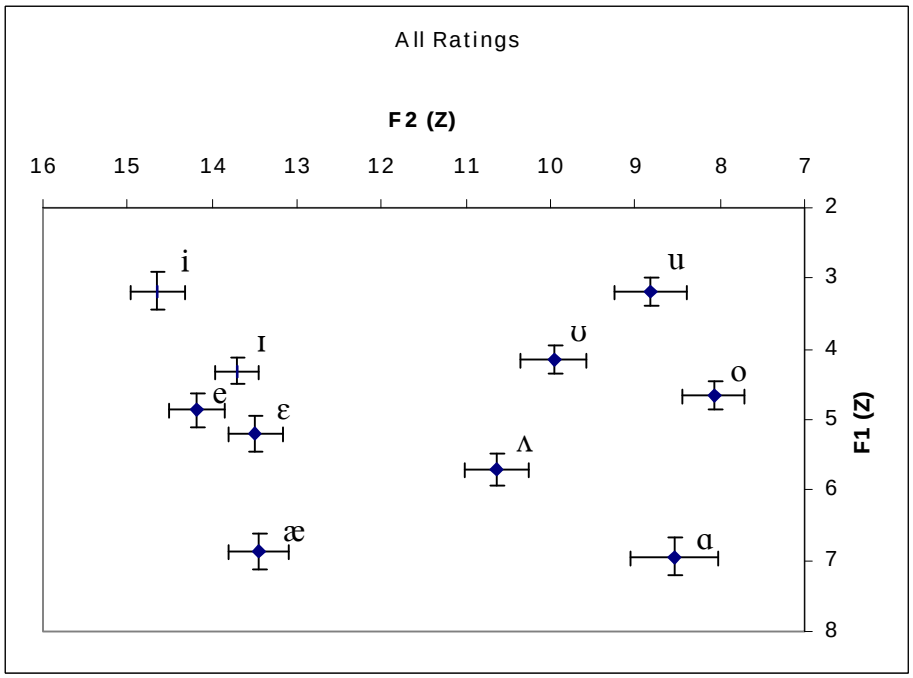


Figure 1: The mean vowel space of normal-hearing listeners, calculated using all of the ratings provided.

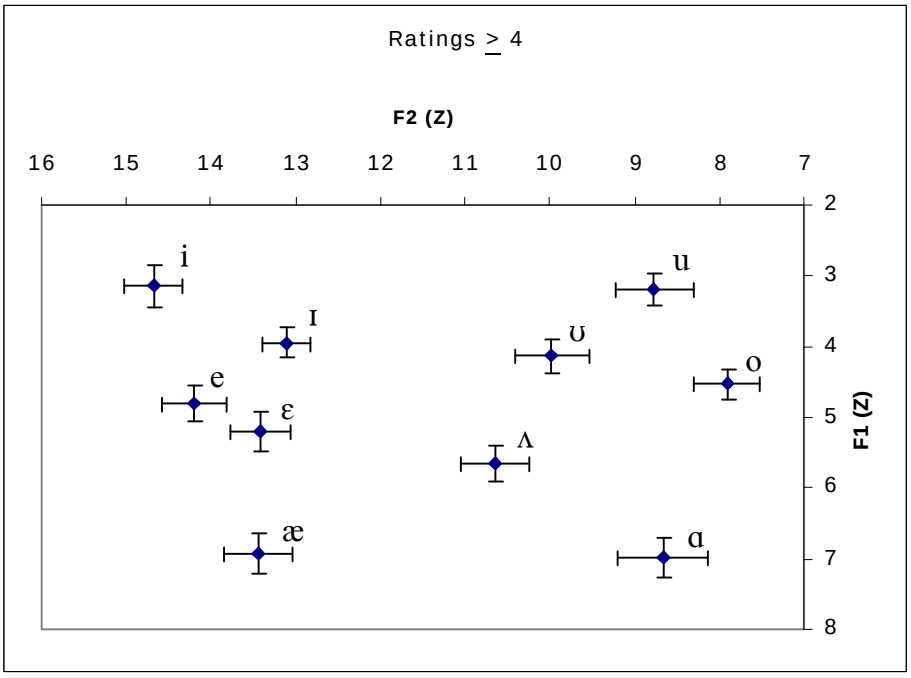


Figure 2: The mean vowel space of normal-hearing listeners, calculated using only ratings of four or above.

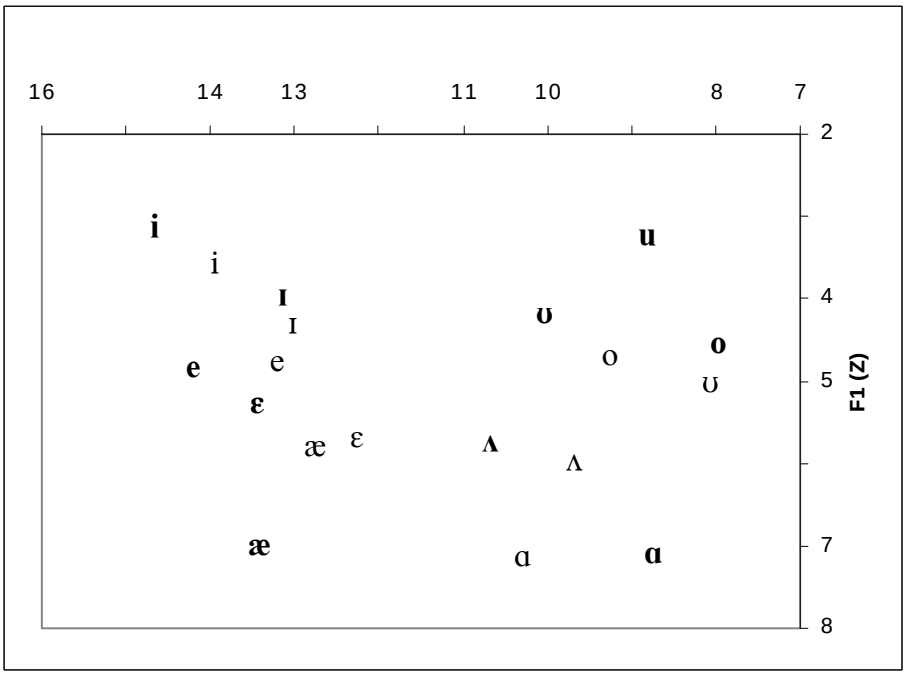


Figure 3: A comparison of the vowel category centers from the MOA task (in bold) with the vowel production centers (in plain text) from Hillenbrand et al. (1995).

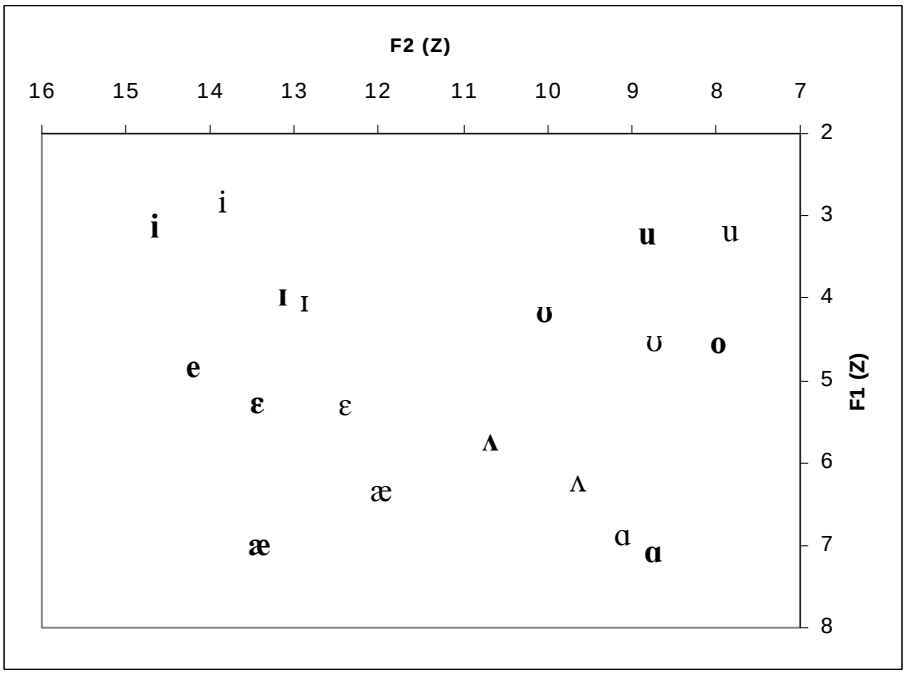


Figure 4: A comparison of the vowel category centers from the MOA task (bold) with the vowel production centers (plain text) from Peterson & Barney (1952).

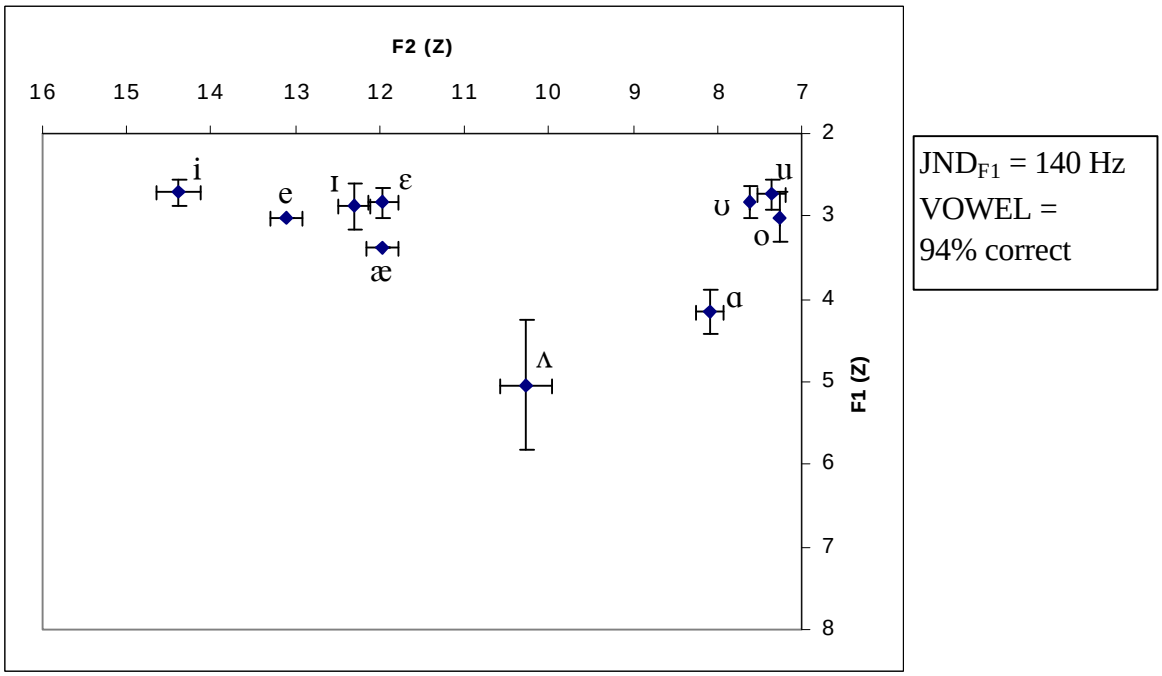


Figure 5: The vowel space of subject CI1.

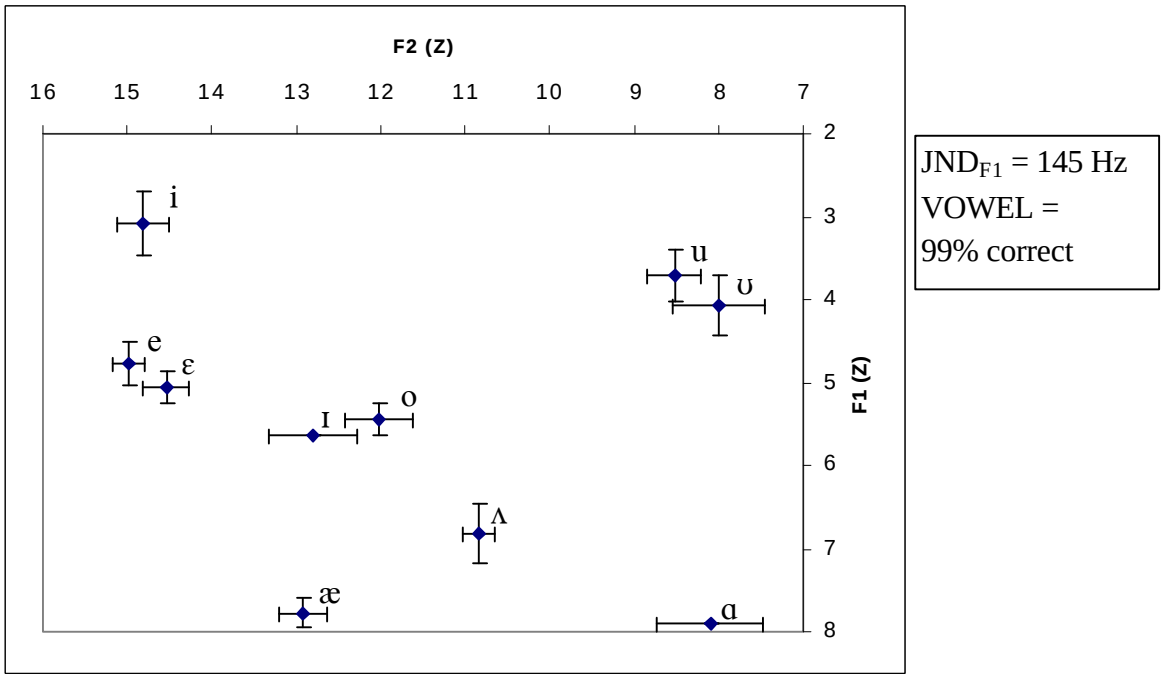
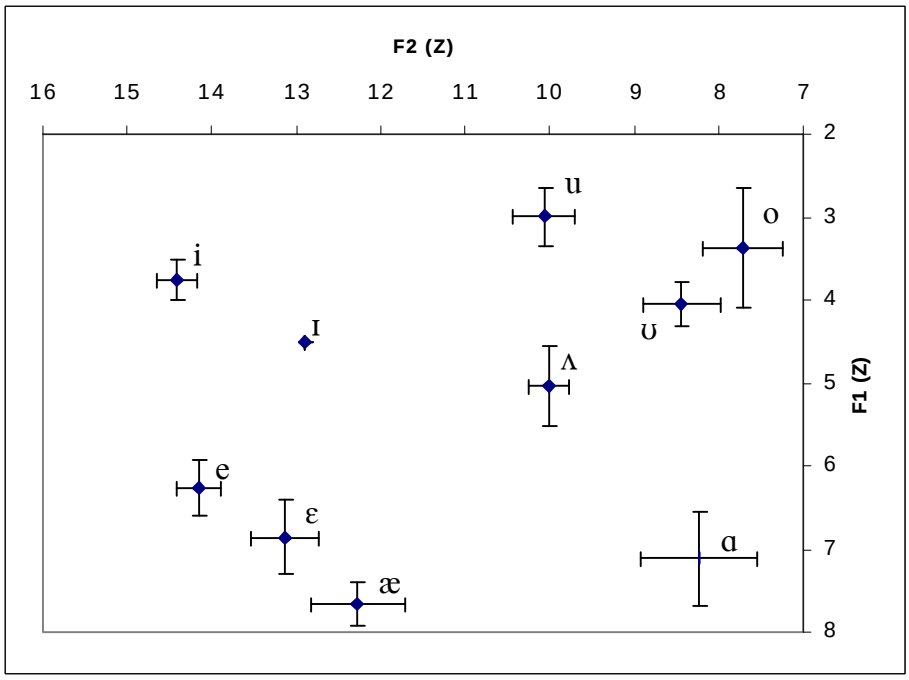


Figure 6: The vowel space of subject CI2.



JND_{F1} = 140 Hz
VOWEL =
71% correct

Figure 7: The vowel space of subject CI3.

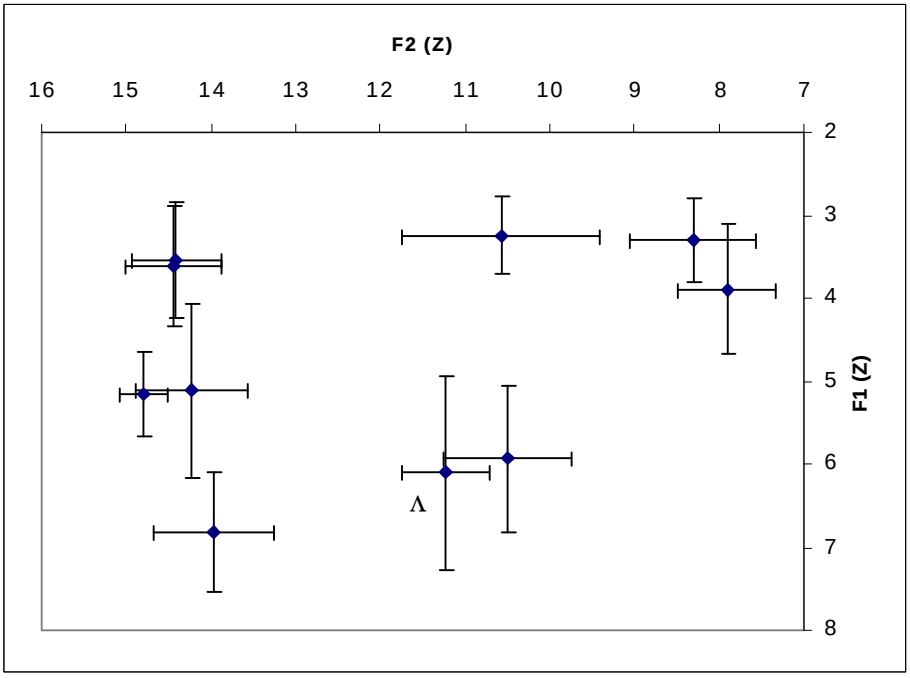
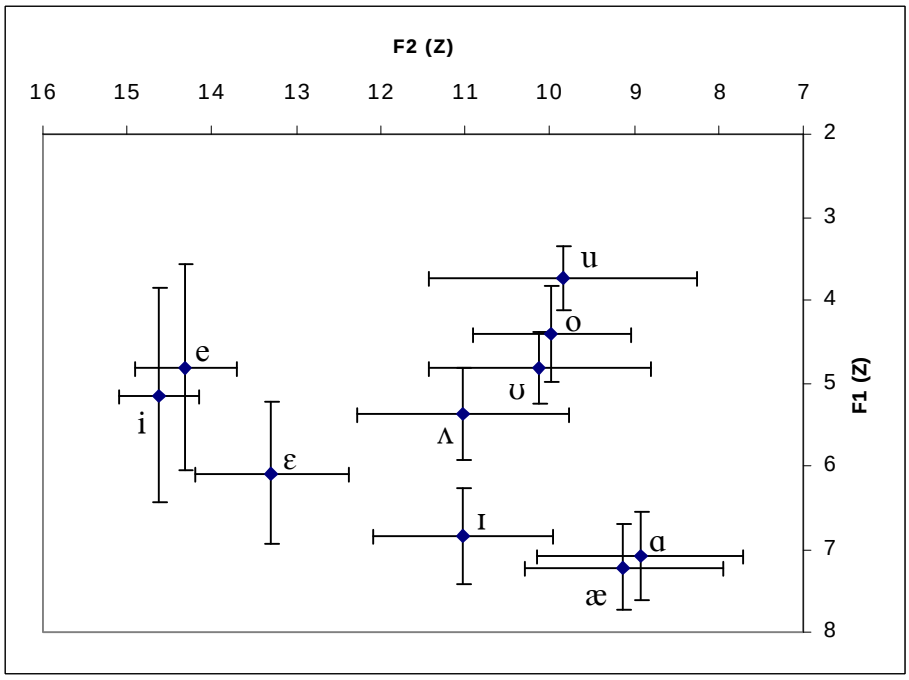
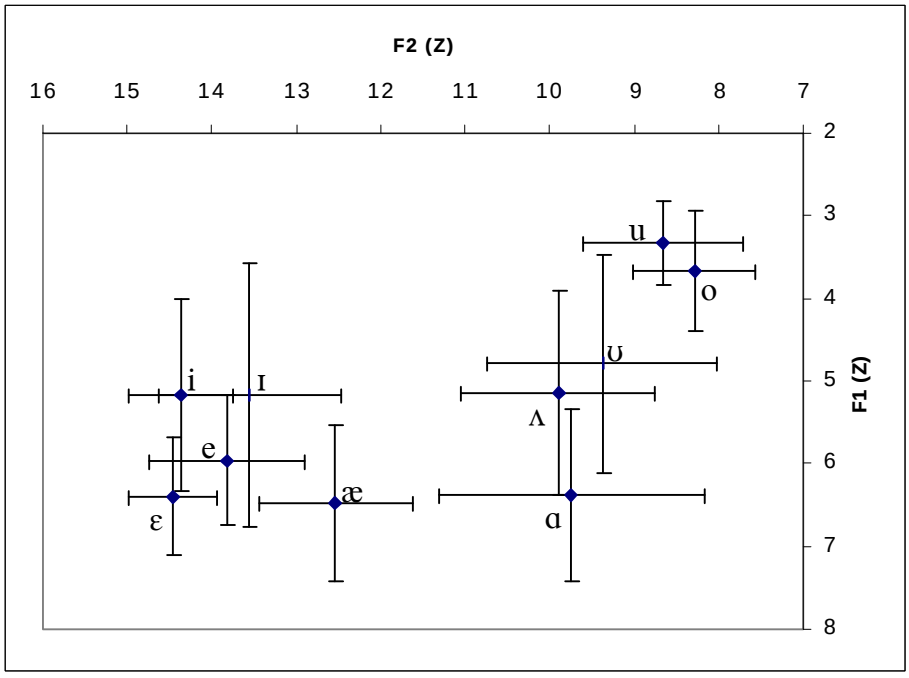


Figure 8: The vowel space of subject CI4.



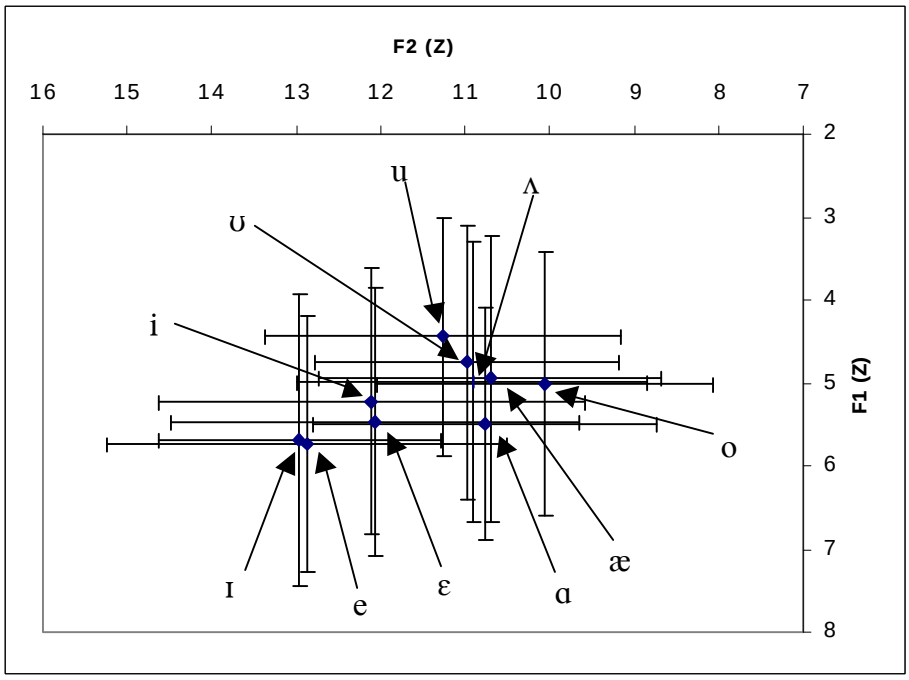
JND_{F1} = 135 Hz
VOWEL =
34% correct

Figure 9: The vowel space of subject CI5.



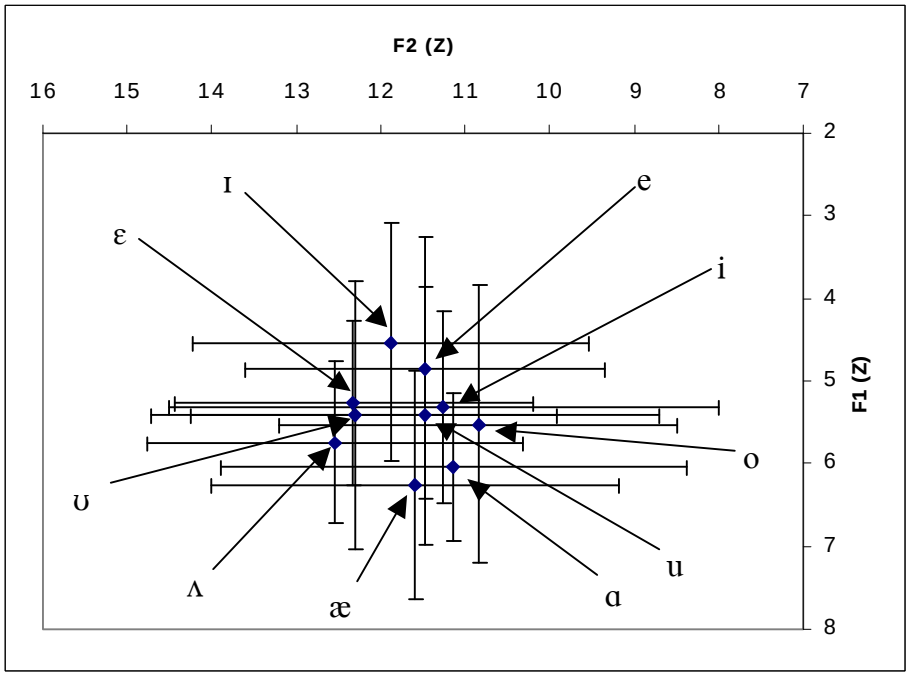
JND_{F1} = 250 Hz
VOWEL =
70% correct

Figure 10: The vowel space of subject CI6.



JND_{F1} = 350 Hz
VOWEL =
23% correct

Figure 11: The vowel space of subject CI7.



JND_{F1} = 550 Hz
VOWEL =
22% correct

Figure 12: The vowel space of subject CI8.

The results indicate that most CI users adapted to their frequency shifted input, although they varied widely in the degree to
vowel spaces, we find several that are composed of categories that overlap very little, and that appear to be in roughly the same regions in perceptual space as those obtained from
of such “normal” CI spaces are CI2 and CI3. In their spaces, front and back vowels occupy their “normal” regions of perceptual space. The sizes of the categories in these spaces are not much greater than those of hearing listeners. In contrast, the vowel spaces for CI7 and CI8 show much larger categories with a great deal of overlap between front and back vowels. The size of these categories and their arrangement in ave great difficulty in using spectral information to discriminate among vowel sounds, or in identifying particular vowels in running speech. The spaces of the cation and size.

which gives the mean sizes of categories (in Bark) for each CI space, in the F1 and F2 dimension, along with the mean sizes averaged over both F
across all CI users (“CI”) and the equivalent scores for all normal- differences between the CI users and the normal- n the order of a 3:1 ratio, and all were significant at the p _ 0.01 level (F1: *t* *df* = 81.1; F2: = 6.9, *df* F2 combined: *t* *df* = 163.6). The vowel spaces of CI2 and CI3, the relatively “normal” spaces, have ory sizes (along with CI1), while CI7 and CI8 have the highest. Large vowel categories typically overlap with one or more neighboring categories in perceptual space, which is he results of the F1 identification and vowel identification tests are consistent with this hypothesis: the subjects whose vowel result is reflected in Spearman rank correlations between the category size measures and the F1 correlated significantly ($p \leq 0.05$), while the vowel identification test corr p _ 0.05) with all three measures. All of the other correlations were only marginally significant.

Name	F1 (Z)		F1 and F2 (Z)
CI1		0.17	0.2
	0.23	0.37	
CI3	0.37		0.37
CI4		0.65	0.7
	0.7	1.06	
CI6	1.01		0.99
CI7		2.1	1.85
	1.33	2.48	
Mean CI	0.78		0.9
Mean NH		0.37	0.3

Table 4: The mean size of individual CI user's categories in the F1 dimension, the F2 dimension, and both dimensions, along with the mean category sizes across all normal-hearing (NH) and CI users' categories.

	Measure					
Test	F1		F2		F1 and F2	
	r	*p*	*r*	*p*	*r*	*p*
JND F1	0.78	0.04	0.63	0.1	0.72	0.06
VOWEL	-0.87	0.02	-0.93	0.01	-0.92	0.01

Table 5: Correlation between the category size measures and the F1 and vowel identification tests.

Discussion

Despite the very large individual differences observed among cochlear implant subjects in all three tests used in this study, the results revealed an orderly relationship between the vowel spaces of these subjects and their scores on the F1 identification and the vowel identification tests. In the introduction we discussed two different potential limitations to vowel identification by cochlear implant users: limited frequency discrimination and limited ability to adapt to more-basal-than-normal stimulation. Overall, the major problem in vowel identification appears to be the former. Only one of the eight CI users, CI1, showed a systematic shift of her vowel space that seemed consistent with limited auditory plasticity. For each vowel, she selected regions with lower F1 and F2 than those selected by normal hearing listeners. This result is consistent with the proposal that she has not completely adapted to more-basal-than-normal stimulation and thus selects vowels with very low formants as the best exemplars, to compensate for the frequency shift that is imposed by the cochlear implant.

Other than CI1, none of the CI subjects showed any systematic shifts in their vowel spaces, suggesting that they were able to adapt to the frequency shift introduced by the cochlear implant. However, limitations in frequency discrimination played an important role in these subjects' ability to identify vowels. For example, CI7 and CI8 were the two CI users with the poorest JND_{F1} values: 350 Hz and 550 Hz respectively. Their vowel spaces showed substantial overlap among most vowel regions and their vowel identification scores were the lowest among the CI user group, 22% and 23% correct. Conversely, the CI users with the best (smallest) JND_{F1}'s, such as CI1 and CI2, tended to have high vowel identification scores and little overlap among vowel regions.

Our results are consistent with those of Hawks and Fourakis (1998). Their cochlear implant subjects also showed widely divergent amounts of overlap among vowel categories that were mapped using an identification task with synthetic stimuli. In terms of adaptation to more-basal-than-normal stimulation, the present results are encouraging because they suggest that most cochlear implant users are able to adapt to the shifts typically introduced by their devices. However, CI1's results remind us that not all listeners may be able to adapt in a similar way, and perhaps more listeners would find it difficult to adapt if the spectral shift was greater than that of the cochlear implant subjects in this study.

In conclusion, the data reported in this paper strongly suggest that vowel perception by cochlear implant users may be limited by the listener's formant frequency discrimination skills, in combination with

his/her ability to adapt to more basal- -normal stimulation. The present findings are also relevant to the

over time) can provide much information about consonant identity. In future studies we intend to explore -basal than-
to investigate the nature of the improvement in speech perception observed in most postlingually deaf

applying the same methods used in this study to prelingually deafened CI users. In this case, we would not expect to see major
prelingually deafened CI users may be, on average, even worse than that observed in postlingually deafened CI users. Finally, it may be interesting to determine whether the list
have less neural and behavioral plasticity than the other listeners, or whether their lack of adaptation is due to a shallower insertion, which would result in a greater spectral shift to be overcome by the listener. Ta
together, these studies will help us explain the enormous individual differences in speech perception by CI users, possibly paving the way for improved devices and intervention strategies in this clinical population.

References

Flanagan, J. (1957). E
vowel sounds. *Journal of the Acoustical Society of America 29*, 533 534.

Hawks, J.W. and Fourakis, M.S. (1998). Perception of synthetic vowels by cochlear implant recipients. , *104 (3* -5.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American *Journal of the Acoustical Society of America*, , 3099-

on, K., Flemming, E., and Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. , *69* -528.

Kewley Port, D. and Watson, C. S. (1994). Formant-
Journal of the Acoustical Society of America 95, 485 496.

Klatt D. H., & Klatt L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. 87, 820-

dialects of English. In P. Eckert (Ed.), *New ways of analyzing sound change* pp. 1-

Journal of the Acoustical Society of America 24, 175 184.

Rosen, S., Faulkner, A., and Wilkinson, L. (1999) Adaptation by normal listeners to upward spectral shifts *Journal of the Acoustical Society of America*, , 3629 3636.

Shannon, R.V, Fu, Q.- L., Chatterjee, M., Wygonski, J., Galvin, J., Zeng, F.-
Wang, X. (1999). Speech processors for auditory prostheses, *Quarterly Progress Report #2* 10.

Skinner, M.W., Clark, G.M., Whitford, L.A., Seligman, P.M., Staller, S.J., Shipp, D.B., Shallop, J.K., Everingham, C., Menapace, C.M., Arndt, P.L., Antogenelli, T., Brimacombe, J.A., Pijl, S., Daniels, P., George, C.R., McDermott, H., & Beiter, A.L. (1994). Evaluation of a new spectral peak coding strategy for the Nucleus 22 channel cochlear implant system. *American Journal of Otology, 15* (Suppl. 2), 15-27.

Skinner, M.W., Holden, L.K., and Holden, T.A. (1995). Effect of frequency boundary assignment on speech recognition with the SPEAK speech-coding strategy. *Annals of Otology, Rhinology, & Laryngology – Supplement, 166*, 307-11.

Tyler, R. S., Preece, J. P., & Lowder, M. W. (1987). The Iowa audiovisual speech perception laser video disc. *Laser Videodisc and Laboratory Report*, University of Iowa at Iowa City, Department of Otolaryngology - Head and Neck Surgery.

Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K. and Rabinowitz, W.M. (1991) Better speech recognition with cochlear implants. *Nature*, vol. 352, 18 July 1991, pp. 236-238.

Wolfram, W. and Schilling-Estes, N. (1998). *American English*. Malden, MA: Blackwell Publishers.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

A Real Time PC Based Cochlear Implant Speech Processor with an Interface to the Nucleus 22 Electrode Cochlear Implant and a Filtered Noiseband Simulation[1]

Adam R. Kaiser and Mario A. Svirsky[2]

DeVault Otologic Research Laboratory
Department of Otolaryngology-Head & Neck Surgery
Indiana University School of Medicine
Indianapolis, IN 46202

[1] This work was supported by NIH/NIDCD Training Grant DC-00012, RO1-TC03937 and by grants from the Deafness Research Foundation and the AHRF. Earlier versions of this work were reported at the 1999 Conference on Implantable Auditory Prostheses.

[2] The authors make no claim about the suitability or safety of the device described herein for any purpose, nor do they assume any liability for the use of the information about this interface. Individual researchers must evaluate and develop protocols to ensure subject safety according to all applicable laws and guidelines.

A Real Time PC Based Cochlear Implant Speech Processor with an Interface to the Nucleus 22 Electrode Cochlear Implant and a Filtered Noiseband Simulation

Abstract: Cochlear implants are electronic devices that have enabled individuals with severe to profound hearing losses to regain some hearing. Nearly all of those who receive cochlear implants (CIs) regain the sensation of sound. There is, however, substantial variability in speech perception performance among users of cochlear implants (Staller et al., 1997). One factor that may contribute to the individual differences in performance is the speech processing strategy used. Traditionally, researchers implemented experimental real time strategies on specialized digital signal processors (DSPs). This approach requires specialized programming knowledge and consequently can be expensive. In this respect, we have simplified the development of real time processing strategies by using an IBM compatible laptop computer to perform all signal processing, and have implemented only a few interface functions on a DSP. Both the continuous interleaved sampling (CIS) and n-of-m speech processing schemes were implemented on a PC in C++. In addition to encoding and transferring speech cues to a cochlear implant, the processor described can also implement an acoustic noiseband based model of a cochlear implant for presentation to listeners with normal hearing.

Introduction

Cochlear implants are electronic auditory prostheses that have enabled individuals with severe to profound hearing losses to regain the sensation of sound. They perform this function through a process beginning with the reception of an acoustic sound wave and ending in the electronic stimulation of the cochlea. First, a microphone is used to receive sound and to convert it to an electrical representation. Second, in multi-channel speech processors, this analog representation is converted to a more easily manipulated digital form. Thirdly, the speech processor encodes and transforms the digitized speech according to the strategy(s) implemented on a particular device. In general, these strategies use this digitally filtered signal to assign a stimulation level to electrode(s) implanted in the cochlea. The stimulation level and the electrode on which it is to be used are encoded and sent via a radio frequency link to a receiver implanted beneath the subject's skin. The implanted device referred to as a receiver/stimulator decodes the information and finally directs the delivery of an electrical stimulus. The research platform described here performs the functions that are normally performed by the external speech processor and enables a researcher to have complete control over this process. The interface described is specific to the Nucleus 22 electrode cochlear implant system, but it can be readily adapted to other platforms as well.

Improvements in speech processing strategies have resulted in increased speech perception performance for users of cochlear implants (Eddington, 1980; Loeb & Kessler, 1995; Kieffer et al., 1997; Wilson et al., 1991). Inherent in the development of such strategies is the need to evaluate subject performance using the new signal processing strategies and stimulation techniques. The complexity of digitizing sound, processing it, tailoring stimulation for a specific patient, and finally encoding the information in a format compatible with a subject's particular CI device is a formidable task. For some purposes researchers have chosen to evaluate new strategies by processing speech off line, generating files containing information that will be used to stimulate patients at a later date (for one example see Fu & Shannon, 1999). This methodology enables complete algorithm flexibility and relieves the experimenter from implementing a real time system; however, it does not allow the subject to practice using the new processing strategy in conversation, nor does it allow the subject to adjust parameters (such

as volume) in real time. For other purposes, researchers have chosen to implement a real time system on a dedicated DSP platform. A summary of such interfaces can be found in (Eddington et al., 1998). This method also enables complete flexibility and real time parameter adjustments but requires a detailed knowledge of device programming. The platform described here allows both complete algorithmic flexibility and real time parameter adjustment. In addition, since the speech processing algorithms are implemented in C++ on a PC, they are more easily adapted for specific signal processing research applications by less specialized programmers and engineers than required for DSP-based development.

While the bulk of speech processing and patient-specific tailoring can be performed on a personal computer, one cannot avoid implementing functions to communicate with the subjects implanted receiver/stimulator in specialized hardware. For the Nucleus 22 device, only four communication functions are needed to implement a wide variety of stimulation strategies. These functions set the stimulation rate and inter pulse interval, direct a stimulation pulse at a given stimulation level to be sent to a specific electrode, and lastly relay interface status back to the PC. Buy using off the shelf components to the greatest extent possible, we have attempted to simplify the construction of this interface. With the exception of a radio frequency pulse generator, the described interface is implemented in software from off the shelf components.

Materials

Figure 1 depicts a block diagram of the hardware used to implement the processor. The core of the Windows 98 PC based platform is based on a DELL Inspiron 3500 laptop equipped with an Intel 350 MHz PII processor. It has 64 Megabytes of RAM, a 4.5 Gigabyte hard drive, and a sound card capable of implementing DirectX 7.1 sound digitization and sound presentation functionality. In preliminary work, the SoundBlaster Live Value (Model SB4670) has been used for this purpose on a desktop PC. Additionally, a Motorola DSP56309EVM evaluation module was used to drive a radio frequency pulse generator according to the communication protocol required of the Nucleus 22 receiver/stimulator. The DSP and PC communicate via a serial cable between a PC com port and the JTAG/OnCE port on the DSP. This connection is needed during program upload and initiation. Real time communication, however, is maintained through a Domain Technologies, Inc. Host Interface Adapter (DSP563xxEVM Host Port Interface Type B). This adapter allows fast bi-directional communication via the host PC's printer port, which must be configured in EPP mode at address 0x378, and the DSP's host port.

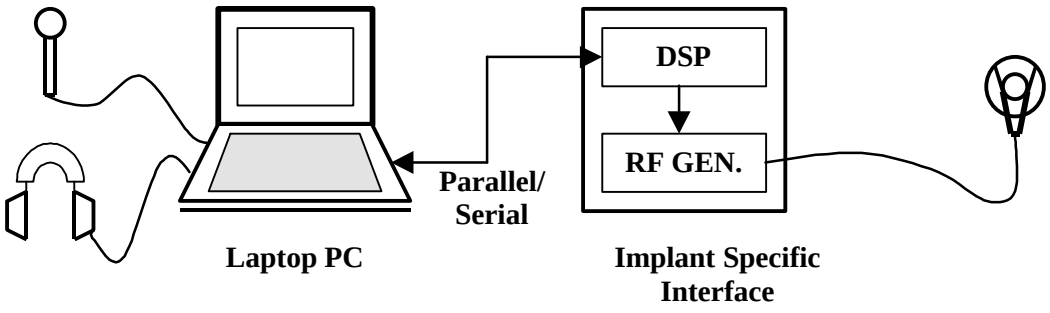


Figure 1: A microphone or alternatively any analog sound source is connected to the PC via the sound card. The PC performs all A/D conversions and DSP routines, leaving the implant-specific interface to generate the RF pulses transferred to the receiver stimulator. The implant-specific interface is initialized via the PC's serial port. After initialization, commands that set the stimulation rate, inter pulse interval, and stimulation parameters are sent via the PC's parallel port. The PC may also request the status of the DSP's internal FIFO buffer to ensure proper stimulation timing.

Several software packages were used to implement the interface. The Microsoft Visual C++ compiler was used in conjunction with the runtime version of the Microsoft DirectX 7.1 software development kit. This software was used to implement all of the speech processing algorithms. Matlab version 5.2.1 was used to calculate filter parameters for the digital filters. Two programs were used to develop, debug and load software on the Motorola DSP evaluation board. Domain Technologies, Inc. EVM563xx version 3.0 software was used to load and monitor the program that implements the implant specific interface on the DSP. The implant specific interface program code was written in assembly language and compiled using the Motorola DSP56300 Assembler Version 6.2.0. PC side communications to the host port adapter were implemented using the interface functions printed in the DSP563xxEVM Host Interface Adapter Reference Manual. No additional drivers were used for this purpose.

Methods

In an effort to keep the PC platform general and to maintain flexibility, several general guidelines were followed during development. First, the interface was designed to allow the easy implantation of the CIS and n-of-m signal processing strategies in addition to developing new ones. Second, researchers without DSP programming experience should be able to modify the programs to suit their own purposes. Third, all speech processing should be implemented on the PC leaving only FIFO (first –in, first- out buffer) maintenance and communication protocol implementation to the implant specific interface. Fourth, the device itself should be the limiting factor in performance, not the interface. Fifth, subject specific information should be loaded from a patient-specific file at runtime.

The resulting platform consists of two main functional sections. The first section to be described is the personal computer on which we implemented of the speech processing algorithms. Each step in the signal processing pathway will be discussed in order from input to output. There are several parameters that may be configured in a configuration file similar to the example in appendix A that is automatically loaded at runtime. These parameters will be referred to in ALL CAPS. When they are specified in the parameter file they must be followed by the appropriate parameter(s) in the following line. Following the description of the PC side algorithms, the functionality of the implant specific interface will then be outlined.

The steps in signal processing performed with this platform are summarized in Figure 2. Signal processing begins by using the sound card to sample the data at 22,050 Hz at 16 bits of resolution. Care should be taken to utilize as much of this dynamic range as possible. The data is then converted to single precision floating point. Pre-emphasis is performed using an N = 1 Butterworth filter with Fc = 1200 Hz. This filter tends to decrease the prominence of the low frequency components of the incoming signal and has been implemented in a variety of research interfaces at this and other institutions (Eddington, 1998; Shannon, 1999). The Fc is set in software and is not yet configurable in the parameter file.

The next step is to perform automatic gain control / dynamic range compression (AGC/DRC). The algorithm used here is based heavily on techniques developed by researchers at MIT (Eddington et al., 1993). It is beyond the scope of this paper to fully describe the operation of this section. Basically, this step normalizes the overall volume of the incoming signal. First, an estimate of the envelope of the incoming signal is made based on a full wave rectified version of the incoming signal. The envelope of the signal is then adjusted based on the previous estimate of the envelope and the current rectified signal. When the rectified signal is greater than the previous envelope estimate, the envelope estimate is increased to the level of the rectified signal at a rate specified by the single time constant AGCATTACK. When the rectified signal is less than the previous envelope estimate, the envelope estimate is decreased to the level of the rectified signal at a rate specified by the single time constant AGCRELEASE. Both constants are specified in seconds in the parameter file. Once the current envelope estimate has been calculated, a gain is selected such that the current envelope estimate multiplied by the gain follows the

piecewise linear function described by the points in DRCTABLE. The values specified in this table are listed as input/output pairs in dB (20*log(amplitude)). They are referenced to a full scale input of zero dB. Full scale is determined by the input to the analog to digital converter. At a minimum, the parameter file must specify an input/output pair for and input of 0 dB and for –100 dB. Pairs should be listed in by decreasing input volume. The output sample from this stage is equal to the input sample multiplied by the gain determined using the relationship above.

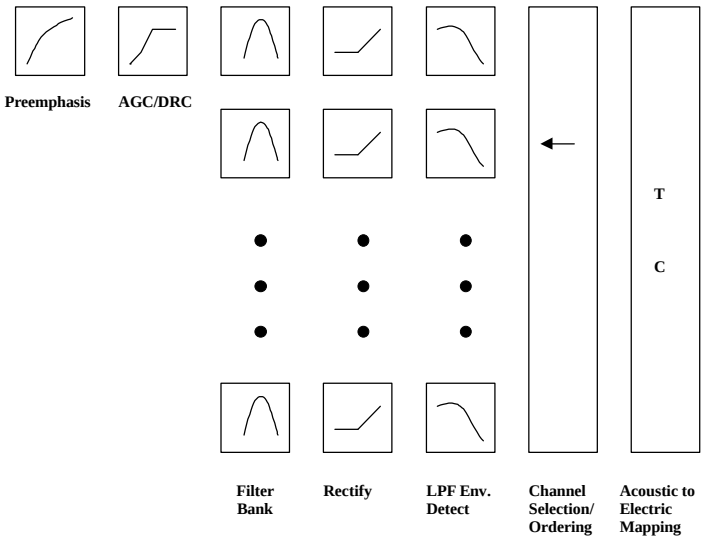


Figure 2: Signal Processing Steps. Data is sampled at 22,050 Hz and 16 bits of resolution prior to conversion to single precision floating point. Preemphasis and AGC/DRC is then performed, followed by frequency-specific filtering and strategy-specific channel selection. The resulting data is scaled logarithmically and mapped to the patient's threshold and comfort levels.

A filter bank with NUMCHAN channels is next used to divide the incoming signal into its corresponding frequency components. These filters are N=4 infinite impulse response (IIR) filters implemented in a single direct form II stage. The characteristics of the filter are specified by the filter coefficients following the words IIRFILTERTAPS in the specification file. An example Matlab program to generate and save the filter coefficients in the correct order appears in Appendix B. The frequency characteristics of the resulting filter are specified by the coefficients listed in the following order for each consecutive channel: a_0, a_1, a_2, a_3, a_4, b_0, b_1, b_2, b_3, b_4. These coefficients correspond to the transfer function in equation 1. Note that a_0 must be unity.

$$H(z) = \frac{\sum_{k=0}^{4} b_k z^{-k}}{\sum_{k=0}^{4} a_k z^{-k}}.$$

Equation 1: Filter description for a single channel of an N channel implementation.

Once filtered, the resulting signal is half wave rectified and subsequently low pass filtered using an N = 2 IIR filter with coefficients listed immediately after the IIROUTPUTLPF flag in the following order a_0, a_1, a_2, a_4, b_0, b_1, b_2. This filter is also implemented in a single Direct Form II stage. These coefficients correspond to the transfer function in equation 1. Note that a_0 must be unity in this case as well.

The processing steps up to this point are identical for both the CIS and the n-of-m processing schemes. The following step of channel selection, however, is dependent upon which PROCESSORTYPE is specified in the configuration file. When this parameter equals 1, CIS is selected. When this parameter is equal to 2, n-of-m is selected. CIS simply sequences through each of the channels one at a time and, as will be described further, selects the appropriate electrode and patient-specific stimulation level at which to stimulate. When n-of-m is selected, the channel with the most energy in its corresponding input frequency band is selected from those that have not been stimulated in the last NFORNOFM-1 stimulation cycles. The NFORNOFM parameter defines the number of maxima to be used in the n-of-m stimulation scheme. Since signal processing is performed at 22,050 Hz and the total STIMULATIONRATE is typically 1250 pulses per second, on average, only one in STIMULATIONRATE/22,050 samples streaming from the LPF filters is used in the channel selection process.

The platform we have developed also has the functionality to measure the minimum detectable level (T), and the maximum comfortable level (C). This option can be selected by specifying PROCESSORTYPE = 0 to execute this function. Stimulation levels can be sequentially tried by selecting the appropriate channel and either the T or C level using the left and right arrow keys. The level can be adjusted by using the up and down arrow keys. A one half second stimulation at the rate specified by the STIMULATIONRATE will be delivered to the corresponding electrode with each press of the F12 key. Care must be taken when setting the stimulation rate. For example, creating a map for use with a 20 channel n-of-m processor with 6 maxima requires a stimulation rate of 250Hz during T and C adjustment but a rate of 6*250=1250Hz when actually using the n-of-m processing scheme. One must correctly adjust STIMULATIONRATE to reflect the maximum rate that can be expected on any one channel for each speech processing scheme.

The last step in the PC portion of signal processing maps the amplitude of the channel selected during the channel selection stage to an appropriate simulation level. The numbers immediately following the ELECTRODETANDC flag in the specification file define the mapping parameters for each channel in a single row. The parameters specify the electrode, mode, T level, C level, acoustic minimum, and finally an acoustic maximum in that order. The electrode simply refers to the actual electrode on the implanted receiver/stimulator. The mode refers to the mode of stimulation. For common ground stimulation, a mode of 0 should be selected. For bipolar stimulation a mode of 1 should be selected. For bipolar + 1 stimulation the mode should be 2. The remaining modes follow in this pattern. Both the T and C levels specify the amplitude and phase duration of the stimulation pulses according to the 6.40b stimulation level specification.[3] The acoustic minimum specifies the acoustic level of the low pass filter (LPF) output in dB (20*log(filteroutput)) that will cause stimulation at the T level. The acoustic maximum specifies the acoustic level which will cause stimulation at the C level. Stimulation above the C level is prevented, but stimulation at levels as low as T/2 are simply extrapolated from T and C in conjunction with the acoustic maximum and minimum. If a stimulation pulse is specified for a simulation level below T/2 a level of T/2 is presented to maintain power to the implant.

There are several controls other than T and C adjustments that the experimenter may make in real time. First, F1 and F2 respectively decrease or increase the volume of the signal in increments of 5% of

[3] For a complete specification contact Cochlear Corporation, 61 Inverness Drive East, Suite 200, Englewood, CO 80112 U.S.A.

each electrodes dynamic range. This is accomplished by adjusting the C level by the fraction of the dynamic range between the T and the C level specified by the volume. For example, a volume of .75, a T level of 100, and a C level of 200 would result in an effective C level of 175. Second the acoustic value that is to be mapped to the T level can be increased with F4 and decreased with F3. This control has the effect of increasing the dynamic range of acoustic signals mapped to electrical stimulation, and therefore volume, with each press of F4. The converse is true for F3. The last control is a sensitivity adjustment that maintains the range of acoustic levels mapped to electrical stimulation but changes both the acoustic maxima and minima simultaneously. F7 effectively decreases the volume by increasing the acoustic maxima and minima while F8 does the opposite. All of these controls operate across all of the channels simultaneously. Only the T and C level adjustments are made on a channel by channel basis.

While this program is primarily intended to implement speech processing strategies for users of cochlear implants, it can also be used to generate a filtered noiseband simulation of these strategies for listeners and researchers with normal hearing. Rosen (1999) and others have used similar models in their research. The first five steps in the CI speech processor and the filtered noiseband simulation are equivalent. The processes differ beyond the low pass filters as diagramed in Figure 3. The remaining steps begin by using the channel selected during the scheme specific channel selection process. If the algorithm has chosen to stimulate a particular channel, the amplitude of the LPF output is sampled and used to modulate a stream of white noise. This same amplitude sample is used to modulate the amplitude of the white noise for a period of time equal to 1/STIMULATIONRATE in the case of CIS, or NFORNOFM/STIMULATIONRATE in the case of the n-of-m processing strategy. If the channel is not selected after this time, then its amplitude is set to 0. The next stage filters the streaming periods of noise and/or silence using the identical filters specified in the input filterbank stage. The resulting output is then summed with the output from the remaining channels and streamed to the sound card as output.

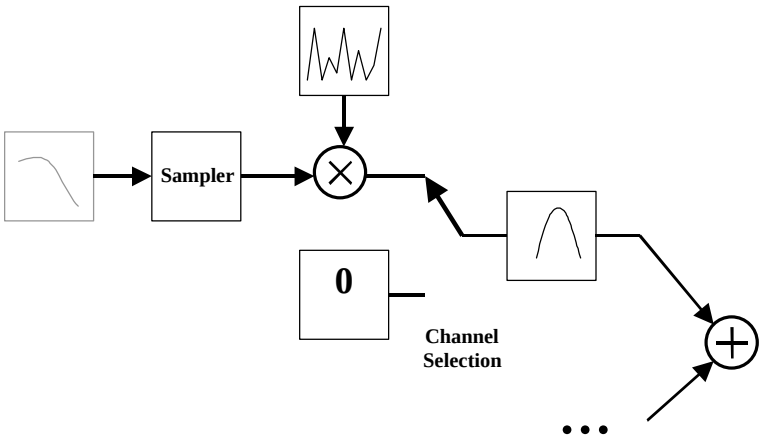


Figure 3: Filtered noiseband simulation. The input to the filter bank for each channel is based on white noise modulated with a sampled version of the LPF output. This sample is updated at a rate based on the simulation strategy and simulation rate.

The second functional component of the platform is implemented in the implant specific interface. After the PC has processed the acoustic speech signal and has determined the appropriate electrode and stimulation level to use, the signal is then sent via the parallel port to the implant specific interface. The following four function prototypes describe the functions that have been developed to facilitate the transfer of information and timing.

1.void Nuc22SendStimPulse(unsigned int elect, unsigned int mod, unsigned int amplitude, unsigned int phase)

As implied by its name, this function sends stimulation information to the interface. The electrode is simply the electrode to be stimulated, and the mode is specified as described above. The amplitude specifies the current level, and must be between 1 and 240. The length of each phase of the signal is specified as phase*.4μs and should be between 12 and 1024.

2.void Nuc22SetStimRate(unsigned int rate)

This function sets the pulse rate in Hz. The interface produces pulses at this rate. The end of the second of the biphasic pulse is used as the timing mark. For pairs of pulses that take longer than 1/rate to transmit, the interface will keep track of how far it has fallen behind, and will catch up during the transmission of the subsequent stimulation pulses. Clearly, the average length of time for a pulse to be sent must be less than 1/rate, or the interface cannot keep up, and the stimulation rate will not be met. This is a limitation of the receiver/stimulator and not of the interface.

3.void Nuc22SetIPI(unsigned int ipi)

Nuc22SetIPI sets the length of the inter phase interval as ipi*.4μs. In general this value should be approximately 10μs.

4.unsigned int Nuc22FIFORemaining(unsigned int slots)

This function returns the number of locations left to be filled in the internal FIFO buffers. When streaming from a data file, this function can be used to determine how many pulses can be sent to the FIFO before it overflows. Samples sent to the FIFO when it is full will be discarded.

These basic functions allow the implementation of a variety of stimulation strategies, whether they are implemented from preprocessed stimuli or used by a real time program such as such as the one described here.

Future Directions

The PC based speech processing algorithms described in this report are completely independent of the receiver/stimulator to which they will ultimately interface. Work is underway in our laboratory to interface the PC based signal processing platform described here to other cochlear implant systems.

References

Eddington, D.K. (1980). Speech discrimination in deaf subjects with cochlear implants. *Journal of the Acoustical Society of America*, 68 (3), 885-891.

Eddington, D.K., Rabinowitz, W.M., Svirsky, M.A. & Tierney, J. (1993). Speech Processors for Auditory Prostheses (Fourth quarterly progress report, NIH Contract N01-DC-2-2402) National Institutes of Health, Bethesda, MD: Neural prostheses program.

Eddington, D.K., Garcia, N., Noel, V., Tierney, J., & Whearty, M. (1998). Speech Processors for Auditory Prostheses (Final report, NIH project N01-DC-6-2100). National Institutes of Health, Bethesda, MD: Neural prostheses program.

Fu, Q.J., & Shannon, R.V. (1999). Effects of electrode configuration and frequency allocation on vowel recognition with the Nucleus-22 cochlear implant. *Ear & Hearing*. 20(4), 332-44.

Kiefer, J., Muller, J., Pfenningdorff, T., Schon, F., Helms, J., von Ilberg, C., Baumgartner, W., Gstottner, W., Ehrenberger, K., Arnold, W., Stephan, K., Thumfart, W., & Baur, S. (1997). Speech understanding in quiet and in noise with the CIS speech-coding strategy (MED EL Combi-40) compared to the MPEAK and SPEAK strategies (Nucleus). *Advances in Oto-Rhino-Laryngology, 52*, 286-290.

Lawson, D.T., Wilson, B.S., Zerbi, M., & Finley, C.C. (1996). Speech processors for auditory prostheses (Third quarterly progress report, NIH project N01-DC-5-2103). National Institutes of Health, Bethesda, MD: Neural prostheses program.

Loeb, G.E., & Kessler, D.K. (1995). Speech recognition performance over time with the Clarion cochlear prosthesis. *Annals of Otology, Rhinology, & Laryngology, 166*, (Suppl.), 290-292.

Rosen, S., Faulkner, A., & Wildinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America*, 106, 3629-3636.

Staller, S., Menapace, C., Domico, E., Mills, D., Dowell, R.C., Geers, A., Pijl, S., Hasenstab, S., Justus, M., Bruelli, T., Borton, A.A., & Lemay, M. (1997). Speech perception abilities of adult and pediatric Nucleus implant recipients using spectral peak (SPEAK) coding strategy. *Otolaryngology-Head & Neck Surgery, 117*, 236-242.

Wilson, B.S., Finley, C.C., Lawson, D.T., Wolford, R.D., Eddington, D.K., & Rabinowitz, W.M. (1991). Better speech recognition with cochlear implants. *Nature, 352*(6332), 236-238.

Appendix A

```
PROCESSORTYPE
2

0 = Ts and Cs
1 = CIS
2 = n of m

NFORNOFM
2

STIMULATIONRATE
1250

NUMCHAN
6

DOAGCDRC
0

1 = do both agc and drc
0 = do not do agc or drc

AGCATTACK
0

AGCRELEASE
.250

DRCTABLE
  0  0
  0 -5
 -30 -5
 -50 -50
 -100 -100



ELECTRODETANDC
20 1 115 146 -80 -50
18 1 115 146 -80 -50
16 1 110 135 -80 -50
12 1 105 140 -80 -50
9 1 110 141 -80 -50
6 1 85 112 -80 -50
```

```
IIROUTPUTLPF
1.0000
-1.9496
0.9509
0.3094e-003
0.6187e-003
0.3094e-003

IIRFILTERTAPS
1.000000e+000
-3.824137e+000
5.498750e+000
-3.523986e+000
8.494152e-001
3.075495e-003
0.000000e+000
-6.150989e-003
0.000000e+000
3.075495e-003
       .
       .
       .

1.000000e+000
-3.794601e+000
5.474094e+000
-3.556699e+000
8.786601e-001
1.963396e-003
0.000000e+000
-3.926793e-003
0.000000e+000
1.963396e-003
```

Appendix B

```
%The Following Code generates the IIR bandpass filters
% using Butterworth filters and plots the resulting frequency
% response

clear;
SamplingRate = 22050;
NyquistRate = SamplingRate/2.0;

%Wn specifies the upper and lower frequency bounds in
%the order of each channel
Wn=[[150 555];
  [555 876];
  [876 1387];
  [1387 2190];
  [2190 3446];
  [3446 5500]];

  ylim('manual');
  xlim([.01 11000]);
  ylim([-90 1]);
  hold on

  Wn = Wn./NyquistRate;

for i = 1:6,
  A =1;
  [B, A] = butter(2,Wn(i,:));
  [H F] = freqz(B,A,1000,SamplingRate);
  H = 20*log10(abs(H));
  semilogx(F,H);
  taps = [taps;A(:)];
  taps = [taps;B(:)];
end

%Now to generate the coefficients for the preemphasis filter
[B,A] = butter(1,1200./NyquistRate,'high');
[H F] = freqz(B,A,1000,SamplingRate);
H = 20*log10(abs(H));
semilogx(F,H);

%Now to generate the coefficients for the low pass output filters
[B,A] = butter(2,125./NyquistRate);
[H F] = freqz(B,A,1000,SamplingRate);
H = 20*log10(abs(H));
semilogx(F,H);
hold off;

%Save the coefficients for cutting and pasting into
% the parameter file
fid = fopen('c:\IIRTAPS.txt','w');
fprintf(fid,'%e\n',taps);
fclose(fid);
```