RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 27
(2005)

David B. Pisoni, Ph.D.
Principal Investigator

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

Research Supported by:

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

National Institutes of Health
Research Grant No. DC-00111

and

National Institutes of Health
Training Grant No. DC-00012

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)

Table of Contents

Introduction v

Speech Research Laboratory Faculty, Staff, and Technical Personnel vi

I. Extended Manuscripts 1

• Some Observations on Representations and Representational Specificity in Speech Perception and Spoken Word Recognition
David B. Pisoni and Susannah V. Levi 3

• Modeling the Mental Lexicon as a Complex System: Some Preliminary Results Using Graph Theoretic Measures
Thomas M. Gruenenfelder and David B. Pisoni 27

• Speaker-independent Factors Affecting the Perception of Foreign Accent in a Second Language
Susannah V. Levi, Stephen J. Winters and David B. Pisoni 49

• Indexical and Linguistic Channels in Speech Perception: Some Effects of Voiceovers on Advertising Outcomes
Susannah V. Levi and David B. Pisoni 65

• Spoken Word Recognition Development in Children with Cochlear Implants: Effects of Residual Hearing and Hearing Aid Use in the Opposite Ear
Rachael F. Holt and Karen I. Kirk 81

• When and Why Feedback Matters in the Perceptual Learning of Visual Properties of Speech
Stephen J. Winters, Susannah V. Levi and David B. Pisoni 107

• Sound Similarity Relations in the Mental Lexicon: Modeling the Lexicon as a Complex Network
Vsevolod Kapatsinski 133

• Audiovisual Asynchrony Detection and Speech Perception in Normal Hearing Listeners and Hearing Impaired Listeners with Cochlear Implants
Marcia J. Hay-McCutcheon, David B. Pisoni and Kristopher K. Hunt 153

• Nonword Repetition with Spectrally Reduced Speech: Some Developmental and Clinical Findings
Rose A. Burkholder, Susannah V. Levi, Caitlin M. Dillon and David B. Pisoni 173

• Identification of Bilingual Talkers across Languages
Stephen J. Winters, Susannah V. Levi and David B. Pisoni 191

II. Short Reports and Work-in Progress 217

• Lip-reading Skills in Bilinguals: Some Effects of L1 on Visual-only Language Identification
Rebecca E. Ronquest and Luis Hernandez 219

• Cross-modal Priming of Auditory and Visual Lexical Information: A Pilot Study
Adam B. Buchwald and Stephen J. Winters 227

III. Publications: 2005 233

INTRODUCTION

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

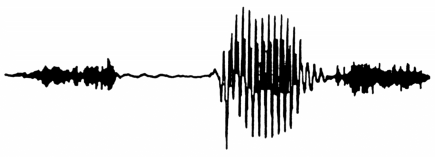
We are distributing progress reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of spoken language processing. We are, of course, very interested in following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

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

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

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

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



SPEECH – THE FINAL FRONTIER

SPEECH RESEARCH LABORATORY
FACULTY, STAFF, AND TECHNICAL PERSONNEL

(January 1, 2005–December 31, 2005)

RESEARCH PERSONNEL

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

Steven B. Chin, Ph.D. ..... Associate Scientist in Otolaryngology–Head and Neck Surgery[3]
Derek Houston, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3]
Tonya Bergeson-Dana, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3]
Marcia Hay-McCutcheon, Ph.D. ..... Assistant Professor of Otolaryngology–Head and Neck Surgery[3]

David L. Horn, M.D. ..... NIH Postdoctoral Trainee[3]
Rachael F. Holt, Ph.D. ..... NIH Postdoctoral Trainee[3]
Stephen J. Winters, Ph.D. ..... NIH Postdoctoral Trainee
Susannah V. Levi, Ph.D. ..... NIH Postdoctoral Trainee
Christopher M. Conway, Ph.D. ..... NIH Postdoctoral Trainee
Jeremy Loebach, Ph.D. ..... NIH Postdoctoral Trainee
Tessa Bent, Ph.D. ..... NIH Postdoctoral Trainee
Adam B. Buchwald, Ph.D. ..... NIH Postdoctoral Trainee
Thomas M. Gruenenfelder, Ph.D. ..... NIH Postdoctoral Trainee

Rose Burkholder, B.S. ..... NIH Predoctoral Trainee
Joshua Radicke, B.S. ..... NIH Predoctoral Trainee
Caitlin M. Dillon, B.A. ..... NIH Predoctoral Trainee
Vsevold Kapatsinski, B.A. ..... NIH Predoctoral Trainee
Rebecca Ronquest, B.A ..... NIH Predoctoral Trainee
Nick Altieri, B.A. ..... Graduate Research Assistant

Melissa Troyer ..... Undergraduate Research Assistant
Jennifer Karpicke ..... Undergraduate Research Assistant

TECHNICAL PERSONNEL

Luis R. Hernández, B.A. ..... Research Associate in Psychology
Darla J. Sallee ..... Administrative Assistant

---

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

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

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

E-MAIL ADDRESSES

Tessa Bent ..... tbent@indiana.edu
Tonya Bergeson-Dana ..... tbergeso@iupui.edu
Adam B. Buchwald ..... abuchwal@indiana.edu
Steven B. Chin ..... schin@iupui.edu
Christopher M. Conway ..... cmconway@indiana.edu
Tom Gruenenfelder ..... tgruenen@indiana.edu
Marcia Hay-McCutcheon ..... rmhaymccu@indiana.edu
Luis R. Hernández ..... hernande@indiana.edu
Rachael F. Holt ..... raholt@indiana.edu
David L. Horn ..... dlhorn@iupui.edu
Derek Houston ..... dmhousto@indiana.edu
Vsevold Kapatsinski ..... vkapatsi@indiana.edu
Jennifer Karpicke ..... jkarpick@indiana.edu
Susannah V. Levi ..... svlevi@indiana.edu
Jeremy L. Loebach ..... jlloebac@iupui.edu
David B. Pisoni ..... pisoni@indiana.edu
Josh Radicke ..... jradicke@indiana.edu
Rebecca Ronquest ..... rronques@indiana.edu
Darla J. Sallee ..... dsallee@indiana.edu
Melissa Troyer ..... mltroyer@indiana.edu
Stephen J. Winters ..... stwinter@indiana.edu

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Some Observations on Representations and Representational Specificity in Speech Perception and Spoken Word Recognition [1]

David B. Pisoni and Susannah V. Levi

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

[1] Preparation of this chapter was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We wish to thank Cynthia Clopper, Daniel Dinnsen, Robert Goldstone, Vsevolod Kapatsinski, Conor McLennan, Robert Port, and Steve Winters for their useful discussion and comments on this chapter.

Some Observations on Representations and Representational Specificity in Speech Perception and Spoken Word Recognition

Abstract. The conventional view of speech perception and spoken word recognition relies on discrete, abstract symbolic units. This conventional view faces several problems which led to the development of new approaches to representing speech. In particular, recent exemplar and episodic models can account for both the robustness of speech perception and the effects of indexical information on speech processing. Instead of only representing speech with conventional abstract symbolic representations, the evidence reviewed here suggests that highly detailed information is encoded and stored in memory as well.

Introduction

The field of speech perception and spoken word recognition has undergone rapid change over the last few years as researchers have begun to realize that many of the properties of speech that are responsible for its perceptual robustness, such as speed, fluency, automaticity, perceptual learning and adaptation, and errorful recovery, reflect general properties shared by other self-organizing systems in physics, biology, and neuroscience (Grossberg, 2003; McNellis & Blumstein, 2001; Sporns, 2003). Theoretical developments in cognitive science and brain modeling, as well as new computational tools, have led to a reconceptualization of the major theoretical problems in speech perception and spoken word recognition. Several new exemplar-based approaches to the study of speech perception and spoken word recognition have also emerged from independent developments in categorization (Kruschke, 1992; Nosofsky, 1986) and frequency-based phonology (Pierrehumbert, 2001; Bybee, 2001). These alternatives offer fresh ideas and new insights into old problems and issues related to variability and invariance (Goldinger, 1998; Goldinger & Azuma, 2003; Johnson, 1997). Although many of the basic research questions in speech perception remain the same, the answers to these questions have changed in fundamental ways reflecting new theoretical and methodological developments (Pardo & Remez, in press). These questions deal with the nature of phonological and lexical knowledge and representation, processing of stimulus variability, perceptual learning and adaptation and individual differences in linguistic performance (see Pisoni & Remez, 2005).

When compared to research in other areas of cognitive and neural science, speech perception is unique because of the close coupling and synchrony between speech production and perception. Speech exists simultaneously in several different domains: the acoustic and optical, the articulatory-motor and the perceptual. While the relations among these domains are complex, they are not arbitrary. The sound patterns used in a particular language function within a common linguistic system of contrast that is used in both production and perception. Thus, the phonetic contrasts generated in speech production by the vocal tract are precisely the same acoustic differences that serve a distinctive function in perceptual analysis by the listener (Stevens, 1972). As a result, any theoretical account of speech perception must also consider aspects of speech production and acoustics as well as optics. The articulatory spaces mapped out in speech production are closely coupled with the perceptual spaces used in speech perception and spoken word recognition (Fowler & Balantucci, 2005).

The fundamental problem in speech perception and spoken language processing is to describe how the listener recovers the talker's intended message. This complex problem has been typically broken down into several more specific subquestions: What stages of perceptual analysis intervene between the presentation of the speech signal and recognition of the intended message? What types of processing operations occur at each stage? What are the primary perceptual processing units and what is the nature

and content of the neural representations of speech in memory? Finally, what specific perceptual, cognitive, and linguistic mechanisms are used in speech perception and spoken language processing?

In this chapter, we provide an overview of some recent developments that have been underway in the field as they bear directly on issues surrounding representation and representational specificity in speech perception and spoken word recognition. Because of space limitations, our presentation is selective and is not meant to be an exhaustive survey of the field (see Pisoni & Remez, 2005). It is important to emphasize here, however, our strong belief that the changes that have occurred recently will very likely have profound and long-lasting effects on research, theory and clinical application in the years to come. Put in a slightly different way, there is a revolution going on in the field and it is important to understand the reasons for these changes in thinking and the consequences for the future.

Conventional View of Speech

Background

Different disciplines approach the study of speech perception and spoken language processing in fundamentally different ways reflecting their diverse interests, goals and theoretical assumptions. Linguists have one set of goals in mind while psycholinguists, cognitive scientists and neuroscientists have another set of goals. Historically, generative linguists adopted a formalist view and focused their research on two related problems: describing the linguistic knowledge that native speakers have about their language (their so-called linguistic competence) and explaining the systematic regularities and patterns that natural languages display. To accomplish these goals, linguists made several foundational assumptions about speech which embrace a strong abstractionist, symbol-processing approach. The linguistic approach to speech assumes that speech is structured in systematic ways and that the linguistically significant information in the speech signal can be represented efficiently and economically as a linear sequence of abstract, idealized, discrete symbols using an alphabet of conventional phonetic symbols. Linguists also assumed that the regularities and patterns observed within and between languages could be described adequately by sets of formal rules that operate on these abstract symbols. The segmental representations of speech that linguists constructed were assumed to be idealized and redundancy-free because they were designed to code only the linguistically significant differences in meaning between minimal pairs of words in the language (Twaddell 1952). These representations excluded other redundant or accidental information that may be present in the speech signal, but which is not linguistically contrastive. Two examples of this conventional view are given below.

> *". . . there is so much evidence that speech is basically a sequence of discrete elements that it seems reasonable to limit consideration to mechanisms that break the stream of speech down into elements and identify each element as a member, or as probably a member, of one or another of a finite number of sets." (Licklider, 1952, p. 590)*

> *"The basic problem of interest to the linguist might be formulated as follows: What are the rules that would make it possible to go from the continuous acoustic signal that impinges on the ear to the symbolization of the utterance in terms of discrete units, e.g., phonemes or the letters of our alphabet? There can be no doubt that speech is a sequence of discrete entities, since in writing we perform the kind of symbolization just mentioned, while in reading aloud we execute the inverse of this operation; that is, we go from a discrete symbolization to a continuous acoustic signal." (Halle, 1956, p. 510)*

The conventional segmental view of speech as a linear sequence of abstract, idealized, discrete symbols has been the primary method used for coding and representing the linguistic structure of spoken words in language. This approach to speech has been adopted across a wide range of related scientific disciplines that study speech processing such as speech and hearing sciences, psycholinguistics, cognitive and neural sciences and engineering (Peterson, 1952). The theoretical motivation for this approach goes back many years to the early Panini grammarians and it has become an inextricable part of all linguistic theories. Words have an internal structure and they differ from each other in systematic ways reflecting the phonological contrasts and morphology of a particular language. Although not often made explicit, several important basic theoretical assumptions are made in this particular view of speech that are worth mentioning because they bear directly on several broader theoretical issues related to the nature and content of lexical representations.

First, the conventional linguistic approach to the representation of speech assumes that a set of discrete and linear symbols can be used to represent what is essentially continuous, parametric and gradient information in the speech signal (Pierrehumbert & Pierrehumbert, 1990). Second, it is universally assumed by almost all linguists that the symbols representing phonetic segments or phonemes in speech are abstract, static, invariant, and context-free having combinatory properties like the individual letters used in alphabetic writing systems. Although speech can be considered as a good example of the “particulate principle of self-diversifying systems,” (Abler, 1989) a property of natural systems like genetics and chemical interaction that make "infinite use out of finite media," ambiguity and some degree of uncertainty still remain in the minds of some linguists and speech scientists about precisely what the elemental primitives of speech actually are even after many years of basic and applied research on speech. Are the basic building blocks of speech acoustic segments or features that emerge from speech perception or are they the underlying sensory-motor articulatory gestures used in speech production or are they both or something else?

Finally, the conventional view of speech relies heavily on some set of psychological processes that function to "normalize" acoustically different speech signals and to make them functionally equivalent in perception (Joos, 1948). It is generally assumed by both linguists and speech scientists that perceptual normalization is needed in speech perception in order to reduce acoustic-phonetic variability in the speech signal making physically different signals perceptually equivalent by bringing them into conformity with some common standard or referent (see Pisoni, 1997).

Problems with the Conventional View of Speech Perception

The fundamental problems in speech perception today are the same set of basic problems that have eluded definitive solution for more than four and a half decades (Fant, 1973; Stevens, 1998). Although the intractability of these long-standing problems has led to a voluminous body of literature on the production and perception of speech, researchers are still hard-pressed to describe and explain precisely how listeners perceive speech. Indeed, not only are speech scientists still unsure about the exact nature of the linguistic units arrived at in perceptual processing of speech, but little attention has been directed towards how perceptual analysis of the speech waveform makes contact with representations of words in the lexicon or how these representations are used to support spoken language understanding and comprehension. The acoustic consequences of coarticulation and other sources of contextually conditioned variability result in the failure of the acoustic signal to meet two formal conditions, linearity and invariance, which in turn give rise to a third related problem, the absence of segmentation into discrete units (first discussed by Chomsky & Miller, 1963).

Linearity of the Speech Signal. One fundamental problem facing the conventional view is linearity. The linearity condition states that for each phoneme in the message there must be a corresponding stretch of sound in the utterance (Chomsky & Miller, 1963). Furthermore, if phoneme X is followed by phoneme Y in the phonemic representation, the stretch of sound corresponding to phoneme X must precede the stretch of sound corresponding to phoneme Y in the physical signal. The linearity condition is clearly not met in the acoustic signal. Because of coarticulation and other contextual effects, acoustic features for adjacent phonemes are often "smeared" across individual phonemes in the speech waveform and a clear acoustic division between "adjacent" phonemes is rarely observed (Liberman, Delattres, & Cooper, 1952). Although segmentation is possible according to strictly acoustic criteria (see Fant, 1962), the number of acoustic "segments" is typically greater than the number of phonemes in the utterance. This smearing, or "parallel transmission" of acoustic features, results in stretches of the speech waveform in which acoustic features of more than one phoneme are present (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). For this reason, Liberman and his colleagues at Haskins Laboratories have argued that speech is not a simple cipher or alphabet, but is, instead, a complex code in which "speech sounds represent a very considerable restructuring of the phonemic 'message'" (p.4).

Acoustic-Phonetic Invariance. Another condition that the speech signal fails to satisfy is the principle of invariance (Chomsky & Miller, 1963). This condition states that for each phoneme X, there must be a specific set of critical acoustic attributes or features associated with X in all contexts. These "defining features" must be present whenever X or some variant of X occurs and they must be absent whenever some other phoneme occurs in the representation (Estes, 1994; Smith & Medin, 1981; Murphy, 2002). Because of coarticulatory effects in speech production, the acoustic features of a particular speech sound routinely vary as a function of the phonetic environment in which it is produced. For example, the formant transitions for syllable-initial stop consonants which provide cues to place of articulation (e.g., /b/ vs. /d/ vs. /g/) vary considerably depending on properties of the following vowel (Liberman, Delattre, Cooper & Gerstman, 1954). These transitions do not uniquely specify place of articulation across all vowels. If formant transitions are the primary cues to the perception of place of articulation for stop consonants, they are highly context-dependent. In short, the problem of acoustic-phonetic invariance is one of explaining how perceptual constancy for speech sounds is achieved and maintained when reliable acoustic correlates for individual phonemes in the speech waveform are absent (Blumstein & Stevens, 1981; Studdert-Kennedy, 1974).

Not only is invariance rarely observed for a specific segment across different phonetic environments within a talker, it is also absent for a particular segment in a particular context across speakers. For example, men, women, and children with different vocal tract lengths exhibit large differences in their absolute formant values (Peterson & Barney, 1952).

Speech Segmentation. The combination of non-linearity of the speech signal and context-conditioned variability leads to a third problem in speech perception, namely, the segmentation of the speech waveform into higher-order units of linguistic analysis such as syllables and words. Because of the lack of linearity and acoustic-phonetic invariance, the speech signal cannot be reliably segmented into acoustically defined units that are independent of adjacent segments and free from the conditioned effects of sentence-level contexts. For example, in fluent speech it is difficult to identify by means of simple acoustic criteria where one word ends and another begins.

Assumptions about segmentation and word recognition are probably not independent from assumptions made about the structure and organization of words in the lexicon (see Bradley & Forster, 1987; Luce, 1986). Precisely how the continuous speech signal is mapped on to discrete symbolic

representations by the listener has been and continues to be one of the most important and challenging problems to solve. In speech perception this is what is referred to as the "mapping problem."

The description of the mapping problem in speech was first characterized by Charles Hockett in his well-known Easter-egg analogy.

> *"Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point the belt carries the row of eggs between the two rollers of a wringer, which quite effectively smash them and rub them more or less into each other. The flow of eggs before the wringer represents the series of impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread out albumen, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer." (Hockett, 1955, p. 210)*

In the field of human speech perception and spoken word recognition, the basic mapping problem has simply been ignored by speech researchers who simply assumed that the continuous speech signal could be represented and encoded as a sequence of discrete symbols and that any further processing by the nervous system used these symbolic representations (Licklider, 1952; Peterson, 1952).

Indeed, a major stumbling block is that the conventional view has routinely assumed a bottom-up approach to speech perception and spoken word recognition where phonemes are first recognized and then parsed into words (Lindgren, 1965). An alternative view of speech perception that we discuss does not suffer from this problem because it allows for a top-down approach where words are recognized as whole units first, and then segmentation into phonemes follows as a natural consequence as required by the specific behavioral task and processing demands on the listener.

New Approaches to Speech Perception and Spoken Word Recognition

Views of the mental lexicon have changed significantly in recent years (Goldinger & Azuma 2003; Goldinger, 1998; Elman, 2004; Johnson, 1997). While traditional theories of word recognition and lexical access assumed that the mental lexicon consisted of a single canonical entry for each word (Oldfield, 1966; Marslen-Wilson, 1984; Morton, 1979), recent episodic approaches to the lexicon have adopted ideas from "multiple-trace" theories of human memory which propose that multiple entries for each word are encoded and stored in lexical memory in the form of detailed perceptual traces that preserve fine phonetic detail of the original articulatory event (Goldinger, 1996; Goldinger, 1998). In contrast to the conventional abstractionist, symbol-processing views of the lexicon, current episodic approaches to spoken word recognition and lexical access emphasize the continuity and tight coupling between speech perception, speech production, and memory in language processing (Goldinger, 1996, 1997, 1998).

Nonanalytic Cognition

Over the last twenty years, a large number of studies in cognitive psychology on categorization and memory have provided evidence for the encoding and retention of episodic or "instance-specific"

information (Jacoby & Brooks, 1984; Brooks, 1978; Tulving & Schacter, 1990; Schacter, 1990, 1992). According to this nonanalytic approach to cognition, stimulus variability is viewed as "lawful" and informative in perceptual analysis (Elman & McClelland, 1986). Memory involves encoding specific perceptual episodes, as well as the processing operations used during recognition (Kolers, 1973; Kolers, 1976). The major emphasis of this view of cognition is the focus on particulars and specific instances, rather than on abstract generalizations or symbolic coding of the stimulus input into idealized categories. Thus, the intractable problem of variability found in speech perception can be approached in fundamentally different ways by nonanalytic accounts of perception and memory.

We believe that the findings from studies on nonanalytic cognition are directly relevant to several long-standing theoretical questions about the nature of perception and memory for speech. When the criteria used for postulating episodic or nonanalytic representations are examined carefully (Brooks, 1978), it becomes obvious that speech displays a number of distinctive properties that make it amenable to this approach (Jacoby & Brooks, 1984). Several properties that encourage a nonanalytic processing strategy are high stimulus variability, complex stimulus-category relations, classifying inputs under incomplete information, and classifying inputs of structures with high analytic difficulty. These criteria are summarized briefly below.

High Stimulus Variability. Stimuli with a high degree of acoustic-phonetic variability are compatible with nonanalytic representations. Speech signals display a great deal of physical variability primarily because of factors associated with the production of spoken language. Among these factors are within- and between-talker variability, such as changes in speaking rate and dialect, differences in social contexts, syntactic, semantic and pragmatic effects and emotional state, as well as a wide variety of context effects due to the ambient environment such as background noise, reverberation and microphone characteristics (Klatt, 1986). These diverse sources of variability produce large changes in the acoustic-phonetic properties of speech and they need to be accommodated in theoretical accounts of the categorization process in speech perception. Variability is an inherent property of all biological systems including speech and it cannot be ignored, designed out of experimental protocols, or simply thought of as an undesirable source of noise in the system. Variability has to be taken seriously and approached directly.

Complex Stimulus-Category Relations. Complex relations between stimuli and their category membership can also be captured by nonanalytic processing strategies. In speech, the relation between the physical acoustic stimulus and its categorization as a string of symbols is complex because of the large amount of variability within a particular speaker across different phonetic contexts and the enormous variability across speakers. Despite these differences, categorization is reliable and robust (Twaddell, 1952). The conventional use of phonemes as perceptual units in speech perception entails a set of complex assumptions about category membership. These assumptions are based on linguistic criteria involving principles such as complementary distribution, free variation and phonetic similarity. In traditional linguistics, for example, the concept of a phoneme as a basic primitive of speech is used in a number of quite different ways. Gleason (1961), for example, characterizes the phoneme as a minimal unit of contrast, as the set of allophones of a phoneme, and as a non-acoustic abstract unit of a language. Thus, like other category domains studied by cognitive psychologists, speech sounds display complex stimulus-category relations that place strong constraints on the class of categorization models that can account for these operating principles.

Classifying Stimuli with Incomplete Information. Classifying incomplete or degraded stimuli is also consistent with nonanalytic analysis. Speech is a system that allows classification under highly degraded or incomplete information, such as silent-center vowels (Jenkins, Strange, & Trent, 1999),

speech processed through a cochlear implant simulator (Shannon et al., 1995), speech mixed with noise (Miller, Heise, & Lichten, 1951), or sinewave speech (Remez, Rubin, Pisoni, & Carrell, 1981). Correct classification of speech under these impoverished conditions is possible because speech is a highly redundant system which has evolved to maximize the transmission of linguistic information. In the case of speech perception, numerous studies have demonstrated the existence of multiple speech cues for almost every phonetic contrast (Raphael, 2005). While these speech cues are for the most part highly context-dependent, they also provide reliable information that can facilitate recognition of the intended message even when the signal is presented under poor listening conditions. This feature of speech perception permits very high rates of information transmission using sparsely-coded and broadly-specified categories (Pollack, 1952, 1953).

Classification of Stimuli with High Analytic Difficulty. Stimuli with high analytic difficulty are those which differ along one or more dimensions that are difficult to quantify or describe. Because of the complexity of speech and its high acoustic-phonetic variability, the category structure of speech is not amenable to simple hypothesis testing. As a result, it has been extremely difficult to construct a set of explicit formal rules that can successfully map multiple speech cues onto discrete phoneme categories. Moreover, the perceptual units of speech are also highly automatized; the underlying category structure of a language is learned in a tacit and incidental way by young children.

Episodic Approaches to Speech Perception

Not only has the focus in speech perception changed in recent years, but the conceptions of the mental lexicon and the nature of the representation of words in lexical memory have also undergone substantial revisions and development based on new findings and theoretical proposals from several different disciplines. The recent episodic approaches to the lexicon considered here assume that spoken words are represented in lexical memory as a collection of specific individual perceptual episodes or tokens rather the conventional abstract symbolic word types that have been universally assumed in the past. Evidence supporting episodic exemplar-based approaches to the mental lexicon has accumulated over the last few years as researchers from a number of related disciplines recognize the potential theoretical power and utility of this conceptual framework. Recent studies on the processing of stimulus variability provide evidence for episodic models of speech perception and spoken word recognition.

According to episodic views of perception and memory, listeners encode “particulars,” that is, specific instances or perceptual episodes, rather than generalities or abstractions (Kruschke, 1992; Nosofsky, 1986). Abstraction "emerges" from computational processes at the time of retrieval (Nosofsky, 1986; Estes, 1994). A series of studies carried out in our lab has shown that “indexical” information about a talker's voice and face and detailed information about speaking rate are encoded into memory and become part of the long-term representational knowledge that a listener has about the words of his/her language (Pisoni, 1997). Rather than discarding talker-specific details of speech in favor of highly abstract representations, these studies have shown that human listeners encode and retain very fine episodic details of the perceptual event (Pisoni, 1997).

In acquiring the sound system of a language, children not only learn to develop abilities to discriminate and identify sounds, they also learn to control the motor mechanisms used in speech articulation to generate precisely the same phonetic contrasts in speech production that they have become attuned to in perception. One reason that the developing perceptual system might preserve very fine episodic phonetic details of speech, as well as the specific characteristics of the talker's voice, would be to allow young children to accurately imitate and reproduce speech patterns heard in their surrounding language learning environment (Studdert-Kennedy, 1983). Imitation skills of this kind would provide

children with an enormous benefit in rapidly acquiring the phonology of the local dialect from speakers they are exposed to early in life.

In contrast to the conventional, abstractionist approach, episodic models assume that listeners store a very large number of specific instances and then use them in an analogical rather than analytic way to categorize novel stimuli (Brooks, 1978; Whittlesea, 1987). Recent findings showing that some sources of variability disrupt language processing and that familiarity with the details of the voice benefit language processing provide converging support for the claim that very detailed, instance-specific information about speech is encoded, represented and stored in memory.

Evidence for Detailed Episodic Representations

Over the last 15 years, we have been carrying out a research program on different sources of variability in speech, specifically, variability from different talkers, speaking rates and speaking modes, to determine how these factors affect spoken word recognition. Our findings suggest that many long-standing theoretical assumptions held by researchers about basic perceptual units of speech such as features, phonemes, and syllables need to be substantially revised. In particular, assuming the existence of only abstract symbolic representations of speech cannot account for the new results showing that variability matters in speech perception and that detailed episodic information affects language processing and memory.

Encoding and Storage of Variability in Speech Perception A number of studies from our research group have explored the effects of different sources of variability on speech perception and spoken word recognition. Instead of reducing or eliminating variability in the stimulus materials, as most speech researchers have routinely done over the years, in a series of novel studies we specifically introduced variability from different talkers and different speaking rates to directly study the effects of these variables on perception (Pisoni, 1993).

Our research on this problem first began with several observations of Mullennix, Pisoni and Martin (1989) who found that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers that were used to generate the test words in the stimulus ensemble. In one condition, all the words in a test list were produced by a single talker; in another condition, the words were produced by 15 different talkers, which included both male and female voices. Across three different signal-to-noise ratios, identification performance was always better for words produced by a single talker than words produced by multiple talkers. Trial-to-trial variability in the speaker's voice affected recognition performance. These findings replicated results originally reported by Peters (1955) and Creelman (1957) and suggested to us that the perceptual system is highly sensitive to talker variability and therefore must engage in some form of "recalibration" each time a novel voice is encountered during the set of test trials using multiple voices.

In a second set of experiments, Mullennix et al. (1989) measured naming latencies to the same set of words presented under single and multiple-talker two test conditions. They found that subjects were not only slower to name words presented in multiple-talker lists but they were also less accurate when their performance was compared to words from single-talker lists. Both sets of findings were surprising in light of the conventional view of speech perception because all the test words used in the experiment were highly intelligible when presented in the quiet. The intelligibility and naming data from this study immediately raised a number of additional questions about how the different perceptual dimensions of the speech signal are processed and encoded by the human listener. At the time, we followed the conventional view of speech that assumed that the acoustic attributes of the talker's voice

were processed independently of the linguistic properties of the signal, although no one had ever tested this assumption directly.

In another series of experiments, Mullennix and Pisoni (1990) used a speeded classification task to assess whether attributes of a talker's voice are perceived independently of the phonetic form of the words. Subjects were required to attend selectively to one stimulus dimension (e.g., talker voice) while simultaneously ignoring another stimulus dimension (e.g., phoneme). Across all conditions, Mullennix and Pisoni found increases in interference from both perceptual dimensions when the subjects were required to attend selectively to only one of the stimulus dimensions. The pattern of results suggested that words and voices were processed as integral dimensions; that is, the perception of one dimension (e.g., phoneme) affects classification of the other dimension (e.g., voice) and vice versa. Subjects could not selectively ignore irrelevant variation in the non-attended dimension. If both perceptual dimensions were processed separately, as we originally assumed, interference from the non-attended dimension should not have been observed. Not only did we find mutual interference between the two dimensions suggesting that the perceptual dimensions were perceived in a mutually-dependent manner, but we also found that the pattern of interference was asymmetrical. It was easier for subjects to ignore irrelevant variation in the phoneme dimension when their task was to classify the voice dimension than it was for them to ignore the voice dimension when they had to classify the phonemes.

The results from these novel perceptual experiments were surprising to us at the time given our original assumption that the indexical and linguistic properties of speech are perceived independently. To study this problem further, we carried out a series of memory experiments to assess the mental representation of speech in long-term memory. Experiments on serial recall of lists of spoken words by Martin, Mullennix, Pisoni, and Summers (1989) and Goldinger, Pisoni, and Logan (1991) demonstrated that specific details of a talker's voice are not lost or discarded during early perceptual analysis but are perceived and encoded in long-term memory along with item information. Using a continuous recognition memory procedure, Palmeri, Goldinger, and Pisoni (1993) found that detailed episodic information about a talker's voice is also encoded in memory and is available for explicit judgments even when a great deal of competition from other voices is present in the test sequence.

In a series of other recognition memory experiments, Goldinger (1998) found strong evidence of implicit memory for attributes of a talker's voice which persists for a relatively long period of time (up to a week) after perceptual analysis has been completed. Moreover, he also showed that the degree of perceptual similarity between voices affects the magnitude of repetition priming effects, suggesting that the fine phonetic details are not lost and the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations (see Goldinger, 1997).

Other experiments were carried out to examine the effects of speaking rate on perception and memory. These studies, which were designed to parallel the earlier experiments on talker variability, also found that the perceptual details associated with differences in speaking rate were not lost as a result of perceptual analysis. In one experiment, Sommers, Nygaard, and Pisoni (1992) found that words were identified more poorly when speaking rate was varied (i.e., fast, medium and slow), than when the same words produced at a single speaking rate. These results were compared to another condition in which differences in amplitude were varied randomly from trial to trial in the test sequences. In this case, identification performance was not affected by variability in overall signal level.

The effects of speaking rate variability have also been observed in experiments on serial recall. Nygaard, Sommers, and Pisoni (1992) found that subjects recalled words from lists produced at a single speaking rate better than the same words produced at several different speaking rates. Interestingly, the

differences appeared in the primacy portion of the serial position curve suggesting greater difficulty in the transfer of items into long-term memory. The effects of differences in speaking rate, like those observed for talker variability in our earlier experiments, suggested that perceptual encoding and rehearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by low-level perceptual sources of variability. If these sources of variability were automatically filtered out or normalized by the perceptual system at early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments. Taken together, the findings on variability and speaking rate suggest that details of the early perceptual analysis of spoken words are not lost as a result of early perceptual analysis. Detailed perceptual information becomes an integral part of the mental representation of spoken words in memory. In fact, in some cases, increased stimulus variability in an experiment may actually help listeners to encode items into long-term memory because variability helps to keep individual items in memory more distinct and discriminable, thereby reducing confusability and increasing the probability of correct recall (Goldinger, Pisoni, & Logan, 1991; Nygaard, Sommers, & Pisoni, 1992). Listeners encode speech signals along many perceptual dimensions and the memory system apparently preserves these perceptual details much more reliably than researchers believed in the past.

Reinstatement in Speech Perception and Spoken Word Recognition. Our findings on the perception of talker variability and speaking rate encouraged us to examine perceptual learning in speech more carefully, specifically, the rapid tuning or perceptual adaptation that occurs when a listener becomes familiar with the voice of a specific talker (Nygaard, Sommers & Pisoni, 1994). This particular problem has not received very much attention in the field of human speech perception despite its obvious relevance to problems of speaker normalization, acoustic-phonetic invariance and the potential application to automatic speech recognition and speaker identification (Fowler, 1990; Kakehi, 1992; Bricker & Pruzansky, 1976). An extensive search of the research literature on talker adaptation by human listeners revealed only a small number of behavioral studies on this topic and all of them appeared in obscure technical reports from the late 1940s and early 1950s (Mason, 1946; Miller, Wiener, & Stevens, 1946; Peters, 1955).

To determine how familiarity with a talker's voice affects the perception of spoken words, Nygaard, Sommers, and Pisoni (1994) had two groups of listeners learn to explicitly identify a set of ten unfamiliar voices over a nine-day period using common names (i.e., Bill, Joe, Sue, Mary). After this initial learning period, subjects participated in a word recognition experiment designed to measure speech intelligibility. Subjects were presented with a set of novel words mixed in noise at several signal-to-noise ratios. One group of listeners heard the words produced by talkers that they were previously trained on, and the other group heard the same words produced by a new set of unfamiliar talkers. In the word recognition task, subjects were required to identify the words rather than recognize the voices, as they had done in the first phase of the experiment.

The results of the speech intelligibility experiment showed that the subjects who had heard novel words produced by familiar voices were able to recognize the novel words more accurately than subjects who received the same novel words produced by unfamiliar voices. Differences in inherent intelligibility between the two sets of words was not a confounding factor. An additional study with two new sets of untrained listeners confirmed that both sets of voices were equally intelligible, indicating that the difference in performance found in the original study was due to training.

The findings from this voice learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by a familiar talker. Thus, speech

perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that was obtained in an entirely different experimental task.

What kind of perceptual knowledge does a listener acquire when he listens to a speaker's voice and is required to carry out an explicit name recognition task like our subjects did in this experiment? One possibility is that the procedures or perceptual operations (Kolers, 1973) used to recognize the voices are encoded and retained in some type of "procedural memory" and these perceptual analysis routines are invoked again when the same voice is encountered in a subsequent intelligibility test. This kind of procedural knowledge might increase the efficiency of the perceptual analysis for novel words produced by familiar talkers because detailed analysis of the speaker's voice would not have to be carried out over and over again as each new word was encountered. Another possibility is that specific instances-- perceptual episodes or exemplars of each talker's voice are encoded and stored in memory and then later retrieved during the process of word recognition when new tokens from a familiar talker are encountered (Jacoby & Brooks, 1984).

Whatever the exact nature of this procedural knowledge turns out to be, the important point to emphasize here is that prior exposure to a talker's voice facilitates subsequent recognition of novel words produced by the same talkers. Such findings demonstrate a form of source memory for a talker's voice that is distinct from the retention of the individual items used and the specific task that was originally employed to familiarize the listeners with the voices (Glanzer, Hilford, & Kim, 2004; Roediger, 1990; Schacter, 1992). These findings provide additional support for the view that the internal representation of spoken words encompasses both a phonetic description of the utterance, as well as information about the structural description of the source characteristics of the specific talker. The results of these studies suggest that speech perception is carried out in a "talker-contingent" manner; the indexical and linguistic properties of the speech signal are closely coupled and are not dissociated into separate, independent channels in perceptual analysis.

Anti-Representationalist Approaches

Another more radical approach to representations has been proposed recently by a group of artificial intelligence (AI) researchers working on behavior-based autonomous robotics and biological intelligence (Brooks, 1991a,b; Beer, 2000; Clark,1999). According to this perspective, called “embodied cognition,” mind, body and world are linked together as a “coupled” dynamic system (Beer, 2000; Clark, 1999). Conventional abstract symbolic representations and information processing involving the manipulation of abstract symbols are not needed to link perception and action directly in real-world tasks, such as navigating around in novel unpredictable environments. Modest degrees of intelligent behavior have been achieved in robots without computation and without complex knowledge structures representing models of the world (Brooks, 1991a,b). Intelligent adaptive behavior reflects the operation of the whole system working together in synchrony without control by a dedicated central executive that is needed to access and manipulate abstract symbolic representations and guide behavior based on internal models of the world.

These are strong claims and important criticisms to raise about the most central and basic foundational assumptions of classical cognition and traditional information processing approaches to perception, memory, learning and language. While the bulk of the research efforts on embodied and situated cognition has come from the field of AI and is related to constructing autonomous robots and establishing links between perception and action in simple sensory-motor systems, the recent arguments against conventional abstract symbolic representations and the mainstream symbol-processing views of cognition and intelligence have raised a number of challenging theoretical issues that are directly relevant

to current theoretical assumptions about representations and processes in speech perception and spoken word recognition. With regard to the problems of representations in speech perception and spoken word recognition, these issues are concerned directly with questions about "representational specificity" and the nature of lexical representations assumed in spoken word recognition and comprehension. Such an anti-representation view of spoken language has been proposed recently by Port who argues that discrete representations are not needed for real-time human speech perception (Port & Leary, 2005).

Although the anti-representation theorists working in AI have argued that it is not necessary to postulate conventional symbolic representations or even to assume complex mediating states corresponding to internal models of the world for the relatively simple sensory and motor domains they have explored so far, there are several reasons to believe that their global criticisms of the conventional symbol-processing approach to cognition may not generalize gracefully to more complex knowledge-based cognitive domains (Markman & Dietrich, 2000). Compared to the simple sensory-motor systems and navigational behaviors studied by researchers working on autonomous robotics, there is good consensus that speech perception and spoken language processing are "informationally-rich" and "representationally-hungry" knowledge-based domains (Clark, 1997) that shares computational properties with a small number of other complex self-diversifying systems. These are systems like language, genetics, and chemistry that have a number of highly distinctive powerful combinatorial properties that set them apart and make them uniquely different from other natural complex systems that have been studied in the past.

Several years ago, William Abler (1989) examined the properties of self-diversifying systems and drew several important parallels with speech and spoken language. He argued that human language displays structural properties that are consistent with other "particulate systems" such as genetics and chemical interaction. All of these systems have a small number of basic "particles" such as genes or atoms that can be combined and recombined to create infinite variety and unbounded diversity without blending of the individual components or loss of perceptual distinctiveness of the new patterns created by the system.

It is hard to imagine that any of the anti-representationalists would seriously argue or even try to maintain that speech and spoken language are non-representational or non-symbolic in nature. The mere existence of reading, orthographies and alphabetic writing systems can be taken as strong evidence and serve as an existence proof that some aspects of speech and spoken language can be represented discretely and efficiently by a linear sequence of abstract symbols. Looking at several selected aspects of speech and the way spoken languages work, it is obvious that spoken language can be offered as the prototypical example for a symbol-processing system. Indeed, this is one of the major "design features" of human language (Hockett, 1960).

Evidence for Symbolic Representations in Speech Perception

For a number of years, there has been an on-going debate concerning the role of segmental representations in speech perception and spoken word recognition. Several theorists have totally abandoned an intermediate segmental level of representation in favor of direct access models of spoken word recognition (Gaskell & Marslen-Wilson, 1997; Klatt, 1979). In these models, words are recognized without an analysis of their "internal structure" into units like phones, allophones, phonemes, diphones, or demisyllables. In this section, we present arguments against this position and summarize evidence from several different areas supporting the existence of discrete segmental units in speech perception and spoken word recognition.

The first general line of evidence we offer in support of segmental representations in speech perception comes from linguistics. One of the fundamental assumptions of linguistic analysis is that the continuously varying speech waveform can be represented as a sequence of discrete units such as features, phones, allophones, phonemes, and morphemes. This assumption is central to all current conceptions of language as a system of rules that governs the sound patterns and sequences used to encode meanings (Chomsky & Halle, 1968). The widespread existence of a range of phonological phenomena such as alternation, systematic regularity, and diachronic and synchronic sound changes require, ipso facto, that some type of segmental level be postulated in order to capture significant linguistic generalizations that exist within and between languages. In describing the sound structure of a given language, then, a level of segmental representation is required in order to account for the idiosyncratic and predictable regularities of the sound patterns of that language (see Kenstowicz & Kisseberth, 1979). Whether these segmental units are actually used by human listeners in the real-time analysis of spoken language is another matter.

The second general line of evidence in support of the segmental representations in speech perception is psychological in nature. One source of evidence comes from observations of speakers of languages with no orthography who are attempting to develop writing systems. In his well-known article on the psychological reality of phonemes, Edward Sapir (1933) cites several examples of conscious awareness of the phonological structure language. Read (1971) also described a number of examples of children who have invented their own orthographies spontaneously. The children's initial encounters with print show a systematic awareness of the segmental structure of language, thereby demonstrating an ability to analyze spoken language into representations with discrete segments. Several theorists have also proposed that young children's ability to learn to read an alphabetic writing system like English orthography is highly dependent on the development of phonemic analysis skills, that is, perceptual and linguistic skills that permit the child to consciously analyze speech into segmental units (Liberman, Shankweiler, Fischer, & Carter, 1974; Rozin & Gleitman, 1977; Treiman, 1980).

The existence of language games based on insertion of a sound sequence, movement of a sound sequence, or deletion of a sound sequence all provide additional support for the existence of segmental units in the internal structure of words (see Treiman, 1983, 1985). The presence of rhymes and the metrical structure of poetry also entail an awareness that words have an internal structure and organization and that this structure can be represented as linear sequence of discrete symbolic units distributed in time.

An examination of errors in speech production also provides additional evidence that words are represented in the lexicon in terms of segments. The high frequency of single segment speech errors such as substitutions and exchanges provide evidence of the phonological structure of the language (Fromkin, 1973, 1980; Garrett, 1976, 1980; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1982). It is difficult, if not impossible, to explain these kinds of error patterns without assuming some kind of segmental representation in the organization of the lexicon used for speech production.

Other evidence comes from studies of the phoneme-restoration effect (Samuel, 1981 a,b; Warren, 1970), a phenomenon demonstrating the on-line synthesis of the segmental properties of fluent speech. Many studies have also been carried out using the phoneme monitoring task in which subjects are required to detect the presence of a specified target phoneme while listening to sentences or short utterances (see Foss, Hawood, & Blank, 1980). Although some earlier findings suggested that listeners first recognize the word and then carry out an analysis of the segments within the word (Foss and Swinney, 1973; Morton and Long, 1976), other studies indicate that subjects can detect phonemes in nonwords that are not present in their lexicon (Foss & Blank, 1980; Foss & Gernsbacher, 1983). Thus,

subjects can detect phonemes based on two sources of knowledge: information from the sensory input and information constructed from their knowledge of the phonological structure of the language (Dell & Newman, 1980).

Finally, in terms of perceptual data, there is a growing body of data on misperceptions of fluent speech (Bond & Garnes, 1980; Bond & Robey, 1983; Bond, 2005). The errors collected in these studies also suggest that a very large portion of the misperceptions involve segments rather than whole words or features.

Recognizing the existence of segments in the representation of phonology does not imply that this is the only information included in the representation. Indeed, this is precisely what we argued against earlier. Indexical properties of speech also contribute to the representation and processing of speech and language, especially under highly degraded listening conditions when multiple sources of knowledge are routinely used to perceive and interpret highly impoverished, partially-specified speech signals.

Representational Specificity

In our view, the current debate that emerges from the embodied cognition criticisms of conventional, symbolic representations is not about whether spoken language is a symbol processing system or whether there are representations and internal states. In the case of language, the evidence is pretty clear; low-dimensional segmental units exist at multiple levels of language. The principal theoretical issues revolve around describing more precisely the exact nature of the phonetic, phonological and lexical representations used in speech perception, production and spoken language processing and the degree of representational specificity these representations preserve.

Two major questions emerge: (a) how much detail of the original speech signal is encoded by the brain and nervous system in order to support language processing and (b) how much detail can be discarded as a consequence of phonological and lexical analysis? Some sources of information in speech are clearly more important and linguistically significant than others and understanding these particular properties of the speech signal may provide new insights into both representation and process and may help to resolve many of the long-standing issues in the field. Moreover, the results from numerous perceptual studies with human listeners over the last 50 years indicate that the distinctive properties of speech vary with the specific task demands placed on the listener as well as properties of the talker. Thus, there may not be one basic unit of perception or only one common representational format in speech perception and spoken word recognition. It is very likely there are multiple units and several different representations that are used in parallel (see Pisoni & Luce, 1987).

Interface Between Speech Perception and Spoken Word Recognition

The conventional symbol-processing approach to speech has a long history dating back to the early days of telephone communications (Allen, 1994, 2005; Fletcher, 1953). The principal assumption of this bottom-up approach to spoken language processing is that spoken word recognition is logically based on prior phoneme identification and that spoken words are recognized by recovering and identifying sequences of phonemes from the acoustic-phonetic information present in the speech waveform. In the early days of speech research, the basic building blocks of speech—the perceptual primitives, were universally assumed to be the discrete segments and symbols—phones or phonemes that were derived from linguistic analysis of speech (Fano, 1950; Licklider, 1952; Peterson, 1952).

According to this conventional approach, speech perception is equivalent to phoneme perception. As the thinking went at the time, if a listener could recognize and recover the phonemes from the speech waveform like reading discrete letters on the printed page, he/she would be successful in perceiving the component words and understanding the talker's intended message (Allen, 2005). This bottom-up reductionist approach to speech perception was readily embraced and universally adopted by engineers, psychologists, and linguists and this view of speech perception is still widely accepted in the field of speech science even today despite the technical and conceptual difficulties that have been encountered over the last 50 years in trying to identify reliable discrete physical units in the speech waveform that correspond uniquely to the component sound segments of the linguistic message resulting from perceptual analysis. The primary problem of this bottom-up approach is its inability to deal with the enormous amount of acoustic-phonetic variability that exists in the speech waveform.

The conventional bottom up "segmental view" of speech perception and spoken language processing was significantly transformed and recast in a fundamentally different way in the early 1980's by Marslen-Wilson (Marslen-Wilson & Welsh, 1978). He argued convincingly that the primary objective of the human language comprehension system is the recognition of spoken words rather than the identification of individual phonemes in the speech waveform (see also Blesser, 1972). Marslen-Wilson proposed that the level at which lexical processing and word recognition is carried out in language comprehension should be viewed as the functional locus of the interactions between the initial bottom-up sensory input in the speech signal and the listener's contextual-linguistic knowledge of the structure of language. Thus, spoken word recognition was elevated to a special and privileged status within the conceptual framework of the Cohort Theory of spoken language processing developed by Marslen-Wilson and his colleagues (Marslen-Wilson, 1984). Speech perception is thus no longer simply phoneme perception, but it is also the process of recognizing spoken words and understanding sentences.

Cohort theory has been extremely influential in bringing together research scientists working in what at the time were two quite independent fields of research on spoken language processing—speech and hearing scientists who were studying speech cues and speech sound perception and psycholinguists who were investigating spoken word recognition, lexical access and language comprehension. The theoretical assumptions and strong claims of cohort theory served to focus and solidify research efforts on common problems that were specific to speech perception and spoken language processing as well as a set of new issues surrounding the organization of words in the mental lexicon (Grosjean & Frauenfelder, 1997). Segments and phonemes "emerge" from the process of lexical recognition and selection rather than the other way around. Lexical segmentation, then, may actually be viewed as a natural by-product of the primary lexical recognition process itself (Reddy, 1975).

Closely related to Cohort Theory is the Neighborhood Activation Model (NAM) developed by Luce and Pisoni (1998). NAM confronts the acoustic-phonetic invariance problem more directly by assuming that a listener recognizes a word "relationally" in terms of oppositions and contrasts with phonologically similar words. Like the Cohort Model, the focus on spoken word recognition in NAM avoids the long-standing problem of recognizing individual phonemes and features of words directly by locating and identifying invariant acoustic-phonetic properties. A key methodological tool of the NAM has been the use of a simple similarity metric for estimating phonological distances of words using a one-phoneme substitution rule (Greenberg & Jenkins, 1964; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). This computational method provided an efficient way of quantifying the "perceptual similarity" between words in terms of phonological contrasts among minimal pairs.

As Luce and McLennan (2005) have recently noted in their discussion of the challenges of variation in speech perception and language processing, all contemporary models of the spoken word

recognition assume that speech signals are represented in memory using conventional abstract representational formats consisting of discrete features, phones, allophones or phonemes. Current models of spoken word recognition also routinely assume that individual words are represented discretely and are organized in the mental lexicon. All of the current models also assume that the mental lexicon contains abstract idealized word "types" that have been normalized and made equivalent to some standard representation. None of the current models encode or store specific instances of individual word "tokens" or detailed perceptual episodes of speech (but see Kapatsinski, forthcoming for an alternative). Not only are the segments and features of individual words abstract, but the lexical representations of words and possible nonwords, are assumed to consist of abstract types, not specific experienced tokens.
The only exception to this general pattern of thinking about speech as a sequence of abstract symbols was the LAFS model proposed by Klatt (1979). The LAFS model assumed that words were represented in the mental lexicon as sequences of power spectra in a large multidimensional acoustic space without postulating intermediate phonetic representations or abstract symbols (also see Treisman 1978a, 1978b). The recognition process in LAFS was carried out directly by mapping the power spectra of sound patterns onto words without traditional linguistic features or an intermediate level of analysis corresponding to discrete segments or features. In many ways, LAFS was ahead of its time in terms of its radical assumptions that intermediate segmental representations are not needed in spoken word recognition and that an optimal system does not discard potentially useful information.

Frequency and Usage-Based Views from Linguistics

Concerns about the inadequacies of the conventional, abstractionist representations of speech have also been expressed recently by a small group of linguists who have been promoting frequency- and usage-based accounts of a range for phenomena in phonetics, phonology, and morphology. For example, Pierrehumbert (1999) argued that conventional accounts of language are unable to capture several generalizations about phonological regularity and change in language. Instead, she argues that a probabilistic or stochastic approach deals better with language-particular phonetic targets (e.g., location of cardinal vowels in the vowel space or VOT differences), phonotactics (e.g., new generalizations about word-internal consonant clusters in relation to the probability of the individual parts occurring in word-initial or final position), and morphological alternations (e.g. vowel changes like in serene/serenity).

Bybee has also recently suggested that fine phonetic details of specific instances of speech are retained in phonological representations (Bybee, 2005). In Bybee's model, individual tokens/exemplars are stored in memory and the frequency of these tokens accounts for resistance to morphological leveling (e.g., keep/kept~*keeped versus weep/wept~weeped), phonetic reduction (e.g., the frequent "I don't know"), and grammaticalization (e.g., gonna < "going to" from the general motion verb construction "journeying to", "returning to", "going to", etc.) (Bybee 1998, 1999, 2005). Even Donca Steriade, who has carried out extensive research in phonology within the formalist tradition has suggested recently that acoustic-phonetic variability in speech needs to be captured and represented in some fashion in linguistic representations and analysis that reflect actual experience with specific instances an individual tokens of speech (Steriade, 2001 a,b).

Conclusions

Evidence from a wide variety of studies suggests that speech is not initially perceived and transformed into idealized abstract context-independent symbolic representations like sequences of letters on the printed page. Instead, highly detailed perceptual traces representing both the "medium" (detailed source information) and the "message" (content of the utterance) are encoded and stored in memory for later retrieval in the service of word recognition, lexical access and spoken language

comprehension. A record of the processing operations and procedures used in perceptual analysis and recognition remains after the primary recognition process has been completed and this residual information is used again when the same source information is encountered in another utterance. Speech is not simply transformed or recoded into an abstract idealized symbolic code like the linear sequence of discrete segments and features resulting from a linguist's phonetic transcription. The fine phonetic details of the individual talker's articulation in production of speech are not lost or discarded as a result of early perceptual processing; instead, human listeners retain dynamic information about the sensory-motor procedures and the perceptual operations and these sources of information become an integral part of the neural and cognitive representation of speech in long-term lexical memory. The representation of speech is not an either/or phenomenon where abstraction and detailed instance-specific exemplars are mutually exclusive; evidence for both detailed episodic traces and abstract segments exist and both must be represented in memory.

The most important and distinctive property of speech perception is its perceptual robustness in the face of diverse physical stimulation over a wide range of environmental conditions that produce large changes and transformations in the acoustic signal. Listeners adapt very quickly and effortlessly to changes in speaker, dialect, speaking rate and speaking style and are able to adjust rapidly to acoustic degradations and transformations such as noise, filtering, and reverberation that introduce significant physical perturbations to the speech signal without apparent loss of performance. Investigating these remarkable perceptual, cognitive and linguistic abilities and understanding how the human listener recognizes spoken words so quickly and efficiently despite enormous variability in the physical signal and listening conditions is the major challenge for future research in speech perception and spoken word recognition.

References

Abler, W.L. (1989). On the particulate principle of self-diversifying systems. *Journal of Social Biological Structure*, *12***,** 1-13.

Allen, J.B. (1994). How do humans process and recognizer speech? *IEEE Trans. Speech audio*, *2*, 567-577.

Allen, J.B. (2005). *Articulation and intelligibility*. Morgan and Claypool Publishers, San Rafael.

Beer, R.D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*, 91-99.

Blesser, B. (1972). Speech perception under conditions of spectral transformations: I. Phonetic characteristics. *Journal of Speech and Hearing Research*, *15*, 5-41.

Blumstein, S.E. & Stevens, K.N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, *10*, 25-32.

Bond, Z.S. (2005). Slips of the ear. In D.B. Pisoni & R.E. Remez, (eds.), *The handbook of speech perception*, pp. 290-310. Blackwell Publishing Ltd, Oxford, UK.

Bond, Z.S. & Garnes, S. (1980). Misperceptions of fluent speech. In R.A. Cole (ed.), *Perception and production of fluent speech*. pp. 115-132. Erlbaum, Hillsdale, NJ.

Bond, Z.S. & Robey, R.R. (1983). The phonetic structure of errors in the perception of fluent speech. In N.J.U. Lass, ed. *Speech and language: Advances in basic research and practice* (Vol. 9). pp. 249-283. Academic Press, New York.

Bradley, D.C. & Forster, K.I. (1987). A reader's view of listening. *Cognition*, *25*, 103-134.

Bricker, P.D. & Pruzansky, S. (1976). Speaker Recognition. In N.J. Lass (ed.), *Contemporary Issues in Experimental Phonetics*, pp. 295-326. Academic Press, New York.

Brooks, L. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (eds.), *Cognition and categorization*, 169-211. Erlbaum, Hillsdale, NJ.

Brooks, R.A. (1991a). New approaches to robotics. *Science*, *253*(5025), 1227-1232.

Brooks, R.A. (1991b). Intelligence without representation. *Artificial Intelligence*, *47*, 139-159.

Bybee, J.L. (1998). The emergent lexicon. *Chicago Linguistic Society, 34*, 421-435.

Bybee, J.L. (1999). Usage-based phonology. In M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan, & K. Wheatley (eds.), *Functionalism and Formalism in Linguistics, Volume I: General papers; Volume II: Case studies*, pp. 211-242. John Benjamins, Amsterdam, Netherlands.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press, Cambridge.

Bybee, J.L. (2005). *The impact of use on representation: grammar is usage and usage is grammar*. Presidential address, Annual Meeting of the Linguistic Society of America, Oakland.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. Harper and Row, New York.

Chomsky, N. & Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In R.D. Luce, R. Bush, & E. Galanter (ed.), *Handbook of mathematical psychology (Vol 2)*, pp. 269-321. John Wiley & Sons, New York.

Clark, A. (1997) *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*, 345-351.

Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, *29*, 655.

Dell, G.S. & Newman, J.E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, *19*, 608-623.

Elman, J.L. (2004). An alternative view of the mental lexicon. *TRENDS in Cognitive Sciences*, *8*, 301-306.

Elman, J.L. & McClelland, J.L. (1986). Exploiting lawful variability in the speech waveform. In J.S. Perkell & D.H. Klatt (eds.), *Invariance and Variability in Speech Processing*. pp. 360-385. Erlbaum, Hillsdale, NJ.

Estes, W.K. (1994) *Classification and Cognition*. Oxford psychology series, No. 22. Oxford University Press, New York, NY.

Fano, R.M. (1950). The information theory point of view in speech communication. *Journal of the Acoustical Society of America*, *22*, 691-696.

Fant, G. (1973). *Speech sounds and features*. The MIT Press, Cambridge, MA.

Fant, G. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos*, *5*, 3-17.

Fletcher, H. (1953). *Speech and Hearing in Communication*. Robert E. Krieger Publishers: Huntington, NY.

Foss, D.J. & Blank, M.A. (1980). Identifying the speech codes. *Cognitive Psychology*, 12, 1-31.

Foss, D.J. & Gernsbacher, M.A. (1983). Cracking the dual code: Towards a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, *22*, 609-632.

Foss, D.J., Harwood, D.A., & Blank, M.A. (1980). Deciphering decoding decisions: Data and devices. In RA Cole (ed.), *Perception and production of fluent speech*. pp. 165-199. Erlbaum, Hillsdale NJ.

Foss, D.J. & Swinney, D.A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, *12*, 246-257.

Fowler, C.A. & Balantucci, B. (2005) The relation of speech perception and speech production. In D.B. Pisoni & R.E. Remez (eds.), *The handbook of speech perception*, pp. 633-652. Blackwell Publishing Ltd: Oxford, UK.

Fowler, C.A. (1990). Listener-talker attunements in speech. *Haskins Laboratories Status Report on Speech Research*, *101-102*, 110-129.

Fromkin, V. (1980). *Errors in linguistic performance*. Academic Press: NY.

Fromkin, V. (1973). *Speech errors as linguistic evidence*. Mouton: The Hague.

Garrett, M.F. (1980). Levels of processing in sentence production. In B. Butterworth (ed.), *Language production* (Vol. 1). pp. 177-221. Academic Press: NY.

Garrett, M.F. (1976). Syntactic processes in sentence production. In R.J. Wales & E. Walker (eds.), *New approaches to language mechanisms*. pp. 231-256. North-Holland: Amsterdam.

Gaskell, M.G. & Marslen-Wilson, W.D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613-656.

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1176-1195.

Gleason, H.A. (1961). *An introduction to descriptive linguistics*. Holt, Rinehart, and Winston: New York.

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Goldinger, S.D. (1997). Talker variability in speech processing. In K. Johnson & J.W. Mullennix (eds.), *Talker Variability in Speech Processing.* pp. 33-66. Academic Press: San Diego.

Goldinger, S.D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166-1183.

Goldinger, S.D. & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics, 31,* 305-320.

Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the locus of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 152-162.

Greenberg, J.H. & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word, 20*, 157-177.

Grosjean, F. & Frauenfelder, U.H. (ed.) (1997). *Spoken Word Recognition Paradigms*: Special Issue of "Language and Cognitive Processes", Psychology Press, Hove: England.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics, 31*, 423-445.

Halle, M. (1956). [Review of the book *Manual of Phonology* by C.D. Hockett]. *Journal of the Acoustical Society of America*, *28*, 509-510.

Hockett, C.D. (1960). The origin of speech. *Scientific American*, *203*, 88-96.

Hockett, C.F. (1955). *Manual of phonology*. Indiana University Publications in Anthropology and Linguistics (No. 11), Bloomington, IN.

Jacoby, L.L. & Brooks, L.R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (ed.), *The psychology of learning and motivation,* pp. 1-47. Academic Press: NY.

Jenkins, J.J., Strange, W., & Trent, S.A. (1999) Context-independent dynamic information for the perception of coarticulated vowels. *Journal of the Acoustical Society of America*, *106*, 438-448.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J.W. Mullennix (eds.), *Talker Variability in Speech Processing*. pp. 145-166. Academic Press, San Diego, CA.

Joos, M.A. (1948). Acoustic phonetics. *Language*, *24*, 1-136.

Kakehi, K. (1992). Adaptability to differences between talkers in Japanese monosyllabic perception. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka, eds. *Speech perception, production, and linguistic structure*, pp. 135-142. Ohmsha Publishing: Tokyo.

Kapatsinski, V.M. (Forthcoming). Towards a single-mechanism account of frequency effects. *Proceedings of LACUS 32: Networks*, Hanover: NH.

Kenstowicz, M. & Kisseberth, C. (1979). *Generative phonology* . Academic Press: New York/London.

Klatt, D.H. (1986). The problem of variability in speech recognition and in models of speech perception. In J.S. Perkell & D.H. Klatt (eds.), *Invariance and Variability in Speech Processing.* pp. 300-319. Erlbaum, Hillsdale: NJ.

Klatt, D.H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279-312.

Kolers, P.A. (1976). Pattern-analyzing memory. *Science*, *191*, 1280-1281.

Kolers, P.A. (1973). Remembering operations. *Memory and Cognition*, *1*, 347-355.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431-361.

Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, *65*. 497-516.

Liberman, A.M., Delattre, P.C., Cooper, F.S., & Gerstman, L.J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, *68***,** 1-13.

Liberman, I.Y., Shankweiler, D., Fischer, F.W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, *18*, 201-212.

Licklider, J.C.R. (1952). On the process of speech perception. *Journal of the Acoustical Society of America*, *24*, 590-594.

Lindgren, N. (1965). Machine Recognition of Human Language. *IEEE Spectrum*, March and April 1965.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception, Technical Report No. 6.* Department of Psychology, Speech Research Laboratory, Bloomington, IN.

Luce, P.A. & McLennan, C.T. (2005). Spoken word recognition: The challenge of variation. In D.B. Pisoni & R.E. Remez (eds.), *The handbook of speech perception*, pp. 591-609. Blackwell Publishing Ltd: Oxford UK.

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

Markman, A.B. & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences*, *4*, 470-475.

Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. A tutorial review. In H. Bouma & D.G. Bouwhis (eds.), *Attention and Performance X: Control of Language Processes.* pp. 125-150. Erlbaum: Hillsdale, NJ.

Marslen-Wilson, W.D. & Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63

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 and Cognition, 15*, 676-684.

Mason, H. (1946). Understandability of speech in noise as affected by region of origin of speaker and listener. *Speech Monographs*, *13*, 54-58.

McNellis, M.G. & Blumstein, S.E. (2001). Self-organizing dynamics of lexical access in normals and aphasics. *Journal of Cognitive Neuroscience, 13*, 151-170.

Miller, G.A., Heise, G.A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test material. *Journal of Experimental Psychology*, *41*, 329-335.

Miller, G.A., Wiener, F.M., & Stevens, S.S. (1946). *Combat instrumentation. II. Transmission and reception of sounds under combat conditions*. Summary Technical Report of NDRC Division 17.3. NDRC (government): Washington, DC.

Morton, J. (1979). Word recognition. In J. Morton & J.C. Marshall (eds.), *Structures and Processes.* pp. 108-156. MIT Press: Cambridge.

Morton, J. & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, *15*, 43-52.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, *47*, 379-390.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, *85*, 365-378.

Murphy, G.L. (2002). *The big book of concepts*. MIT Press: Cambridge, MA.

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

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. (1992). Effects of speaking rate and talker variability on the representation of spoken words in memory. *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, Oct. 12-16, pp. 209-212.

Oldfield, R.C. (1966). Things, words and the brain. *Quarterly Journal of Experimental Psychology, 18,* 340-353.

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. Journa*l of Experimental Psychology: Learning, Memory, and Cognition, 19*, 309-328.

Pardo, J.S. & Remez, R.E. (in press). The perception of speech. In M. Traxler & M.A. Gernsbacher (eds.), *The Handbook of Psycholinguistics*. Elsevier, New York.

Peters, R.W. (1955). *The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise* (Joint Project Report No. 56, pp. 1-9). US Naval School of Aviation Medicine: Pensacola, FL.

Peterson, G. (1952). The information-bearing elements of speech. *Journal of the Acoustical Society of America*, *24*, 629-637.

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.

Pierrehumbert, J.B. (1999). What people know about sounds of language. *Studies in the Linguistic Sciences*, *29*, 111-120.

Pierrehumbert, J.B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure.* pp. 137-158. John Benjamins: Amsterdam.

Pierrehumbert, J.B. & Pierrehumbert, R.T. (1990). On attributing grammars to dynamical systems. *Journal of Phonetics*, *18*, 465-477.

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. Academic Press, San Diego.

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. & Luce, P.A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*, 21-52.

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.

Pisoni, D.B. & Remez, R.E. (Eds.). (2005). *The Handbook of Speech Perception*. Blackwell Publishing: Malden, MA.

Pollack, I. (1953). The information of elementary auditory displays II. *Journal of the Acoustical Society of America*, *25*. 765-769.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America*, *24*. 745-749.

Port, R. & Leary, A. (2005). Against formal phonology. *Language*.

Raphael, L.J. (2005). Acoustic cues to the perception of segmental phonemes. In D.B. Pisoni & R.E. Remez (eds.), *The handbook of speech perception*, pp. 182-206. Blackwell Publishing Ltd: Oxford, UK.

Read, C. (1971). Preschool children's knowledge of English phonology. *Harvard Educational Review*, *41*, 1-34.

Reddy, R.D. (1975). *Speech recognition*. Academic Press: NY.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues, *Science*, *212*(4497), 947-950.

Roediger, H.L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043-1056.

Rozin, P. & Gleitman, L.R. (1977). The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In A.S. Reber & D.L. Scarborough (eds.), *Toward a psychology of reading*. pp. 55-141. Erlbaum: Hillsdale, NJ.

Samuel, A.G. (1981a). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*, 474-494.

Samuel, A.G. (1981b). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1124-1131.

Sapir, E. (1933). La réalité psychologique des phonemes. *Journal de Psychologie Normale et Pathologique*, *30*, 247-265.

Schacter, D.L. (1992). Understanding implicit memory: A cognitive neuroscience approach. *American Psychologist*, *47*, 559-569.

Schacter, D.L. (1990). Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. *Annals of the New York Academy of Sciences*, *608*, 543-571.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science*, *270* (5234), 303-304.

Shattuck-Hufnagel, S. & Klatt, D.H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, *18*, 41-45.

Smith, E.E. & Medin, D. (1981). *Categories and Concepts*. Harvard University Press: Cambridge, MA.

Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1992). Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. *Proceedings 1992 International Conference on Spoken Language Processing,* Banff, Canada, Oct. 12-16, pp. 217-220.

Sporns, O. (2003). Network analysis, complexity, and brain function. *Complexity, 8*, 56-60.

Stemberger, J.P. (1982). *The lexicon in a model of language production*. Unpublished doctoral dissertation, University of California, San Diego.

Steriade, D. (2001a). Directional asymmetries in place assimilation: A perceptual account. In E. Hume & K. Johnson (eds.), *The role of speech perception in phonology*, pp. 219-250, Academic Press, San Diego.

Steriade, D. (2001b). *The phonology of Perceptibility Effects: the P-map and its consequences for constraint organization*. Unpublished manuscript, UCLA.

Stevens, K.N. (1998). *Acoustic Phonetics*. MIT Press: Cambridge, MA.

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E.E. David & P.D. Denes (Eds.), *Human Communication: A Unified View.* pp. 51-66. McGraw-Hill: New York.

Studdert-Kennedy, M. (1983). On learning to speak. *Human Neurobiology*, *2*, 191-195.

Studdert-Kennedy, M. (1974). The perception of speech. In T.A. Sebeok (ed.), *Current trends in linguistics (Vol. XII)*, pp. 2349-2385. Mouton: The Hague.

Treiman, R. (1985). Onsets and rimes as units of spoken syllables. Evidence from children. *Journal of Experimental Child Psychology*, *39*, 161-181.

Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, *15*, 49-74.

Treiman, R. (1980). *The phonemic analysis ability of preschool children*. Unpublished doctoral dissertation, University of Pennsylvania.

Treisman, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review, 78*, 420-425.

Treisman, M. (1978b). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior, 17*, 37-59.

Tulving, E. & Schacter, D.L. (1990). Priming and human memory systems. *Science*, *247*, 301-306.

Twaddell, W.F. (1952). Phonemes and allophones in speech analysis. *The Journal of the Acoustical Society of America*, *24*, 607-611.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, *176*, 392-393.

Whittlesea, B.W.A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, *13*, 3-17.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Modeling the Mental Lexicon as a Complex System: Some Preliminary Results Using Graph Theoretic Measures[1]

Thomas M. Gruenenfelder and David B. Pisoni

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

[1] The research reported in this paper was supported by NIH Grants DC00111 and DC00012. The authors would like to thank Nick Altieri, Vsevolod Kapatsinski, Shane Mueller, and Mike Vitevitch for valuable discussions of the work reported in this paper.

Modeling the Mental Lexicon as a Complex System: Some Preliminary Results Using Graph Theoretic Measures

Abstract. The mental lexicon used for spoken word recognition was modeled as a complex system using tools of graph theory. Words were represented as nodes in the model, and an edge was placed between two nodes if the corresponding words could be changed into one another via a single phoneme deletion, addition, or substitution. The resulting graph had a small-world, scale-free structure. However, the scale-free property reflected the fact that words have different lengths and are created from a relatively small set of phonemes, rather than reflecting the way the network evolves over time. Various network properties of words were also found to be correlated with listeners' performance in an open-set word identification task and in a word repetition task. Those properties included the number of lexical neighbors of a word at different network distances from that word, the mean shortest distance from a word to all other words within the mental lexicon, and a word's clustering coefficient, a measure of the probability that a word's neighbors are also neighbors of one another. The results suggest that including these new measures did not significantly improve the ability to predict the accuracy with which a word can be recognized in various levels of noise using only the word's neighborhood size. In contrast, repetition latencies tended to be longer for words with higher clustering coefficients. Furthermore, this effect did not appear to be modulated by the word's neighborhood size. Possible reasons for the effects of this non-local, global variable are discussed.

Introduction

A number of recent studies have modeled a diverse set of complex systems as graphs or networks (see Albert and Barabási, 2002, for a review). These systems include the structure of the Internet (Faloutsos, Faloutsos, & Faloutsos, 1999) and of the World Wide Web (Huberman, & Adamic, 1999; Huberman, Pirollo, Pitkow, & Lukose, 1998; Lawrence & Giles, 1998, 1999), metabolic interactions (Jeong, Tombor, Albert, Oltavi, & Barabási, 2000), protein-protein interactions (Wuchty, 2001), citation patterns in scientific papers (Newman, 2001), neural networks (Achacosa & Yamamoto, 1992), contacts among potential disease carriers (Liljeros, Edling, Amaral, Stanley, & Aberg, 2001), and different aspects of language, including people's representations of word meanings (Steyvers & Markham, 2004) and the co-occurrence of words in sentences (Dorogovtsev & Mendes, 2001; Ferrer & Solé, 2001). The World Wide Web, for example, can be modeled as a graph in which each web site is represented by a *node*. An *edge* between two nodes is created if the web site represented by one node has a link to that represented by the second node. Graphs can be either *directed*, in which each edge has a particular direction, from one node to the other (web site A links to web site B), or *undirected*, in which case each edge has no specific direction (persons A and B are married to one another).

A question typically asked in such studies is whether the system of interest shows a "small world" (Albert & Barábasi, 2002; Barabási & Albert, 1999; Watts & Strogatz 1998), "scale-free" (Albert & Barábasi, 2002; Barabási & Albert, 1999) structure. In a small-world network, the mean shortest path length between any two arbitrary nodes in the network—that is, the minimum number of edges that must be traversed to get from one of the two nodes to the other—is small relative to the total number of nodes in the network. More precisely, the mean shortest path length grows much more slowly than the number of nodes. Albert, Jeong, and Barabási (1999), for example, found that in the University of Notre Dame intranet, which at the time consisted of over 300,000 documents, any arbitrary document could be

reached from another arbitrary document by traversing on average 11 links. In 1998, in the World Wide Web as a whole, which at the time consisted of over one billion documents, the mean shortest path length between any two documents was estimated to be 19 links (Albert et al., 1999). In a well-known study, Milgram (1967) asked people (the sender) in one part of the United States to forward a letter to another person (the target, who was not known to the sender) in another part of the United States by sending the letter to someone known to the sender who in turn the sender thought might know the target person. This intermediate recipient of the letter, in turn, would either forward it directly to the target, if known, or to another intermediate recipient. Milgram found that for those letters that eventually reached their target, the mean number of intermediate recipients, out of a population at the time of 175 million, was 6. All these values are much smaller than would be predicted for a random network, that is, a network with the same total number of connections, but where the connections were placed between randomly selected pairs of nodes.

Many small-world networks also have a higher than chance clustering coefficient (Albert & Barabási 2002; Watts & Strogatz 1998). The clustering coefficient (CC) is a measure of the probability that two nodes, B and C, are connected, given that a third node A is connected to both B and C. In other words, the CC is a measure of the probability that any two neighbors of a given node are themselves neighbors.

Scale free structure refers to a property of the network's degree distribution (Albert & Barabási, 2002; Barabási & Albert, 1999). A node's degree is the number of edges going into (the in-degree) or out of (the out-degree) the node, or both into and out of the node. The degree distribution is the frequency distribution of node degrees in the network. A scale-free network has a degree distribution characterized by a power law, $N(k) \sim k^{-\gamma}$, where N(k) is the degree distribution, k is the degree (i.e., the number of edges going into or coming out of the node), and the exponent, γ, is typically between 2 and 3. In other words, on a log-log plot, the degree distribution is a straight line with a slope between -2 and -3. In such a degree distribution, most of the nodes have only a very few edges coming into or going out of them. A small number of nodes, however, frequently referred to as hubs, are connected to a very large number of edges.

Demonstrating that a network has a small-world, scale-free structure is important because it potentially has several implications for how the network developed over time (Albert & Barabási, 2002; Barabási & Albert, 1999; Barabási, Albert, & Jeong, 1999). In particular, Barabási and his colleagues have argued that a small-world, scale-free structure indicates that the network evolved over time (1) by adding new nodes, and (2) through a process called "preferential attachment." When a new node is added to a network, a process must exist for it to form edges to other nodes. In preferential attachment, a new node forms edges with an already existing node with a probability that is proportional to the number of edges that existing node already has, i.e., with a probability proportional to the existing node's degree. Such a growth process results in rich nodes—nodes with a large number of edges—getting richer—getting even more edges, a phenomenon sometimes referred to in the psychological literature as the "rich get richer" principle.

The present paper models the human mental lexicon of spoken words as a complex network. The mental lexicon refers to the representations in our brains of the various characteristics of every word we know, including semantic, orthographic, and acoustic-phonetic characteristics. The focus here is on the last of these properties, the acoustic-phonetic properties, or the word's sound structure. Our long-term lexical knowledge is part of what enables the rapid and efficient recognition of speech under a wide range of listening conditions. By comparing the acoustic input to stored representations in the mental

lexicon, and applying some algorithm or heuristic for selecting the best match, listeners are able to recognize each of the individual words spoken by a talker (Luce & Pisoni, 1998).

Traditionally, linguists describe a word's sound as a sequence of phonemes. A phoneme is the smallest unit of sound that distinguishes meaning within a given language. It is an idealized abstract unit in the sense that it does not distinguish all differences in sound but only those necessary to differentiate meaning. The American English word "cat," for instance, though its exact pronunciation varies from utterance to utterance, can be described as consisting of the three phonemes, /k/, /æ/, and /t/, usually written as /kæt/, which is referred to here as the phonetic transcription.[2] The phoneme /k/ distinguishes it from a number of other American English words, such as "mat" (/mæt/), "pat" (/pæt/), and so on. The phoneme /æ/ distinguishes it from other American English words, such as "kit" (/kIt/), and similarly for the phoneme /t/. In American English there are approximately 12 vowel phonemes and 24 consonantal phonemes, the exact number varying by dialect and phonetician.

More than twenty years ago, Nusbaum, Pisoni, and Davis (1984) created an on-line lexicon of nearly 20,000 American English words, based on Webster's Pocket Dictionary (*Webster's Seventh Collegiate Dictionary*, 1967). This lexicon has become known as the Hoosier Mental Lexicon, or HML. The lexicon contains every word in that dictionary, but with homophones and morphemic derivatives eliminated. For instance, the word "dear" (/dIr/) appears in the lexicon; the homophone "deer" (also /dIr/) does not. The word "ask" (/æsk/) appears; the morphemic derivatives "asks," "asking," and "asked" do not. The on-line lexicon contains a phonetic transcription for each word, as well as information such as the word's frequency in printed English (Kucera & Francis, 1967), its orthography (i.e., spelling), its syntactic role(s) (noun, verb, adjective, and so on), and length in number of phonemes.

Following an earlier paper by Vitevitch (2004), an undirected graph was constructed from this lexicon in the following manner. Each word was represented as a node. An edge was created between two nodes if the word represented by one can be turned into the word represented by the other through the deletion, addition, or substitution of a single phoneme. (In the remainder of this paper, this rule is referred to as the Deletion-Addition-Substitution or DAS rule.) Otherwise, no link was placed between the two nodes. This rule has long been used for operationally distinguishing similar sounding words from dissimilar sounding words (Greenberg & Jenkins 1964; Landauer and Streeter, 1973). Two words that can be changed into one another using the DAS rule are referred to as lexical neighbors. A word's total collection of neighbors is referred to as its lexical neighborhood. As an illustration of the rule, the neighborhood of the word "bait" (/bet/) includes the words "late" (/let/), "rate" (/ret/), "bit" (/bIt/), "bail" (/bel/), and "bake" (/bek/), but not the word "sane" (/sen/) or the word "bare" (/bɛr/).

Two primary questions were then asked about the properties of this graph. Does the structure of the lexicon, modeled this way, show a small-world, scale-free structure? Finding that it does have such a structure has potentially strong implications for theories of how children acquire language (Vitevitch, 2004). In particular, this result would suggest that children acquire new words using a process of preferential attachment, learning words that are neighbors of words that they already know. We are not the first to ask if the mental lexicon shows a small-world, scale-free structure. Vitevitch, using the same base lexicon (Nussbaum et al., 1984) and the same single phoneme DAS rule as we used, modeled the mental lexicon as a graph and concluded that it does in fact follow a small-world, scale-free structure. Thus, our initial work provides an opportunity to replicate Vitevitch.

[2] Throughout this paper, the International Phonetic Alphabet (IPA) symbols are used for phonemes. Following the normal conventions in the word recognition literature, phonemic transcriptions are enclosed within forward slashes, /…/, and orthographic transcriptions within quotation marks "…"

The second, and more important, question concerns whether there are global properties of the lexicon, or non-local properties of individual words that affect people's ability to identify particular words? In other words, are there network metrics that correlate with the ease with which individual words are perceived? It is well-known that the larger a word's neighborhood, the more difficult it is to recognize that word (Elman & McClelland, 1986; Luce & Pisoni, 1998; Marslen-Wilson, 1987, 1989; McClelland & Elman, 1986; Norris, 1994; Vitevitch & Luce, 1999). Luce and Pisoni (1998), for example, found that the more neighbors a word had, the more likely was that word to be misperceived in noise, the longer it took for people to discriminate that word from spoken non-words, and the longer it took for people to repeat the word after hearing a spoken version of it.

Luce and Pisoni (1998) developed the Neighborhood Activation Model (NAM) of spoken word recognition to explain these effects as well as other findings in the word recognition literature. In NAM, as in most current models of word recognition, words are assumed to exist in a large multi-dimensional acoustic space (Triesman, 1978a, b). There is no detailed description of how the sound patterns of words in this space are organized or structured. The NAM computes the lexical neighborhood of a word using the DAS rule. When the neighborhood is computed in this way, the resulting similarity space is defined only locally for an individual word based on a one-step distance metric without regard to other perceptually similar words in the lexicon. Thus, in the present architecture of NAM and all other word recognition models, words are not organized in any global structure and consequently there is no consideration of how the word's position in the overall lexicon or even of the structure of its local neighborhood can affect its identification.

Modeling the lexicon as a complex network using graph theory can potentially reveal global structural properties of the lexicon that may influence human word recognition. These properties will then need to be accounted for by current models of word recognition. Two properties were of special concern in the present study, a word's mean shortest distance (or mean shortest path length) to all other words, and a word's clustering coefficient. Word A's shortest distance to Word B is the smallest number of edges that must be traversed to reach Word B from Word A. Word A's mean shortest distance is the mean of that value across all words in the lexicon. Unlike the DAS rule, mean shortest distance takes into account the relative position of a word to all other words in the lexicon, not only to its immediate neighbors. The DAS rule is a coarse measure of phonological similarity that may miss important behavioral consequences. Intuitively, the word "bait" (/bet/) is more similar to the word "bare" (/bɛr/) than it is to the word "epileptic" (/ɛpəlɛptIk/), and the more "bare"s there are in the lexicon relative to "epileptic"s, the harder should it be to identify "bait." The DAS rule does not capture this intuition, but a rule based on mean shortest distance does. Similarly, the clustering coefficient provides a measure of the density of interconnections in a word's neighborhood. Typically, models of spoken word recognition posit the initial activation of multiple lexical candidates followed by a competition among those candidates. The competition ultimately leads to the identification of the word. Neural network models, for example, implement this competition with inhibitory connections between candidate words (cf., Elman & McClelland, 1986; McClelland & Elman, 1986). In such models, the overall density of connections in a word's neighborhood, and not just its neighborhood size, may have a significant effect on its recognition.

As already mentioned, we are not the first to model the mental lexicon as a graph, where the DAS rule is used to place edges between nodes that represent words. Here, we attempt to relate network properties of a word to the ease with which it can be recognized by human listeners.

Method

Constructing the Network

An Excel based version of the HML database was used as the basis for constructing the network. The neighborhood of each word was determined by finding all other words in the lexicon that could be converted to the target word by deleting one phoneme of the target word, adding one phoneme to it, or changing one phoneme to a second valid phoneme of American English. The software for determining neighbors can be publicly accessed at the Washington University (St. Louis, MO, USA) Speech and Hearing Lab website: http://128.252.27.56/neighborhood/Home.asp. The output of this software was converted into a list of pairs of words, where each word pair was a pair of neighbors and the pairs collectively were an exhaustive list of all neighbors in the lexicon. This list was used as input to Version 1.0 Pajek (Batagelj & Mrvar, 1998), a program designed for the analysis of large networks, and available for non-commercial use at http://vlado.fmf.uni-lj.si/pub/networks/pajek/. All statistics on the network, unless explicitly stated otherwise, were calculated using Pajek.

The network was constructed as an undirected graph, rather than directed graph, since, given the definition of neighbor, if Word A is a neighbor of Word B, then Word B must also be a neighbor of Word A.

Behavioral Data

Global properties of the network structure were correlated with two sets of behavioral data: word identification and repetition. The behavioral data were collected as part of an earlier study (Luce & Pisoni, 1998) and reanalyzed here. In the identification task, a digitized recording of a spoken, monosyllabic word was played to a listener at one of three Signal to Noise ratios (SNR): -5 dB SPL, +5 dB SPL, and +15dB SPL. The listener's task was simply to identify the spoken word by typing a response on the computer keyboard. The identification task was open-set. That is, the listener did not choose the correct alternative from a limited set of alternatives, but rather from the entire set of American English words. A total of 90 listeners participated in this experiment in partial fulfillment of an introductory psychology course requirement at Indiana University, Bloomington, IN, USA. Data were collected for a total of 908 monosyllabic words. Because of the large number of words in the study and the use of three SNR, in order to keep experimental sessions to a manageable length, each word was identified by a total of 10 listeners at each SNR. The dependent variable was the percent correct identifications of each word across listeners as a function of SNR.

In the word repetition study, a digitized recording of a monosyllabic word was played to the listener, in the clear (that is, without noise). The listener's task was simply to repeat back the word. Because listeners rarely make errors in this task, the dependent variable of interest was response latency, or the time for the listener to begin repeating the word. Latencies were measured from the offset of the spoken word to the beginning of the listener's utterance. Mean latencies to each word across listeners were analyzed in the present study. Eighteen volunteers from the Indiana University community served as listeners in this experiment. Latencies were collected for 939 words. All of the stimuli were monosyllabic words.

Additional details on the procedure used for collecting both the identification and the repetition data can be found in Luce and Pisoni (1998). The major finding of the earlier Luce and Pisoni study was that percent correct identification of a word decreased as the size of the word's neighborhood increased, while repetition latencies increased with neighborhood size. Luce and Pisoni interpreted these results as

support for the hypothesis that spoken words are recognized relationally in the context of other words in the lexicon.

Results

Network Analysis

The column labeled Whole Corpus in Table 1 shows some of the basic properties of the graph constructed from this lexicon. For comparison purposes, the results from Vitevitch (2004) are shown in Table 1 in the column labeled Vitevitch. In the present study, each word had a mean number of 3.18 neighbors. The clustering coefficient was 0.048 and the mean shortest distance from one word to any other word was 6.08. These properties are nearly identical to those reported by Vitevitch, a not surprising finding, given that we started with the same base lexicon as he did.

Table 1. Parameters of the networks based on the whole HML corpus and on monosyllabic words. The column labeled Vitevitch shows the data from Vitevitch (2004). The column labeled Random Networks is taken from Vitevitch (2004).

Parameter	Whole Corpus	Monosyllabic Corpus	Vitevitch	Random Networks
Number of Nodes (n)	19,587	4110	19,340	19,340
Number of Edges (l)	3.18	11.56	3.23	3.23
Mean Shortest Path Length (l)	6.08	4.66	6.05	8.44
Maximum Path Length (D)	29	13	29	19
Clustering Coefficient (CC)	0.048	0.10	.045	.000162
Degree Exponent (γ)	1.97	n/a	1.96	n/a

Vitevitch (2004) also constructed 10 random networks in which he fixed the number of nodes and the number of edges to be the same as in the network constructed from the lexicon. However, instead of using a rule to determine which pairs of nodes were to be connected by an edge, the two nodes connected by each edge were chosen randomly. The properties of these graphs are shown in Table 1 in the column labeled Random Network. The values in this column are the means across the 10 random networks. Notably, the mean clustering coefficient in the 10 random networks was significantly less than the clustering coefficient observed in the network based on the lexicon, and the mean of the mean shortest path length was significantly longer in the random networks. These two observations—that the mean shortest path length was less than that expected by chance and that the clustering coefficient was greater than that expected by chance—suggest that the lexical network created using the single DAS rule follows a small-world structure (Watts & Strogatz, 1998). Vitevitch drew the same conclusions from these observations.

Given the small-world structure of the lexicon, we next asked if it also follows a scale free structure. Recall that a network that has a scale-free structure is characterized by a degree distribution

that follows a power law, with the power law's parameter, γ, being in the range 2 – 3, and where a node's degree is the number of edges to which it connects. Figure 1 shows the degree distribution on log-log coordinates for the HML lexicon.[3]

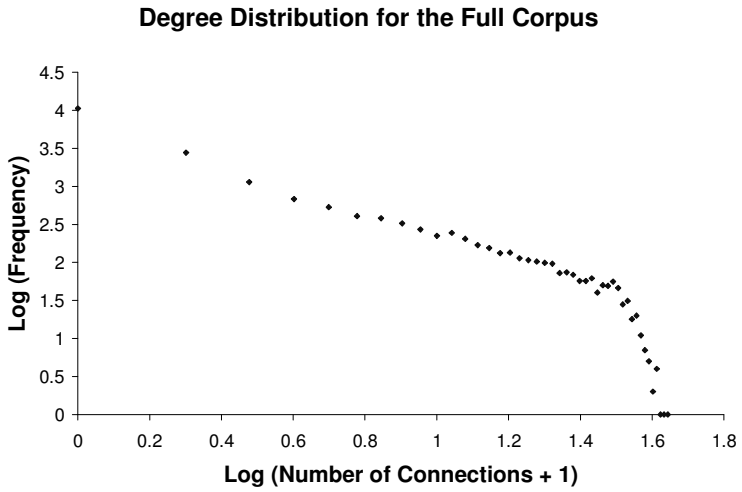


Figure 1. Log-log degree distribution (number of occurrences as a function of degree) for the 19,587-word corpus.

Ignoring momentarily the sharp drop in frequency at degrees higher than approximately 36, the degree distribution appears to be linear on a log-log plot. The Pearson product moment correlation between frequency and degree (including the degrees after the sharp drop off evident in the curve) is -0.85 ($p < .01$), indicating that the degree distribution can be reasonably fit with a straight line. The best fitting straight line for that distribution has a slope of -1.97, a value not far out of the range of -2 to -3 characteristic of scale-free distributions (Albert & Barabási, 2002; Barabási & Albert, 1999). This slope is in very close agreement with the slope of -1.96 found by Vitevitch (2004).

Two further observations, however, suggest that the scale-free properties of the lexicon may be an illusion. First, using the DAS rule, most words in the HML lexicon have no neighbors. This fact is evident in Figure 2, which shows the degree distribution on linear coordinates. Note that the plot begins with degree = 0, not 1. Note also the large number of words with a degree of 0. In fact, over 10,500 words, representing roughly half the lexicon have no neighbors at all when the network was created using the DAS rule. Although it is true that had we built our network using a different definition of lexical neighbor, we might not have found such a large number of isolates, it is also true that the definition of neighbor that we did use has been found to be a powerful predictor of performance on a wide variety of experimental tasks involving spoken word recognition (e.g., Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998; Vitevitch, 2002; Vitevitch & Luce, 1999) and hence seems to be appropriate at least as a starting point for building the graph.

[3] The X axis in Figure 1 is actually the logarithm of the (degree plus 1). As will be discussed in more detail, many words have no neighbors, i.e., have a degree of 0. Since the logarithm of zero is undefined, the adjustment of adding 1 to the degree was made before plotting on log-log coordinates.

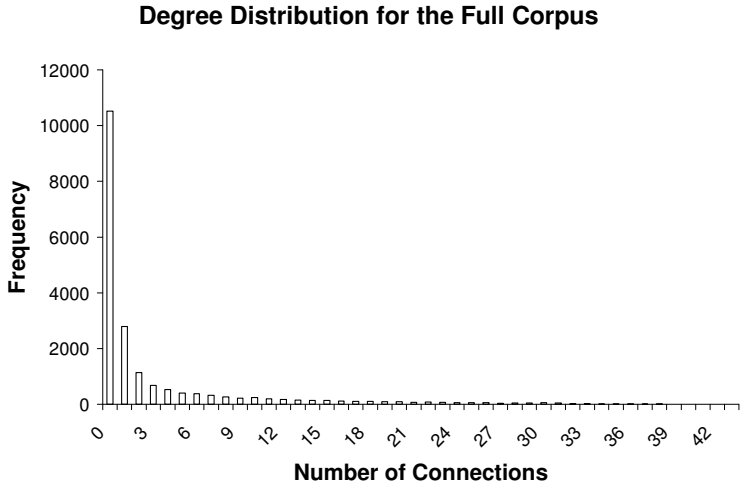


Figure 2. Degree distribution for the full corpus on linear-linear coordinates.

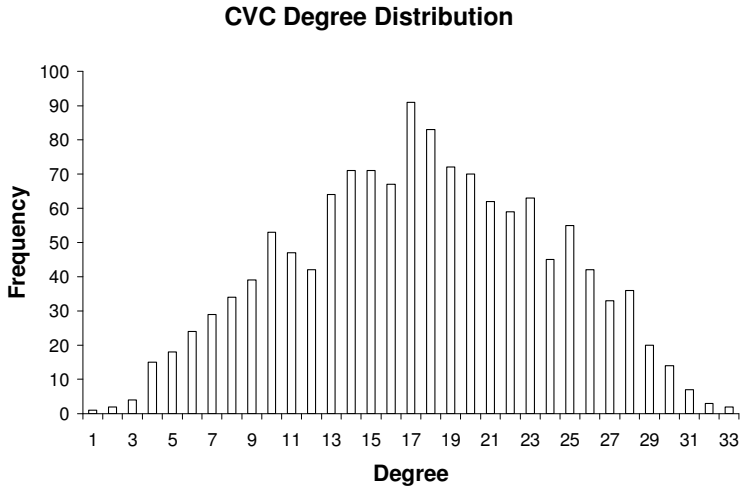


Figure 3. Degree distribution for the CVC network on linear-linear coordinates.

The second and more crucial observation indicating that this network is not scale free involves examination of particular subsets of the data. Figure 3 shows the degree distribution, on linear coordinates, for a network built according to the same deletion-addition-substitution rule as the original network, but including only those 1338 words with a CVC structure, that is, words consisting of an initial consonant, followed by a single vowel, followed by a single final consonant. This distribution clearly does not follow a power law, but more closely resembles a Poisson distribution. This kind of distribution would be expected if links between nodes were placed at random. Figure 4 shows the degree distribution for the subset of the corpus that includes all monosyllabic words, that is words with a single vowel

(CVC, CCVC, CVCC, and so on). This distribution also clearly does not follow a power law. Inspection of Figures 2, 3, and 4 makes it clear that the middle and right hand portions of the overall degree distribution reflect primarily the contribution of monosyllabic words. These words tend to have more neighbors simply because they are shorter in length, not because of how the network grew. The left hand portion of the distribution is determined primarily by multi-syllabic words, which tend to have few neighbors simply because they are longer. Hence, any resemblance of the overall degree distribution of the lexicon to a power law merely reflects differences in the length of words rather than the acquisition and development processes used to create the network. To summarize, there is good evidence that the mental lexicon used for spoken word recognition has a small-world structure, but there is little evidence that it is scale-free.

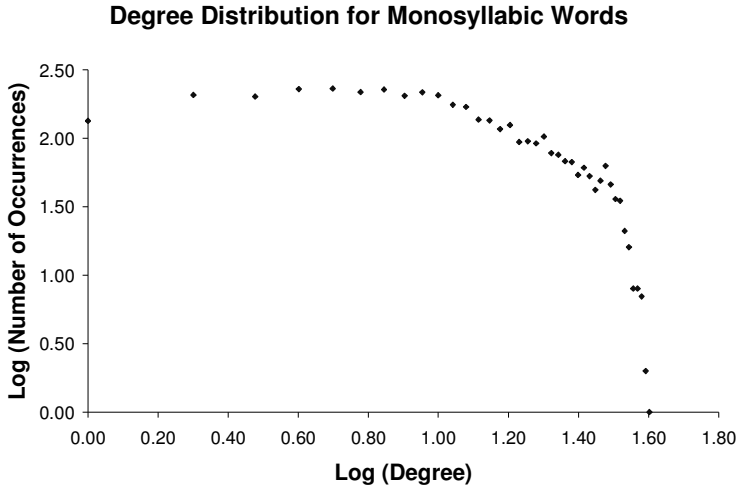


Figure 4. Degree distribution for the monosyllabic network on log-log coordinates.

Correlations with Behavioral Measures of Spoken Word Recognition. Our primary concern in this project is not with establishing whether the mental lexicon follows a scale-free structure, but with determining whether there are fundamental properties of the lexicon, when modeled as a complex network graph, that correlate with the processing and perception of spoken words. Accordingly, we correlated several characteristics of a word's neighborhood in the lexical network with the accuracy with which people could recognize that word in noise, and the speed with which people could repeat that word when it was spoken to them in isolation. It is now well-known that as a word's "local" neighborhood density increases, where density is the number of immediate neighbors, as defined by the DAS rule, weighted by the frequency of occurrence of each of those neighbors, the accuracy of the identification of that word decreases (Luce & Pisoni, 1998; Vitevitch & Luce, 1998, 1999). Percent correct in a word identification task decreases and latency in a word repetition task increases as density increases (e.g., Luce & Pisoni, 1998), a finding that supports models of spoken word recognition that posit a stage of processing in which candidate words are first hypothesized to be activated by the acoustic-phonetic input, and then compete with one another for recognition through, for example, inhibitory connections in a neural network (cf. Goldinger et al., 1989; Vitevitch & Luce, 1998). In our network, these immediate neighbors are adjacent words, separated by a single link. What about words further removed from one another? Does the number of words separated by two links or three links in the lexical network from a target word also affect the recognition of the target word? More generally, does a word's average

distance from all other words in the network affect its recognition? And, does the density of the interconnections within a word's neighborhood, as measured by the clustering coefficient, affect that word's recognition?

Our original corpus on which the network was built consisted of over 19,500 words, of which 10,521 had no neighbors. The two most salient characteristics of these 10,500 hermits was that they were longer than words with neighbors in terms of number of phonemes (7 – 8 phonemes per hermit compared to 4 – 5 for words with at least one neighbor) and they were multi-syllabic. Over 98% of the hermits in this network were multi-syllabic words. In contrast, of the 9066 words with at least one neighbor, only 5087, or 56% were multi-syllabic. In addition, the words for which we have behavioral data are all monosyllabic words. In fact, almost all work on spoken word recognition has been done with short, monosyllabic words, at least in part because recognition data for multi-syllabic words can be more easily contaminated by post-perceptual guessing strategies. For these reasons, for our analyses of spoken word recognition data, we rebuilt the lexical network using neighborhoods only for monosyllabic words. The column labeled "Monosyllabic" in Table 1 shows some characteristics of the new network. Of the 4110 monosyllabic words in our original corpus, only 131, or 3.2%, were lexical hermits. The correlations of behavioral data with network structure we report here are based on this new network built from monosyllabic words. The results, however, are qualitatively the same when the network based on the entire lexicon is used in these calculations.

Word Identification Results. Table 2 shows Pearson product moment correlations between a word's percent correct identification and various measures of its structure in the network, and with the logarithm of its frequency of occurrence in the language. Correlations are shown for each of the three SNR used. A single asterisk indicates a statistically significant correlation at the $p < .01$ level. A double asterisk indicates a statistically significant correlation at the $p < .001$ level. In the table, D*m* refers to the number of words whose shortest path in the network from the target word is exactly *m* links. The value D8 was included as a control condition. Words 8 links removed from a target would not be expected to affect its recognition and hence correlations with D8 would expected be near 0. PL refers to path length. Mean PL*m-n* refers to a word's mean shortest path distance from all words whose shortest distance is from exactly *m* through and including exactly *n* links. Hence, mean PL1-13, because the maximum shortest distance between any two words in the network was 13 links, is the mean distance of the word from all other words in the network (excluding the 3.2% of the words with no immediate neighbors). Mean PL1-3 is the mean shortest distance from the target word to all other words exactly 1, 2 or 3 edges distant from it. This value goes up as the proportion of words 2 or 3 links, as opposed to 1 link, from the target word increases. The measure PL1 – 13 was included in order to assess a word's mean distance from every other word on its recognition. The measure PL2 – 13 was likewise included in order to assess the effects of the word's mean distance from all other words, but without also including effects of the number of its immediate neighbors, an already well-studied variable. The measures PL1 – 3 and PL2 – 3 were included in order to assess possible effects of a word's more immediate neighborhood without diluting those effects by simultaneously including effects of far away communities. Finally, CC is the clustering coefficient, a measure of the probability of two word's being neighbors when each is a neighbor of some other third word. The mean proportions correct for the three SNR of +15, +5 and -5 dB were .77, .57, and .17, respectively, suggesting that any correlations of the predictor variables with the percent correct data would not be artificially lowered due to floor or ceiling effects.

Table 2. Correlations of percent correct identification with log frequency of occurrence and with network measures of distance, path length, and clustering for monosyllabic words for three different signal to noise ratios. See the text for an explanation of abbreviations used. *: $p < .01$; **: $p < .001$.

	Log Freq	**D1**	**D2**	**D3**	**D8**	**CC**	**PL1-13**	**PL1-3**	**PL2-3**	**PL2-13**
SNR +15	0.18**	-0.03	-0.04	-0.04	0.02	-0.03	0.03	0.00	0.01	0.03
SNR +5	0.23**	-0.11**	-0.13**	-0.14**	0.05	-0.08*	0.12**	0.08*	0.10*	0.12**
SNR -5	0.20**	-0.04	-0.06	-0.05	0.03	-0.03	0.05	0.05	0.06	0.05

Replicating previous findings, the correlation between a word's frequency of occurrence in the language, as measured by Log Freq, and the accuracy with which it is identified, while low, were statistically significant, at least for the SNR +5 and SNR -5 conditions. The correlations between identification accuracy and network measures, on the other hand, were uniformly low and statistically non-significant. The correlations between the accuracy with which a word is identified and the number of neighbors it has at short distances (D1, D2, D3) were negative in sign and hence in the expected direction—as the number of neighbors increases, identification accuracy decreases. However, the correlations were statistically indistinguishable from 0.

Table 3. Correlations of percent correct identification with network measures of distance, clustering, and path length for monosyllabic words with a low, medium, and high frequency of occurrence. *: $p < .01$; **: $p < .001$.

	Log Freq	**D1**	**D2**	**D3**	**D8**	**CC**	**PL1-13**	**PL1-3**	**PL2-3**	**PL2-13**
Frequency < 20 (N = 512)										
SNR 15	0.15**	-0.02	-0.03	-0.05	0.02	0.01	0.03	0.04	-0.02	0.03
SNR 5	0.17**	-0.10	-0.12*	-0.12*	0.03	-0.02	0.09	0.05	0.04	0.09
SNR -5	0.15**	-0.04	-0.08	-0.08	0.04	-0.02	0.07	0.01	0.07	0.07
20 <= Frequency <= 99 (N = 228)										
SNR 15	.18*	-.10	-.11	-.10	.09	-.12	.10	.09	.10	.10
SNR 5	.08	-.17*	-.20*	-.21**	.14	-.18*	.20*	.16*	.17*	.20*
SNR -5	.17*	-.11	-.10	-.09	.06	-.14	.09	.17*	.16*	.09
Frequency >= 100 (N=168)										
SNR 15	.04	-.01	.00	.00	-.07	-.08	-.02	.01	.00	-.02
SNR 5	.02	-.13	-.16	-.15	-.05	-.13	.13	.07	.10	.13
SNR -5	-.04	.00	.00	.02	-.03	.04	-.02	-.01	-.01	-.02

Because a word's frequency of occurrence in the language can modulate the effects of other variables on its recognition, we re-analyzed the identification data after dividing the words into low frequency (fewer than 20 occurrences per million), medium frequency (from 21 to 99 occurrences per million), and high frequency (100 or more occurrences per million) words. These correlations are shown

in Table 3. As is the case for the overall analysis, the correlations with network parameters were low and statistically non-significant, with the exception of medium frequency words (and to a lesser extent, high frequency words) presented at a SNR of +5 dB. For these words, identification was less accurate the more neighbors the word had that were separated by 1, 2, or 3 phonemes.

Table 3 also shows that for medium frequency words presented at a Signal-to-Noise ratio of +5 dB SPL, the various PLn – m measures also correlate significantly with identification accuracy. As a word's path length, that is, its mean distance to other words in the lexicon, decreases, so does the ability of listeners to correctly identify it. The importance of all these correlations, however, needs to be considered in light of the how the various measures of network structure inter-correlate with one another, in particular with how the D = 1 measure correlated with the D = 2, D = 3, and the PLn – m measures. Table 4 shows these inter-correlations for the entire monosyllabic corpus. Note first that the metric D = 1 is the same variable that previous investigators have referred to as neighborhood size or neighborhood density, a variable already known to affect percent correct identification of words spoken in noise (and word repetition latencies). As is evident from Table 4, this measure was strongly correlated with the D = 2, D = 3, and PLn – m measures (all R^2's > .65, $p < .001$) and when the effects of density were partialed out from these other measures, they no longer significantly correlated with percent correct identification. This finding suggests that our observation that a word's mean distance to other words in the lexicon correlated with listeners' ability to identify it can be entirely accounted for by the word's local neighborhood size.

Table 4. Inter-metric correlations among distance, clustering, and path length measurements for the monosyllabic corpus.

	Log Freq	**D=1**	**D=2**	**D=3**	**D=8**	**CC**	**PL1-13**	**PL1-3**	**PL2-3**	**PL2-13**
Log Freq	1.00	0.05	0.05	0.04	0.00	0.03	-0.93	-0.02	-0.03	-0.03
D =1		1.00	0.94	0.88	-0.48	-.78	-0.84	-0.81	-0.86	-0.83
D=2			1.00	0.96	-0.53	0.77	-0.91	-0.73	-0.84	-0.91
D=3				1.00	-0.63	0.73	-0.97	-0.56	-0.70	-0.97
D=8					1.00	-0.46	0.78	0.12	0.32	0.78
CC						1.00	-0.73	-0.68	-0.74	-0.73
PL1-13							1.00	0.50	0.66	1.00
PL1-3								1.00	0.96	0.49
PL2-3									1.00	0.65
PL2-13										1.00

Finally, again for medium frequency words presented with an SNR of +5 dB SPL, as shown in Table 3, there was a statistically significant albeit small negative correlation between a word's clustering coefficient and its percent correct identification. As a word's clustering coefficient increased, its percent correct identification decreased. Further consideration of the possible effects of the clustering coefficient is postponed to the description of the results for word repetition latencies, where the effects of the clustering coefficient are more robust.

To summarize, in the present data set, there is little evidence that non-local or global properties of the lexicon improve the ability to predict word identification accuracy above and beyond the ability of word frequency and the single DAS rule. There are several possible reasons why we failed to observe

such an effect. First, the DAS rule is a rather crude measure of the perceptual similarity of two words, as evidenced by the fact that over half the words in our original corpus are, by the DAS rule, hermits—i.e., they have no neighbors. Second, given the method we used to construct the network, two words that differ by two phonemes are separated by a distance of 2 if, and only if, there is a third word that is separated by a single phoneme from each of those two words. If no such word exists, then the two words are separated by a distance of greater than 2 in the network. There seems no *a priori* reason to believe that the two words in the former case are more similar than the two words in the latter case. Hence, a similarity measure more refined than the DAS rule may make the effects of global properties of the mental lexicon on word identification more evident. We are currently exploring this possibility in more detail.

Word Repetition Latencies. Pearson product moment correlations of a word's repetition latency and various network properties are shown in Table 5. Latencies were measured from the offset of the spoken word to the onset of the repetition. Hence, they reflect both the time to perceive the word and the time to initiate the motor program for pronouncing the word. As is the case for the identification data, in the table, D*m* refers to the number of words whose shortest path in the network from the target word is exactly *m* links. PL refers to path length. Mean PL*m-n* refers to a word's mean shortest path distance from all words whose shortest distance is from exactly *m* through and including exactly *n* links. Finally, CC is the clustering coefficient, a measure of the probability of two words being neighbors when each is a neighbor of some other third word.

Table 5. Correlations of word repetition latencies with measures of network distance, path length, and clustering as a function of a word's frequency of occurrence. *: $p < .01$; **: $p < .001$.

Log Freq	D1	PL 1-13	PL 1-3	PL 2-3	PL 2-13	CC
Entire Corpus (N = 939)						
-.07	.26**	-.28**	-.26**	-.27**	-.28**	.20**
Frequency < 20 (N = 467)						
.01	.20**	-.21**	-.20**	-.22**	-.21**	.15**
20 <= Frequency <= 100 (N = 275)						
-.03	.27**	-.30**	-.26**	-.27**	-.30**	.21**
Frequency > 100 (N = 197)						
-.07	.48**	-.48**	-.44**	-.44**	-.48**	.40**

Consistent with the earlier findings reported by Luce & Pisoni (1998), repetition latencies did not correlate with frequency of occurrence in the language. However, they did correlate with several of the network parameters. First, a statistically significant positive correlation was observed between the number of immediate neighbors a word has (D1) and its naming latency. The more local neighbors a word has, the longer was the naming latency. At first glance, the data also suggest that words at distances beyond immediate neighbors also influenced word repetition latencies. Significant negative correlations ($p < .01$) were also observed for naming latency with the mean shortest path to for all words at a distance of 2 to 3 edges (PL 2-3), and for naming latency with average shortest distance to all words in the network (PL 1-13), even when words at a distance of 1 (i.e., immediate neighbors) were excluded (PL 2-13). That is, as mean shortest path lengths increased, the time to repeat a word decreased. These correlations need to be interpreted with some caution, however. Given the methods used to construct the original graph, the number of neighbors a word has at a distance of 2 (and at distance 3, and so forth) must correlate with the number of neighbors it has at distance 1 (because to reach a word in two edges, some other word needs to be reached in one edge), meaning that all PLn-m measures would also tend to

(negatively) correlate with D1. In fact, as can be seen in Table 4, PL2-3, PL1-13, and PL2-13 all did correlate strongly with D1 (all R^2's $> .75$, $p < .001$).When the effects of D1 are partialed out of PL2-3, PL1-13, and PL2-13, the correlations of these variables with repetition latency all become non-significant. Hence, we can conclude that extending distance measures beyond a word's immediate neighbors, as defined by the single phoneme deletion/addition/substitution rule, does little to improve the prediction of repetition latencies.

The results obtained from the analyses of the clustering coefficient (CC) are more revealing. CC correlated significantly with repetition latency, r = +0.20, $p < .02$. As the CC became larger, naming latencies increased in duration. The effect tended to be stronger as frequency of occurrence in the language increased. The correlation of CC with repetition latency was +0.15 (n.s.), +0.21 ($p < .02$), and +0.40 ($p < .001$) for low, medium, and high frequency words, as defined above.

We also found that CC correlated strongly with D1, the number of immediate neighbors (r = +0.76, $p < .001$), leaving open the possibility that the observed correlation between CC and repetition latency was in fact due to the correlation of D1 with repetition latency, or conversely, that the observed correlation between D1 and repetition latency was in fact due to the correlation of CC with repetition latency. To determine whether there are independent effects of CC and D1 on naming latencies, a median-split analysis was performed on the data. Each word for which naming latency data were available was assigned to one of four cells in a 2 X 2 ANOVA, corresponding to whether the word was above or below the median value for CC and above or below the mean value for D1. LowCC/LowD1 words (n = 351) were below the median value for both variables, lowCC/HighD1 (n = 121) words were below the median CC value and above the median D1 value, HighCC/LowD1 words (n = 93) were above the median CC value and below the median D1 value, and HighCC/HighD1 words (n = 374) were above the median value for both variables. The means for these four cells are shown in Table 6. An analysis of variance found no significant interaction between D1 and CC ($F < 1$). The main effects, however, of both D1 and CC, though numerically small, were both highly significant, (for D1, $F(1, 937) = 51.84$, $p < .001$; for CC, $F(1, 937) = 30.80$, $p < .001$).The results of this analysis suggest the presence of an effect of clustering coefficient on the latency to repeat a word that is independent of the effect of the word's number of neighbors on the time to repeat that word. Note that the CC is a non-local, global property of the word that appears to affect response times in at least one spoken word recognition task.

Table 6. Mean word repetition latencies (ms) as a function of neighborhood density and clustering coefficient.

	Clustering Coefficient	
Density	Low	High
Low	294	304
High	316	326

The present analysis makes two contributions to our understanding of the structure of the mental lexicon and how it affects spoken word recognition. First, replicating Vitevitch's (2004) earlier findings, we found that the mental lexicon, when modeled as a complex graph, displays a scale-free structure. However, rather than reflecting growth through preferential attachment, we suggest that this structure reflects the constraints that all words are constructed from a limited number of phonemes and that words differ in length. Second, our ability to recognize an isolated word is affected not only by the number of

neighbors that word has, but also by how interconnected those neighbors are with one another, as measured by the clustering coefficient of a word. We discuss each of these points in turn below.

The Mental Lexicon as a Scale-Free Structure

As we mentioned in the introduction, Vitevitch (2004) has also recently modeled the mental lexicon as a complex network. Our results for the various network measures we examined closely follow his, a finding that is not surprising given that we both began with the same HML lexical database (Nussbaum et al., 1984), and we both used the same single phoneme DAS rule. Vitevitch concluded that the mental lexicon has a small-world, scale-free structure. He further suggested that his findings indicate that the mental lexicon grows via a process of preferential attachment (Albert & Barabási, 2002; Barabási & Albert, 1999). That is, as a child learns new words in his or her language, he/she adds words that are acoustically similar to those already learned. In fact, Storkel (2004), as noted by Vitevitch (2004), has recently reported that words learned early by a child are words that have many neighbors in the adult mental lexicon, consistent with the notion that the child's lexicon grows through a process like preferential attachment.

Our results suggest an alternative explanation of why a power-law like degree distribution occurs for the lexicon. Words are constructed from a small number of basic sounds, i.e., phonemes or particles (cf. Abler, 1989). Furthermore, words can have different lengths; some have more phonemes than others. As a consequence, and given the DAS rule used to construct the lexical network, short words will have more neighbors than longer words. The overall result will be a degree distribution that looks quite similar to a power law distribution, but which arises primarily from random processes, not preferential attachment.

The results from our analysis of monosyllabic words support this explanation. The degree distribution for these words did not approach anything resembling a power law. Likewise, the degree distribution for multi-syllabic words also does not resemble a power law. The power law degree distribution is the result of averaging the degree distribution for monosyllabic words, which contributes the "middle" and "right hand (hub)" sides of the distribution with that for multi-syllabic words, which contributes the "left hand" side of the overall degree distribution.

The present study is not the first investigation to find power law frequency distributions arising from averaging non-power law distributions. Improvements with practice frequently follow a power law, and a great deal of research has been directed at determining why this phenomenon occurs (e.g., Newell & Rosenbloom, 1981). Anderson (2001) and Brown and Heathcote (2003a, 2003b; see also Newell & Rosenbloom, 1981) have shown that a power law distribution frequently results when a number of underlying distributions, none of which itself is a power law distribution, are averaged together. For example, power law learning curves can result when individual learning curves for a number of experimental subjects are averaged together, where each individual's curve is an exponential, but each with a different parameter. In Vitevitch's (2004) case, and in our analyses, the degree distribution of monosyllabic words was effectively averaged with the degree distribution of multi-syllabic words (or more precisely, degree distributions for words of different lengths were averaged together), with a similar result. The extent to which this effect underlies other observations of power law degree distributions in other analyses of complex systems remains to be seen.

Nevertheless, it does seem to be the case that words that children learn early do have more neighbors in the adult lexicon (Storkel, 2004). This finding suggests that a process akin to preferential attachment operating at the level of words' acoustic patterns is operating during the course of language

development. At the same time, however, semantics also plays an important role in shaping the organization of the lexicon. A child will learn those words that are most important to meeting its perceived needs, independent of the underlying sounds that comprise those words. That is, the meaning of the word and its relevance to the child is at least as likely to influence which words are added to the child's lexicon as the acoustic-phonetic similarity of the word to other already learned words in the lexicon.

In summary, a complex network constructed from the lexicon using the DAS rule does display a scale-free structure. However, we suggest that the scale-free structure is as likely to reflect the averaging of degree distributions across words of different lengths as it is to reflect fundamental underlying language acquisition processes using preferential attachment.

Effects of the Clustering Coefficient on Spoken Word Recognition

The second purpose of the current study was to determine if global measures of network structure would provide novel insights into spoken word recognition processes previously overlooked by more traditional analyses. To this end, we examined the correlations of various network measures, including measurement of distance and clustering, on word repetition latencies. It is already known that as the number of words at a distance of 1 from a given word increases, so does its repetition latency and the accuracy of identifying that word in noise (Luce & Pisoni, 1998), and the current study replicated these findings. However, we also found that taking into account the mean distance of a word from all other words in the lexicon, or even its mean distance to relatively nearby words, adds little additional predictive power to this original measure. Hence, beyond a distance of 1, distance measurements based on the DAS rule, showed little relation to repetition latencies. This finding suggests that similarity effects drop off sharply as similarity decreases. Alternatively, the present findings suggest that network distance, where the network is built using the single DAS rule, is not a very robust measure of phonological similarity. A *prima facie* case could be made for the second of these two alternatives. Consider two words, A and B, which differ by two phonemes. If a third word, C, exists, such that A can be converted to C by the single DAS rule and C to B by the single DAS rule, then A and B would be separated by a network distance of 2 in our model. If no such word C exists, then the network distance between A and B would be greater than 2, a seemingly somewhat artificial situation. Despite the superficial reasonableness of the second of the two above alternatives, we are currently investigating other objective measures of perceptual similarity in an effort to select the best explanation for the process of spoken word recognition.

The findings obtained with the clustering coefficient were different, however. We found that the higher a word's clustering coefficient, the longer its repetition latency. In other words, repetition latencies were longer for words whose neighbors were also neighbors of one another. This correlation appears particularly strong for higher frequency words. The clustering coefficient thus appears to be an example of a "global," non-local, emergent property that affects the recognition of a particular word.

This result is of course *post hoc*, and like any *post hoc* finding, this finding needs to be independently verified in a replication. In addition, the word repetition task involves both perceptual and production processes. Participants must first correctly perceive the word and then execute a motor program for pronouncing the word out loud. The current study does not address the issue of whether the effects of the clustering coefficient are on perceptual processes, production processes, or both. In collaboration with Nick Altieri, we are currently undertaking a series of new studies in our laboratory designed to address two questions. First, is the effect of the clustering coefficient on repetition naming latency reliable? Second, assuming the effect is reliable, to what extent does a word's clustering

coefficient affect perceptual processes and to what extent does it affect production processes? Although these experiments are still in an early stage, preliminary results are encouraging in terms of confirming the reliability of the effect. The results also indicate that at least part of the effect is on perceptual processes used at the time of encoding.

What are the implications of the effects of clustering coefficient on repetition latency for models of spoken word recognition and speech production? To a large extent, until the locus and replicability of the effect are better delineated, any detailed answer to the question is a little premature. However, at least one general comment can be made at this time. As noted several years ago by Goldinger, Luce, Pisoni, and Marcario (1992), most contemporary models of spoken word recognition (Elman & McClelland, 1986; Luce & Pisoni, 1998; Marslen-Wilson, 1987, 1990; McClelland & Elman, 1986; Norris, 1994) are "activation-plus-competition" models. In such models, the acoustic input activates a number of possible lexical candidates that are each consistent with that input. Each candidate then competes with the other lexical candidates for recognition. In connectionist terms, the competition is often modeled as mutual inhibition of acoustically similar words (Elman & McClelland, 1986; McClelland & Elman, 1986). Models with such an inhibitory mechanism might expect a higher clustering coefficient to actually facilitate word recognition, a result opposite in direction to the findings we have observed. Facilitation would occur because a target's neighbors, when the clustering coefficient is higher, would tend to inhibit one another, reducing the overall activation of the target's immediate neighbors, thus reducing the inhibition they exert on the target itself.

The negative effects of a high clustering coefficient on spoken word recognition, however, might be more easily understood if viewed from a slightly different perspective. Saying that a word has a high CC is another way of saying that a word's neighbors are not only similar to the word itself, but that they are also similar to one another. In contrast, the neighbors of a word with a low CC are similar to that word but not to one another. This observation in turn implies that a high CC word is going to share any given phoneme sequence with many other lexical neighbors; if it did not, those neighbors would not be neighbors of one another. For a low CC word, on the other hand, there will be at least some phoneme sequences that are not shared with many neighbors. For instance, the high CC word "boot" (/but/) shares the phoneme sequence /ut/ with root, loot, soot, coot. It shares the "sequence" /b_t/ with bit, bat, but, bout. Hence, the acoustical evidence for any given sequence is not likely to be very good at discriminating amongst the various alternatives, making it harder to eliminate alternatives. In the case of low CC words, however, having available partial phonological information about a possible word will keep in play fewer neighbors of the target word, since different neighbors of a target word are neighbors for different reasons. In other words, any partial information is more discriminating in the case of low CC words, making overall identification easier.

In addition to these general theoretical issues, we are also interested in using network concepts to explore the lexical organization and processing of spoken words in profoundly hearing-impaired children with cochlear implants (CI). Kirk, Pisoni, and Osberger (1995) found that in an open-set spoke word recognition identification task, deaf children with cochlear implants recognized high frequency words from sparse neighborhoods more accurately than low frequency words from dense neighborhoods. This pattern of results follows that found with normal hearing adult listeners. Based on such results, Kirk et al. suggested that children with CIs may organize their lexicon in a manner similar to normal hearing adults. However, the speech signal received processed by a CI is a highly degraded signal. This degradation could result in the blurring of some phonetic distinctions. The end-result of such blurring may be that the lexical network of CI listeners is much more densely inter-connected than the lexicon of normal listeners. This higher degree of connectivity could contribute to at least some of the difficulty CI listeners experience with spoken word recognition. Higher connectivity, for example, would tend to lead to higher

average clustering coefficients, which in turn, as suggested by the results reported here, may lead to increased difficulty with spoken word recognition in open-set tasks.

To summarize, the present study was designed to investigate whether modeling the lexicon as a complex network, using the tools of graph theory, could provide some new insights into the processes underlying spoken word recognition. In particular, we were interested in whether spoken word recognition is affected by structural properties that go beyond a word's local, immediate lexical neighborhood and reflect more global properties of the mental lexicon, or connectivity patterns of words in the mental lexicon. On the one hand, we found evidence that the size of a word's neighborhood, as originally defined by Landauer and Streeter (1973) (see also Greenberg & Jenkins, 1964, and Luce & Pisoni, 1998), is a robust predictor of the similarity effects observed in open-set word recognition and word repetition tasks. These findings suggest that a complex systems approach may add little to our earlier understanding of similarity effects in spoken word recognition. This conclusion does need to be qualified by two additional observations. First, we analyzed behavioral data only for monosyllabic words, which included approximately 20% of the original corpus. Hence, our observations may not generalize across the entire mental lexicon. Second, we constructed our network using the DAS rule which admittedly is at best a rather crude measure of the perceptual similarity of two words. On the other hand, even under these constraints, we identified a new global variable, the clustering coefficient, that does appear to affect spoken word recognition nearly as robustly as does local neighborhood size. This second finding, if verified and replicated, suggests that non-local, global properties of a word's position in the mental lexicon can have important consequences for how listeners recognize that word in isolation and in context.

References

Abler, W.L. (1989). On the particulate principle of self-diversifying systems. *Journal of Social & Biological Structures*, *12*, 1–13.

Achacosa, T.B. & Yamamoto, W.S. (1992). *AY's Neuroanatomy of C. elegans for Computation*. CRC Press: Boca Raton, FL.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, *401*, 130 - 131.

Anderson, R.B. (2001). The power law as an emergent property . *Memory & Cognition, 29*, 1061-1068.

Batagelj, V. & Mrvar, A. (1998). Pajek—A program for large network analysis. *Connections*, *21*, 47–57.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, *272*, 173–187.

Brown, S. & Heathcote, A. (2003a). Bias in exponential and power function fits due to noise: Comment on Myung, Kim and Pitt. *Memory and Cognition*, *31*, 656-661.

Brown, S. & Heathcote, A. (2003b). Averaging learning curves across and within participants. *Behaviour Research Methods, Instruments & Computers*, *35*, 11-21.

Dorogovtsev, S. N. & Mendes, J.F.F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London B*, *268*, 2603 – 2606.

Elman, J.L. & McClelland, J.L. (1986). Exploiting lawful in variability in the speech waveform. In J.S. Perkell & D.H. Klatt (Eds.), *Invariance and variability in speech processing*, Hillsdale, NJ: Erlbaum, 360–385.

Faloutsos, M., Faloutsos, P, & Faloutsos, C. (1999). On power-law relationships of the Internet topology. *Computing and Communications Review*, *29*, 251–262.

Ferrer, R. & Solé, R.V. (2001). The small world of human language. *Proceedings of the Royal Society of London B*, *268*, 2261–2265.

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., Luce, P.A., Pisoni, D.B., & Marcario, J.K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1211–1238.

Greenberg, J.H. & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157–177.

Huberman, B.A. & Adamic, L.A. (1999). Growth dynamics of the World-Wide Web. *Nature*, *401*, 131.

Huberman, B.A., Pirollo, P.L.T., Pitkow, J.E. & Lukose, R.M. (1998). Strong regularities in World-Wide Web surfing. *Science*, *280*, 95–97.

Jeong, H., Tombor, B., Albert, R., Oltavi, Z.N. & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651-654.

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

Kucera, F. & Francis, W. (1967). *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Landauer, T.K. & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*, 119–131.

Lawrence, S. & Giles, C.L. (1998). Searching the World-Wide Web. *Science*, *280*, 98–100.

Lawrence, S. & Giles, C.L. (1999). Accessibility of information on the web. *Nature*, *400*, 107-109.

Liljeros, F., Edling, C.R., Amaral, L.A.N., Stanley, H.E., & Aberg, Y. (2001). *Nature*, *411*, 907.

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

Marslen-Wilson, W.D. (1987). Parallel processing in spoken word recognition. *Cognition*, *25*, 71–102.

Marslen-Wilson, W.D. (1989). Access and integration: Projecting sound onto meaning. In W.D. Marslen-Wilson (Ed.), *Lexical access and representation*. Cambridge, MA: Bradford, 3–24.

McClelland, J.L. & Elman, J.L (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Milgram, S. The small world problem. *Psychology Today*, *2*, 60–67.

Newell, A. & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*. *98*, 404–409.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189-234.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*, Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Storkel, H.L. (2004). Do children acquire neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*, 201-221.

Triesman, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review*, *85*, 525–570.

Triesman, M. (1978b). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, *17*, 37–59.

Vitevitch, M.S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 735–747.

Vitevitch, M.S. (2004). Phonological neighbors in a small world: What can graph theory tell us about word learning? Unpublished manuscript, Department of Psychology, University of Kansas, Lawrence, KS.

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.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440-442.

Webster's Seventh Collegiate Dictionary. (1967). Los Angeles: Library Reproduction Service.

Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Molecular & Biological Evolution*, *18*, 1694-1702.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Speaker-independent Factors Affecting the Perception of Foreign Accent in a Second Language[1]

Susannah V. Levi, Stephen J. Winters and David B. Pisoni

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

[1] Preparation of this manuscript was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We wish to thank Jennifer Karpicke and Chirstina Fonte for help with data collection.

Speaker-independent Factors Affecting the Perception of Foreign Accent in a Second Language

Abstract. Previous research on the perception of foreign accent has largely focused on speaker-dependent factors such as Age of Learning and Length of Residence which are specific to an individual speaker. Factors that are independent of particular speakers and their language learning history have also been shown to affect perception of second language speech. The present study examined two speaker-independent factors—lexical frequency and listening context—that affect the perception of foreign-accented speech. Using a seven-point scale, two groups of listeners rated speakers on how much of a foreign accent they displayed. Listeners in the Auditory-Only listening context heard only the target stimuli, while listeners in the Auditory + Orthography listening context were presented with both the auditory signal and an orthographic display of the target word. The results revealed that lexical frequency affects the perception of the degree of foreign accent; higher frequency words were consistently rated as sounding less accented than lower frequency words. The effect of the listening context emerged in two interactions; the Auditory + Orthography context reduced the effects of lexical frequency but increased the perceived differences between native and nonnative speakers. The results suggest that structural and methodological factors independent of the speakers' actual speech articulations or developmental history affect the perception of degree of foreign accent and that such factors should be considered when interpreting the results of studies on the perception of foreign accented speech.

Introduction

The ability to speak a second language fluently depends in large part on how well a speaker has been able to acquire the second language (L2) phonology and to accurately realize the intended phonetic targets. The degree of foreign accent of a speaker, however, is not based exclusively on the amount of acoustic and articulatory mismatches between nonnative and native productions. Degree of foreign accent also reflects a *listener's* perception of the L2 speech. Many of the factors known to affect the perception of foreign-accented speech are speaker-specific factors that are inherent to a particular individual. We will refer to these factors as "speaker-dependent" since they are dependent upon a particular speaker's language learning history and cannot be directly changed or manipulated by an experimenter. Speaker-dependent factors have received considerable attention in the L2 literature. They include Age of Learning (the age at which a speaker begins learning a second language), Length of Residence in an L2 environment, the first language of the speaker, and his/her motivation to attain unaccented or less-accented speech (see Piske, MacKay, & Flege, 2001 for a review).

Additional factors which are not inherent to a particular speaker and are not part of the speaker's language learning history can also affect the perception of degree of foreign accent. These factors can be manipulated or controlled by the researcher and often reflect the specific methodology involved in obtaining measures of degree of foreign accent. We will refer to these as "speaker-independent" factors. For example, Southwood and Flege (1999) suggest that different rating scales may affect judgments of perceived degree of foreign accent. They point out that scales with fewer intervals may produce ceiling effects and therefore are not sensitive enough to differentiate L2 speakers.

Different types of elicitation techniques can also affect the degree of perceived foreign accent. Studies investigating the perception of foreign accent have used a variety of techniques to produce their stimulus materials; these techniques vary in whether the L2 speakers spontaneously generate speech, read printed text (words, sentences, or paragraphs), or repeat samples of speech after hearing the intended target produced by a native speaker. Oyama (1976) and Thompson (1991) have found that read speech is judged as more accented than spontaneous speech.

Studies also differ in whether native speaker controls are included. Native controls serve to confirm that listeners are correctly performing the task by testing that they can distinguish native from nonnative speech. Using native controls also ensures that listeners use a wider range of the rating scale. Characteristics of the listener can affect the perceived degree of foreign accent, as well. Several studies have varied whether naïve listeners (e.g., Flege & Fletcher, 1992; Flege, Munro, & MacKay, 1995) or experienced listeners such as linguists (e.g., Fathman, 1975) or ESL teachers (e.g., Piper & Cansin, 1988) serve as raters. Thompson (1991) found that naïve listeners tended to perceive a greater degree of foreign accent than experienced listeners, although Bongaerts, van Summeren, Planken, & Schils (1997) did not find a significant difference. Taken together, these studies show that speaker-independent factors can also affect the perceived degree of foreign accent.

The current study investigated the effects of two additional speaker-independent factors—lexical frequency and listening context—on the perception of degree of foreign accent using an accent rating task.[2] These two factors were chosen because they have been shown to affect speech perception and language processing of native speech. This study extends these two factors to the perception of foreign-accented speech.

Lexical frequency has been found to play an integral role in language processing and may therefore be expected to affect the perception of degree of foreign accent. Lexical frequency affects spoken word recognition (Howes, 1957; Savin, 1963; Luce & Pisoni, 1998), the recognition of words in a gating paradigm (Grosjean, 1980), and word shadowing (Goldinger, 1997). In a word identification task, Howes (1957) mixed words of varying frequency with multiple signal-to-noise ratios. High frequency words exhibited greater intelligibility by being perceived at less favorable signal-to-noise ratios than were the low frequency words. In a similar study, Savin (1963) examined listeners' response errors. Incorrect responses tended to be words of higher frequency than the target word. In a lexical decision task, Luce & Pisoni (1998) asked listeners to determine whether a target stimulus was a word or a nonword. They found that listeners responded more quickly and more accurately to high frequency words than to low frequency words.

Goldinger (1997) showed that listeners rely more heavily on the acoustic-phonetic information in the speech signal when they perceive low frequency words than when they perceive high frequency words. Using a word shadowing task, Goldinger presented listeners with both high and low frequency words and asked them to repeat the words as quickly as possible. The target words were spoken by several different talkers. Goldinger predicted that the subjects would change their productions to match the different speakers using "spontaneous vocal imitation." The amount of vocal imitation was quantified by comparing how well the response utterances matched the stimulus in fundamental frequency and duration. Goldinger found that low frequency words resulted in higher rates of spontaneous imitation

[2] We consider lexical frequency to be a speaker-independent factor because we consider it to be a property of a linguistic community. Though no two speakers have exactly the same frequency for all of their lexical items, we contend that globally, listeners with similar levels of education (in this case, students at Indiana University) will have similar lexical frequencies. In this experiment, we are examining the effects of frequency on the listener's perception of foreign accented speech and are therefore testing a homogeneous population.

than high frequency words, suggesting that subjects were more sensitive to the surface acoustic-phonetic details in the low-frequency words than the high-frequency words.

Goldinger explained these findings within the framework of Hintzman's (1986, 1988) MINERVA2 model, an exemplar-based model of memory (see also Johnson, 1997; Pierrehumbert, 2001, 2002; Kirchner, 1999, 2004). The MINERVA2 model, like other exemplar models, assumes that every exposure to a stimulus creates a memory trace that includes all perceptual details. When a new token (the probe) is heard, it activates an aggregate of all traces in memory, called the *echo*. This *echo* forms the listener's percept. The intensity of the echo depends upon both the similarity of the traces to the probe and the number of these traces. Thus, for speech and language processing, high frequency words induce "generic" echoes because they have many existing traces in memory and are therefore less influenced by any particular probe which enters the perceptual system. Low frequency words, on the other hand, have many fewer existing traces in memory. Any incoming probe will therefore have a greater influence on the subsequent percept. In Goldinger's word shadowing task, speakers based their repetitions more heavily on the incoming instance-specific information than on traces in memory for low frequency words. Their subsequent productions of low frequency words were therefore affected more by specific properties of the stimulus than high frequency words.

Working within the framework of exemplar models of speech perception, we hypothesized that the degree to which a speaker is perceived to have a foreign accent will be directly related to the amount of acoustic-phonetic mismatch between the signal and its resulting echo. In a nativeness rating task, we expected listeners' perception of L2 speech to rely more heavily on the acoustic-phonetic features of an incoming speech token for low frequency words. Listeners have fewer exemplars of low frequency words in memory and will thus generate less generic echoes in response to productions of those words. Potential acoustic-phonetic mismatches between productions of those words and their corresponding exemplars in memory should therefore be larger for low frequency words, which should in turn be rated as more accented than high frequency words.

The second speaker-independent factor investigated in this study was the listening context. Spoken words were either presented to participants in the auditory modality alone ("Auditory-Only") or with the addition of a simultaneous orthographic display ("Auditory + Orthography"). Knowledge of the intended target in the Auditory + Orthography context should facilitate the perception of degraded speech stimuli (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Davis et al. use the term "pop-out" to refer to a phenomenon where a degraded speech stimulus immediately becomes comprehensible after it is played to listeners in its original, undegraded form. Davis et al. tested the effects of pop-out on noise vocoded speech, a type of speech stimulus that simulates the signal heard by cochlear implant users.[3] In one experiment, they found that listeners were able to correctly report more words from a noise-vocoded target sentence after hearing the sentence in the clear. In another experiment, they found that listeners showed the same advantage, or "pop-out effect", after seeing the written version of a noise-vocoded sentence presented on a computer screen. This combination of effects demonstrates that top-down processing can influence the learning of severely degraded, noise vocoded speech regardless of the modality in which the original undegraded sentences are presented. As Davis et al. concluded, "pop-out must be at a non-acoustic, phonological level or higher" (pg. 230).

Presenting a word to a listener in orthographic form while he/she hears a nonnative production of the word may induce similar "pop-out" effects, since foreign accented speech can be regarded as a form of degraded speech. The effects of this type of pop-out on the perception of degree of foreign accent are

[3] This type of speech is created by filtering the original signal into six logarithmically spaced frequency bands.

unclear, however. One possibility is that simultaneously presenting the auditory and orthographic representations of the target word together will cause nonnative speech samples to be rated as less accented. If a nonnative production of the target word is ambiguous or difficult to understand, presenting the target word with orthography on the screen may promote a type of pop-out effect to occur where the "degraded", nonnative production immediately becomes more intelligible. Once the listener knows the intended utterance, possible ambiguities or confusions about which lexical item the listener should retrieve are lost. In this case, the perception of a high degree of foreign accent may also be significantly attenuated.

A second possibility is that simultaneously presenting auditory and orthographic representations of the target word will cause nonnative speech samples to be rated as more accented. This outcome might occur because knowledge of the target word may serve as a perceptual benchmark and therefore highlight the amount of mismatch between the target and its corresponding exemplars in memory. An actual example from our data serves to illustrate this point. Several of the L2 speakers in the current study consistently produced word final target /s/ as [z]. For these speakers, the target word 'noose' [nus] was produced as [nuz] (identical to 'news', which was not one of the target words). Hearing [nuz] while seeing 'noose' focuses listeners' attention to the mismatches between the expected and observed productions. It might be expected that listeners would rate these speakers as having more of a foreign accent when they hear the word [nuz] in conjunction with seeing 'noose' on the screen than when they simply hear [nuz] alone and could freely conclude that they had heard an accurate production of 'news'.
To summarize, the current study examined the effects of lexical frequency and listening context on the perceived degree of foreign accent of native and nonnative speakers of English. We predicted that higher frequency words would be rated as less accented than lower frequency words. In terms of the listening context, two competing hypotheses were assessed. The addition of orthographic displays may induce pop-out effects, making the stimuli more intelligible, resulting in their being rated as less accented. Alternatively, the presentation of the target word may cause listeners to focus their attention on mismatches between the target utterance and the actual stimuli, resulting in the stimuli being rated as more accented.

Methodology

Materials

Twelve female and ten male German L1/English L2 speakers were recorded in a sound-attenuated IAC booth in the Speech Research Laboratory at Indiana University. Speech samples were recorded using a SHURE SM98 head-mounted unidirectional (cardioid) condenser microphone with a flat frequency response from 40 to 20,000 Hz. Utterances were digitized into 16-bit stereo recordings via Tucker-Davis Technologies System II hardware at 22,050 Hz and saved directly to a PC. A single repetition of 360 English and 360 German words was produced by each speaker. Each word was of the form consonant-vowel-consonant (CVC) and was selected from the CELEX English and German databases (Baayen et al. 1995). Speakers read each word as it was presented to them on a computer monitor in the recording booth. Before each presentation, an asterisk appeared on the screen for 500 ms, signaling to the speaker that the next trial was about to begin. This was followed by a blank screen for 500 ms. After this delay, a recording period began which lasted for 2000 ms. The target word was presented on the screen for the first 1500 ms of this recording period. After the conclusion of the recording period, the screen went blank for 1500 ms, and then another asterisk appeared to signal the beginning of the next recording cycle. Presentation of the test items was blocked by language, but all within-language items were randomized. Items that were produced incorrectly or too loudly were noted and re-recorded in the same manner following each recording block. The total recording time for each

language block was approximately one hour for each speaker. Speakers were given the option of recording both sets of language items on either the same day or on two separate days, but all speakers elected to record all stimuli in a single recording session.

This process yielded recordings which were uniformly 2000 ms long. Since the actual productions of the stimulus words were always shorter than 2000 ms, the silent portions in the recording before and after each production were manually removed using Praat sound editing software. All edited tokens were then normalized to have a uniform RMS amplitude of 66.4 dB. Only the English words which had been both edited and normalized in this way were presented to the listeners in this study.

Of the 22 speakers, nine speakers were eliminated due to dialect differences (Austrian German: N= 3, Southern German: N=2, Romanian-German: N=1), reported speech or hearing disorders (N=2), or for only completing part of the recordings (N=1). Recordings from the remaining seven female and six male speakers were used in this study. All speakers were paid $10/hr for their time.

Thirteen native speakers (six male, seven female) of American English were also recorded producing only the list of English words under the same conditions as the bilingual speakers. These speakers were from various dialect areas of American English (Midland: N=7, West: N=1, South: N=1, North: N=1, More than one dialect area: N=3) (See Labov, Ash, & Boberg, 2006 for descriptions of these dialect labels.) Productions from two of the female speakers were not included in the study due to problems these speakers had with completing the task accurately. Productions from the remaining six male and five female native speakers were included in the study. All of these speakers received partial course credit for their participation.

Words from both languages varied in frequency based on counts from the CELEX database. For the purposes of analysis, the English words were divided into three equal groups of varying frequency. The 120 lowest frequency words all had a CELEX frequency count of less than or equal to 96, while the 120 highest frequency words all had a frequency of greater than or equal to 586. The remaining 120 words thus all had frequency counts between 96 and 586. The frequency count of homophones (e.g., rite, write, right) was taken to be the frequency count of the most frequent homophone; this homophone was also the word that was presented orthographically to the speakers during the recording sessions.

Listeners

A total of 87 listeners participated in this experiment; Forty-two were assigned to the Auditory-Only context and 45 were assigned to the Auditory + Orthography context. Twenty-seven listeners were eliminated (polylingual/nonnative speakers of English: N=6, L2 German: N=8, machine malfunction: N=9, non-American English dialect: N=1, speech/hearing disorder: N=2, not completing: N=1), resulting in 30 listeners for each listening context. None of the remaining listeners had studied German, and only 6 reported having German acquaintances (Friend: N=3, Teaching Assistant: N=2, Professor: N=1). Each listener participated in only one of the two listening contexts. All listeners received partial course credit for their participation.

Procedure

The experiment was implemented on Macintosh G3 computers running a customized SuperCard (version 4.1.1) stack. Listeners sat in front of these computers in a quiet testing room while wearing Beyerdynamic DT-100 headphones. The SuperCard stack played productions of individual words to listeners and then presented them with the on-screen question, “How much of a foreign accent did that

speaker have?" Participants answered this question by clicking the appropriate button in a seven-point rating scale ranging from 0 (="no foreign accent—native speaker of English") to 6 ("most foreign accent") presented on-screen. All listeners were informed that some of the speakers they would hear were native speakers of English and some were nonnative speakers. All listener ratings were converted to normalized z-scores per listener prior to completing any statistical analyses.

The auditory tokens of each word were presented to listeners in one of two different ways. Listeners in the Auditory-Only context heard each word prior to making a judgment of how accented the spoken stimulus was. Listeners in the Auditory + Orthography context, however, saw the orthographic representation of each word on the computer screen for 500 ms before hearing an auditory production of that word. The orthographic representation of the word remained on screen until the conclusion of the auditory stimulus, after which the listener rated its accentedness.

The experiment was divided into two blocks. In each block, 12 words were randomly selected for presentation from each of the eleven monolingual and thirteen bilingual speakers, yielding a total of 288 tokens per block. Listeners thus heard a total of 576 words over the duration of the entire experiment. Each block of words was rated by two different listeners.

The experiment was self-paced and listeners had the option of listening to the target words again before making their responses. After rating each token, participants clicked an on-screen button to play the next token. The entire study took approximately one hour for most listeners to complete. Participants in the Auditory + Orthography context listened to 89.5% of the tokens only once and to 10.5% more than once prior to making their responses. Participants in the Auditory-Only context listened to 78.5% of the tokens only once and to 21.5% of the tokens two or more times. A repeated-measures ANOVA with listening context (Auditory-Only vs. Auditory + Orthography) as a between-subjects factor and with native language of the speaker (L1-English vs. L2 English) as a within-subjects factor revealed main effects of both listening context ($F(1,58)=5.688$, $p=.020$) and native language ($F(1,58)=5.617$, $p=.021$), but no interaction. Listeners listened to stimuli more often in the Auditory-Only context than in the Auditory + Orthography context (means: 21.5% vs. 10.5%, respectively). Furthermore, listeners listened more often to the native speakers of English than to the nonnative speakers of English (16.6% vs. 15.3%).

Results

A repeated measures ANOVA with lexical frequency (low, medium, or high) and native language of the speaker (native vs. nonnative) as within-subjects variables and listening context (Auditory-Only or Auditory + Orthography) as a between-subjects variable was conducted on the z-scores of the nativeness ratings for all listeners. In the presentation of the results, larger z-score ratings indicate a greater degree of foreign accent.

The repeated measures ANOVA revealed a significant main effect of lexical frequency ($F(2, 116) = 44.8$, $p < .001$). Paired-samples t-tests revealed significant pair-wise differences between low vs. medium frequency, between medium vs. high frequency, and between low vs. high frequency (all $p \leq .002$). The direction of this effect indicated that lower frequency words were rated as more accented than higher frequency words. The mean z-scores for the ratings for each frequency group are presented in Table 1. A main effect of native language of the speaker was also found ($F(1, 58) = 1214.8$, $p < .001$). Native speakers were rated as having less foreign accent overall than nonnative speakers (see Table 2). The main effect for listening context was not significant ($F(1, 58) = 2.29$, $p = .135$).

Frequency	Mean (SD)
Low	.048 (.093)
Medium	-.049 (.054)
High	-.090 (.071)

Table 1. Overall means and standard deviations of the z-scores of the ratings by three levels of frequency.

Speaker	Mean (SD)
Native English	-.39 (.14)
Nonnative English	.33 (.12)

Table 2. Overall means and standard deviations of the z-scores for native vs. nonnative speakers.

The analysis also revealed significant interactions between lexical frequency and native language of the speaker ($F(2, 116) = 6.51$, $p = .002$), lexical frequency and listening context ($F(2, 116) = 13.81$, $p < .001$), and native language of the speaker and listening context ($F(1, 58) = 8.76$, $p = .004$). The three-way interaction was not significant.

The interaction between lexical frequency and native language of the speaker is shown in Figure 1. Paired-samples t-tests revealed that this interaction was due to a different pattern of ratings for the medium and high frequency words between the two groups of speakers. For native speakers of English, low frequency words were rated as having more of a foreign accent than medium frequency words, which were in turn rated as more accented than high frequency words (all $p \leq .001$). In contrast, for the nonnative speakers, low frequency words were rated as more accented than both medium and high frequency words ($p < .001$), but there was no significant difference between the medium and high frequency words (p=.213).

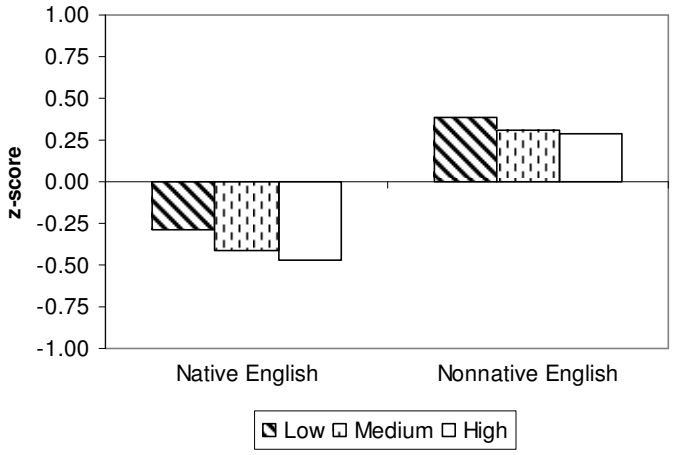


Figure 1. Mean z-score ratings for native and nonnative speakers for each of the three frequency conditions.

Figure 2 shows the interaction between lexical frequency and listening context. Paired samples t-tests revealed that this interaction was also the result of a different pattern of ratings for the medium and high frequency words. In the Auditory-Only context, low frequency words were rated as more accented than medium frequency words, which were in turn rated as more accented than high frequency words (all $p < .001$). In the Auditory + Orthography context, however, only low frequency words were rated as more accented than medium and high frequency words (both $p \leq .005$), while no significant difference was observed between medium and high frequency words ($p = .855$).

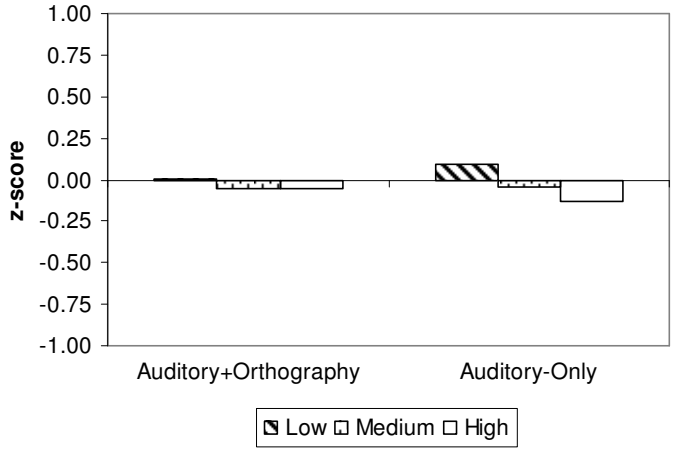


Figure 2. Mean z-score ratings for Auditory + Orthography and Auditory-Only contexts for each of the three levels of frequency.

Post-hoc t-tests on the native language of the speaker by listening context interaction revealed significant differences between the two listening contexts for both speaker groups. The cross-over interaction is illustrated in Figure 3. Native speakers were rated as less accented in the Auditory + Orthography context than in the Auditory-Only context (p= .004), whereas nonnative speakers were rated as less accented in the Auditory-Only context than in the Auditory + Orthography context ($p = .004$).

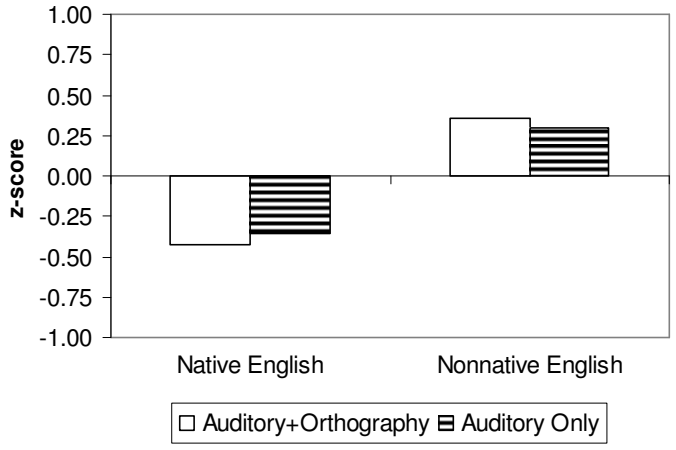


Figure 3. Mean z-score ratings for native and nonnative speakers for each of the two listening contexts.

Discussion

The results of this study demonstrate that two speaker-independent factors, lexical frequency and listening context, affect the perception of foreign accent in spoken words. High frequency words were rated as less accented than low frequency words. This result replicates earlier findings reported by Goldinger (1997) and is consistent with predictions based on exemplar models of speech perception. The more frequently a word occurs in the language, the more often a listener will hear it being spoken, which will in turn lead to encoding more exemplars of the word in memory. Highly variable, unusual, nonnative, "accented" productions of a target word will therefore be more likely to match (or approximate) an exemplar of a high frequency word in memory and therefore sound comparatively less accented to a native listener of English. Low frequency words, on the other hand, will be experienced less often and therefore have many fewer exemplars in memory. Thus, a nonnative production of a word must be a closer acoustic match to the few exemplars in memory in order to be rated as a good exemplar of that word type.

The effect of lexical frequency also entered into an interaction with the native language of the speaker. Lexical frequency had a stepwise effect on accent ratings for natively-produced tokens: high frequency words were rated as less accented than medium frequency words, which were in turn rated as less accented than low frequency words. The effect of frequency was attenuated for the nonnative speech, however. For nonnative tokens, only the low frequency words had significantly higher accent ratings than the medium and high frequency words, which did not significantly differ from one another.

The attenuation of the lexical frequency effect for the nonnative tokens may have been caused by the relationship between incoming acoustic stimuli and their stored exemplars. If degree of perceived foreign accent is dependent upon the number of exemplars in memory that are acoustically similar to the input signals, then stimulus tokens that are acoustically similar to many exemplars in memory will be rated as less accented than those which are acoustically similar to only a few exemplars, as was observed for native tokens. Nonnative tokens, however, are likely to have fewer acoustically similar stored exemplars than native tokens, especially for naïve listeners who have little if any experience with nonnative speech. Because the nonnative tokens lie in sparsely populated areas of the exemplar space in memory, the differences between high and medium frequency words may be eliminated. The frequency effect may remain for native speech because native productions of high and medium frequency words are in densely populated portions of the acoustic space where differences in frequency of the exemplars are likely to be apparent. The reason why the frequency effect remains for the low frequency nonnative productions may be the result of a different processing strategy for low frequency words. Very low frequency words may be processed as nonwords for some of the listeners and therefore receive significantly lower ratings. The lack of the expected frequency effects may also be due to the way the three levels of frequency were created. No a priori notion of high, medium, or low frequency was assumed. Instead, the 360 lexical items were simply ranked by lexical frequency and then divided into three equal groups. The resulting frequency groups were therefore continuous in the level of frequency. Since the difference between high and medium frequency was arbitrary and adjacent, the differences between the two highest levels of frequency may have been too small.

Although the main effect of listening context did not reach significance, it did have an effect on perceived degree of foreign accent through interactions with both lexical frequency and the native language of the speaker. The interaction between listening context and lexical frequency demonstrated that presenting a visual display of the target word on the screen attenuated the effect of lexical frequency. In the Auditory-Only listening context, the perceived degree of foreign accent was significantly different for words of all three levels of frequency. In the Auditory + Orthography context, however, accent

ratings for the high and medium frequency words were not significantly different from one another. The difference between the two listening contexts with respect to frequency is most likely the result of different processing requirements in the two contexts. In the Auditory-Only context, listeners must perform both a word recognition task and a nativeness rating task after hearing a stimulus. Listeners must evaluate the stimulus and compare it with stored exemplars in memory. In the Auditory + Orthography context, the process of auditory word recognition and lexical access are bypassed because the correct word is displayed visually on the computer screen. The attenuation of frequency effects on the perceived degree of foreign accent in the Auditory + Orthography context is consistent with numerous studies showing that effects of lexical frequency which are observed in open-set word recognition tasks disappear in analogous closed-set word recognition tasks (Pollack, Rubenstein, & Decker, 1959; Sommers, Kirk, & Pisoni, 1997; Clopper, Pisoni, & Tierney, in press). Since the Auditory + Orthography listening context eliminates the process of auditory word recognition from influencing ratings of accentedness, the perceived accentedness of a target word in this listening context must be based solely on its acoustic-phonetic properties, rather than on how familiar or unfamiliar the listener may be with the lexical item itself. In other words, in the Auditory + Orthography context, the nativeness ratings are based exclusively on acoustic-phonetic or phonological differences between the stimulus and existing exemplars and not on knowledge of the lexical properties of the items.

Nonetheless, low frequency words were consistently rated as more accented than both medium and high frequency words in both the Auditory + Orthography context and the Auditory-Only context. Because this effect was observed across both listening contexts, low frequency words may be processed in a fundamentally different way than high and medium frequency words, perhaps because listeners remain unfamiliar with them even after they have been informed of the identity of the word. It is also possible that the higher accent ratings for the low frequency words may reflect differences in the productions of these words. In a study that manipulated lexical frequency and neighborhood density, Wright (2003) found that speakers differed in the degree of vowel reduction/centralization as a result of these two lexical factors. In particular, he found that vowels in lexically "easy" words (i.e., high frequency words from sparse lexical neighborhoods) exhibited greater centralization than lexically "hard" words (i.e., low frequency words from dense lexical neighborhoods). Similarly in our data, low frequency words may exhibit less fluency and may include more hyper-articulated segments, causing them to be consistently perceived as less natural and therefore more accented.

Presentation context also influenced the degree of perceived accentedness by interacting with the native language of the speaker. Native speakers were rated as less accented in the Auditory + Orthography context than in the Auditory-Only context. The pattern of results was reversed, however, for nonnative speakers who were rated as more accented in the Auditory + Orthography context than in the Auditory-Only context. This crossover interaction may reflect differences in the relevant task demands placed on the listener. The Auditory + Orthography context allows listeners to bypass word recognition because the orthographic presentation serves to limit the possible "response alternatives". The Auditory + Orthography context, then, requires listeners to only judge the accentedness of a stimulus based on acoustic-phonetic similarity with existing exemplars of a particular word type. In their classic study of speech intelligibility, Miller, Heise, and Lichten (1951) showed that fewer response alternatives in a word-recognition task leads to higher levels of speech intelligibility at the same signal-to-noise ratio. In one experiment, they found that threshold (50 % correct) was reached at -14 dB SNR in a two-word vocabulary task but that a -4 dB SNR was needed for a 256-word vocabulary task.

These findings illustrate that more noise may be added to stimuli when there are fewer response alternatives while maintaining the same amount of intelligibility. Miller et al. argued that speech intelligibility is not determined by the stimulus item alone, but also by its context. Likewise, in the

present study, the intelligibility of a particular stimulus is increased in the Auditory + Orthography context because there is essentially only a single response alternative. The availability of context may account for why the native speakers are judged as less accented in the Auditory + Orthography context than in the Auditory-Only context. In other words, the reduction of response alternatives increases the intelligibility of the individual stimuli.

This explanation does not, however, account for the ratings of the nonnative speakers in the two listening contexts. The nonnative speakers were instead rated as more accented in the Auditory + Orthography context than in the Auditory-Only context. Since the process of word recognition is bypassed in the Auditory + Orthography listening context, accent ratings will be based solely on the acoustic-phonetic or phonological mismatch between a stimulus and stored exemplars. Presenting the target word to listeners orthographically in this context may highlight how poorly a nonnative production of that word matches its stored exemplars. Hence, nonnative productions of words may sound more accented when listeners are informed of the word's identity. In some cases, the auditory percept may even conflict with the orthographic target (e.g., [nuz] with "noose") and therefore result in a significantly higher rating of perceived foreign accent than if the auditory stimulus were presented without its orthographic representation. Data from the number of times listeners chose to repeat stimuli provide converging evidence that the context modulates a listener's judgment. Listeners in the Auditory-Only context listened to stimuli more often than in the Auditory + Orthography context.

The observed interaction of listening context and native language of speaker in this study has an important implication for future nativeness rating studies. Presenting words to listeners in an Auditory + Orthography context makes nonnative speakers sound *more accented* while making native speakers sound *less accented* than in the Auditory-Only context. The Auditory + Orthography listening context makes the accent ratings for the two groups of speakers diverge in the appropriate directions; native speakers are rated as less accented and nonnative speakers as more accented.

Conclusion

The results of the present study demonstrate that two speaker-independent factors—lexical frequency and listening context—affect the perception of degree of foreign accent in isolated spoken words. Listeners consistently perceived high frequency words as less accented than low frequency words. Simultaneously presenting a target word to listeners both auditorily and orthographically attenuated the effect of frequency, however. Furthermore, the addition of orthographic information in the Auditory + Orthography context caused native speakers of English to be rated as less accented and nonnative speakers of English to be rated as more accented than in the Auditory-Only context.

These findings have several implications for future research on accent perception. First, these results demonstrate that researchers need to consider the role that lexical frequency plays in studies that measure degree of foreign accent. If the effects of frequency are to be avoided, an orthographic representation of the target word can be used to attenuate these effects. Second, presenting target words to listeners both auditorily and orthographically yields different measures of perceived degree of foreign accent; in the Auditory + Orthography context, native speakers were rated as less accented while nonnative speakers were rated as more accented. The Auditory + Orthography context thus mitigates the effects of lexical frequency on accent ratings and also helps listeners better distinguish speech samples from native and nonnative speakers.

The results of this study also have several theoretical implications. Our findings show that an "accent" is not just a feature of a speaker's voice or how well a speaker is able to phonetically

approximate native speech, but also depends on the process by which that voice is perceived. We have shown here that this perceptual process is partially dependent upon non-acoustic properties of the signal such as lexical frequency and listening context. Previous work has shown that the intelligibility of L2 speakers reflects not only the actual acoustic accuracy of nonnative productions but also the prior experience and history of the listener. For example, Bent and Bradlow (2003) found that the intelligibility of several groups of nonnative speakers depended on the language background of the listeners. Nonnative listeners in both a matched and mismatched native language background performed equally well in a sentence intelligibility task with proficient nonnative talkers and with native talkers. Native listeners, on the other hand, found all the nonnative talkers to be less intelligible than native talkers. The process of accent perception is therefore shaped and modified to a large extent by a listener's past experiences and developmental history. A speaker may therefore only have an "accent" within a specific perceptual framework and listening context. The perception of a foreign accent thus reflects not only properties of the talker, but also prior experience of the listener and factors that affect the attunement between speaker and listener.

The influences of two speaker-independent factors—lexical frequency and listening context—on the perception of foreign accent in this study can be accounted for in large part by casting the process of accent perception more broadly within the framework of exemplar models of speech perception and spoken word recognition. We have assumed that the perception of foreign accent reflects the degree to which there is an acoustic-phonetic mismatch between a stimulus token and the stored exemplars in the listener's memory. The validity and robustness of this theoretical framework can be tested in future research. One possible way to test this framework is to manipulate the amount of experience that listeners have in listening to L2 speech. It has been noted above that highly experienced listeners (e.g. linguists and ESL teachers) sometimes rate the accents of nonnative speakers more leniently (i.e. as less accented) than naïve listeners do. This result may occur because experienced listeners have more exposure of L2 speech and therefore more L2 speech exemplars in memory than naïve listeners. Therefore, experienced listeners are more likely to find an acoustic match to incoming exemplars of L2 speech and rate these tokens as less accented than naïve listeners do. Increasing the amount of experience that naïve listeners have with L2 speech by presenting them with many tokens of L2 speech should therefore make them more tolerant raters of foreign accent in speech produced by unfamiliar L2 talkers. Experimental studies such as these should help increase our understanding of the process by which foreign accents are perceived and also provide us with a more complete picture of what it means for a speaker to "have a foreign accent".

References

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995) The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Bent, T. & Bradlow, A.R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America, 114,* 1600-1610.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language, *Studies in Second Language Acquisition, 19,* 447-465.

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

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K. & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, *134*, 222-241.

Fathman, A. (1975). The relationship between age and second language productive ability, *Language Learning*, *25*, 245-253.

Flege, J.E. & Fletcher, K.L. (1992). Talker and listener effects on degree of perceived foreign accent, *Journal of the Acoustical Society of America*, *91*, 370-389.

Flege, J.E., Munro, M.J., & MacKay, I.R.A. (1995) Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*. 3125-3134.

Goldinger, S. (1997) Words and voices: perception and production in an episodic lexicon. In K. Johnson & J. Mullennix (eds.), *Talker variability in speech processing*, pp. 33-66 (San Diego: Academic Press.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, *28*. 267-283.

Hintzman, D. (1986) Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93,* 411-428.

Hintzman, D. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528-551.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, *29*, 296-305.

Johnson, K. (1997) Speech perception without speaker normalization: an exemplar model In K. Johnson & J. Mullennix (eds.), *Talker variability in speech processing*, pp. 145-165. San Diego: Academic Press.

Kirchner, R. (1999). Preliminary thoughts on "phonologisation" within an exemplar-based speech processing system. *UCLA Working Papers in Linguistics, 1 (Papers in Phonology 2)*, 207-231.

Kirchner, R. (2004) Exemplar-based phonology and the time problem: a new representational technique. Poster presented at LabPhon 9 Conference, June 28th, 2004.

Labov, W., Ash, S. & Boberg, C. (2006). *The Atlas of North American English: Phonetics, phonology and sound change.* Berlin: Mouton de Gruyter.

Luce, P.A. & Pisoni, D.B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear & Hearing*, *19*, 1-36.

Miller, G.A., Heise, G.A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*, 329-335.

Oyama, S. (1976) A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*, 261-283.

Pierrehumbert, J. (2001) Exemplar dynamics: word frequency, lenition, and contrast. In J. Bybee & P. Hopper (eds.), *Frequency effects and the emergence of linguistic structure*, pp. 137-157. Amsterdam: John Benjamins.

Pierrehumbert, J. (2002) Word-specific phonetics. In C. Gussenhoven & N. Warner, (eds.), *Laboratory Phonology VII,* pp. 101-140. Berlin: Mouton de Gruyter.

Piper, T. & Cansin, D. (1988). Factors influencing the foreign accent, *The Canadian Modern Language Review*, *44*, 334-342.

Piske, T., MacKay, I.R.A., & Flege, J.E. (2001) Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics, 29*. 191-215.

Pollack, I., Rubenstein, H., & Decker, L. (1959) Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31*. 273-279.

Savin, H.B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, *35*, 200-206.

Sommers, M.S., Kirk, K.I., Pisoni, D.B. (1997) Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear & Hearing, 18*. 89-99.

Southwood, M.H. and Flege, J.E. (1999) Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics, 13*. 335-349.

Thompson, I. (1991) Foreign accents revisited: the English pronunciation of Russian immigrants, *Language Learning*, *41*, 177-204.

Wright, R. (2003). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (eds.), *Papers in Laboratory Phonology, VI: Phonetic Interpretation*, pp. 75-87. Cambridge, UK: Cambridge University Press.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Indexical and Linguistic Channels in Speech Perception: Some Effects of Voiceovers on Advertising Outcomes [1]

Susannah V. Levi and David B. Pisoni

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

[1] Preparation of this chapter was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We wish to thank Luis Hernandez and Darla Sallee for technical assistance and help with this manuscript.

Indexical and Linguistic Channels in Speech Perception: Some Effects of Voiceovers on Advertising Outcomes

Abstract. This article examines the effects that voice features have on advertising. Previous research in neurolinguistics and psycholinguistics shows that linguistic and extralinguistic ("indexical") properties of speech are closely coupled in speech perception and spoken language processing. We review research from the advertising and marketing literature that examines which voices are the most suitable for voiceovers, whether speech rate compression is advisable, and under what circumstances voice selection is most important. We integrate these two bodies of literature and conclude that the voices used in advertising should be familiar and consistent across the campaign and the speaking rate may be increased without deleterious effects.

Components of Speech

Marshall McLuhan wrote "the medium is the message." That is, not only is the content of the message itself important in conveying information, but so too is the medium, or the way in which the intended message is conveyed to an audience. When people perceive spoken language, information about the content of the message is transmitted to the listener, along with information about the specific person who produced the message. Because these two sources of information are ineluctably bound together in the speech stream, both channels of information contribute to the final product of perception and both should be considered by advertisers when developing voiceovers.

Speech is a complex, multimodal time-varying pattern. Although both auditory and visual cues function in speech perception, we will focus only on the auditory portion. Spoken language encodes two different sources of information. First, it carries linguistic information about the symbolic content of the talker's intended message. This content contains several levels of linguistic information: phonological (sounds), morphological (units which form words), syntactic (combining words into sentences), and semantic (meaning of an utterance). Taken together, this linguistic information provides the content of an utterance.

The second type of information that is carried in the speech stream is often termed paralinguistic, extralinguistic, or indexical. Indexical information can be thought of as the "medium" through which the message is conveyed. Abercrombie (1967) wrote that "[s]uch 'extra-linguistic' properties of the medium… may fulfill other functions which may sometimes even be more important than linguistic communication, and which can never be completely ignored" (p. 5). Abercrombie divided the indexical properties of speech into three sets: (1) those properties that indicate group membership (e.g., regional, dialectal, and social aspects of speech), (2) those that characterize the individual (e.g., age, gender, and size and shape of the vocal tract), and (3) those that reveal changing states of the speaker (e.g., affective properties such as fatigue, excitement, amusement, anger, suspicion, health, speaking rate). Indexical and linguistic information in speech correspond to what cognitive psychologists often refer to as source and item information, respectively (see Hilford, Glanzer, Kim, & DeCarlo, 2002).

What makes speech a complex signal is that these two properties are carried simultaneously in a single acoustic waveform that is at first produced by an individual speaker and then perceived by a listener who can extract both sources of information. Speech is generated by a speaker's larynx and supralaryngeal vocal tract. The vocal tract which extends from the larynx through the throat and mouth to

the lips acts as an acoustic filter, enhancing certain resonance frequencies (formants) and attenuating others. When speakers produce different sounds in a language, they constrict their vocal tract at different locations. Which frequencies are enhanced or attenuated in the vocal tract is determined both by its length and by the location of the constriction. In the productions of sounds, the relative frequencies provide the linguistic information about the place of constriction of sounds. In contrast, the absolute frequencies that resonate in a particular person's vocal tract are dependent on the length of that person's vocal tract and thus provide talker-specific information. The sound spectrogram in Figure 1 provides a specific example of the integration of linguistic and indexical properties of speech in the production of speech. The formant values produced by the female speaker (first author) are higher than those produced by the male speaker (second author), showing one indexical difference resulting from differences in vocal tract length. The overall movement and relative locations of the formants, on the other hand, provide linguistic information and indicate that the speakers are saying the same utterance. Thus, the same vocal mechanisms produce both linguistic and indexical information simultaneously and both sources of information are encoded and carried in the same signal.

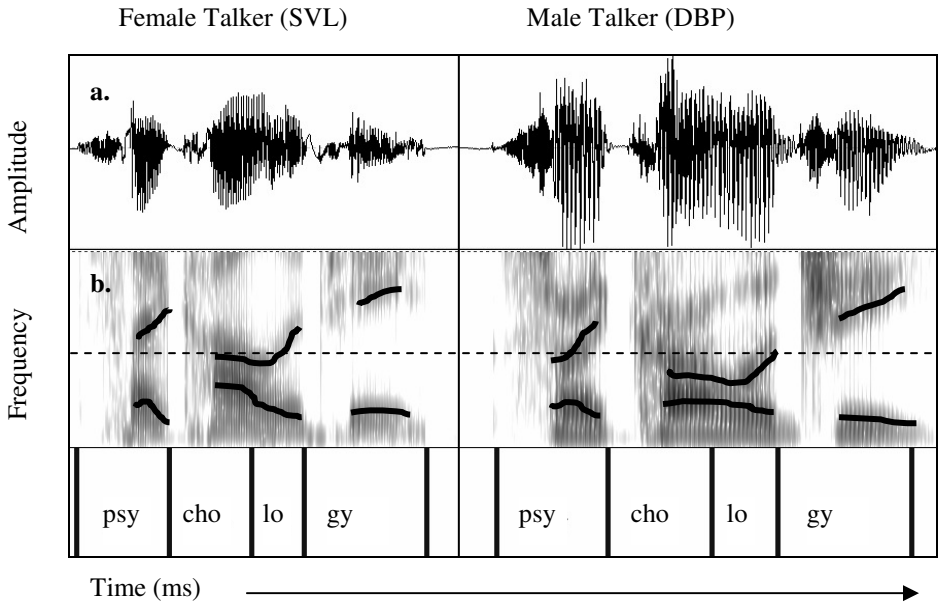


Figure 1. Waveform (a) and spectrogram (b) of the word "psychology" produced by the first author (SVL) and the second author (DBP). Dark lines in the spectrogram represent the first formant (lower curve) and second formant (upper curve).

The perception of these two different aspects of speech is illustrated in Figure 2. The basilar membrane (bottom of Figure 2) is situated in the cochlea in the inner ear and allows a listener to segregate frequencies. The left path in Figure 1 shows the absolute frequencies that are heard by the listener and provide indexical information about an individual talker. The right path represents the relative frequencies which provide linguistic information about the intended message. In this way, both the linguistic and the indexical properties of the speech signal can be perceived and encoded by the listener.

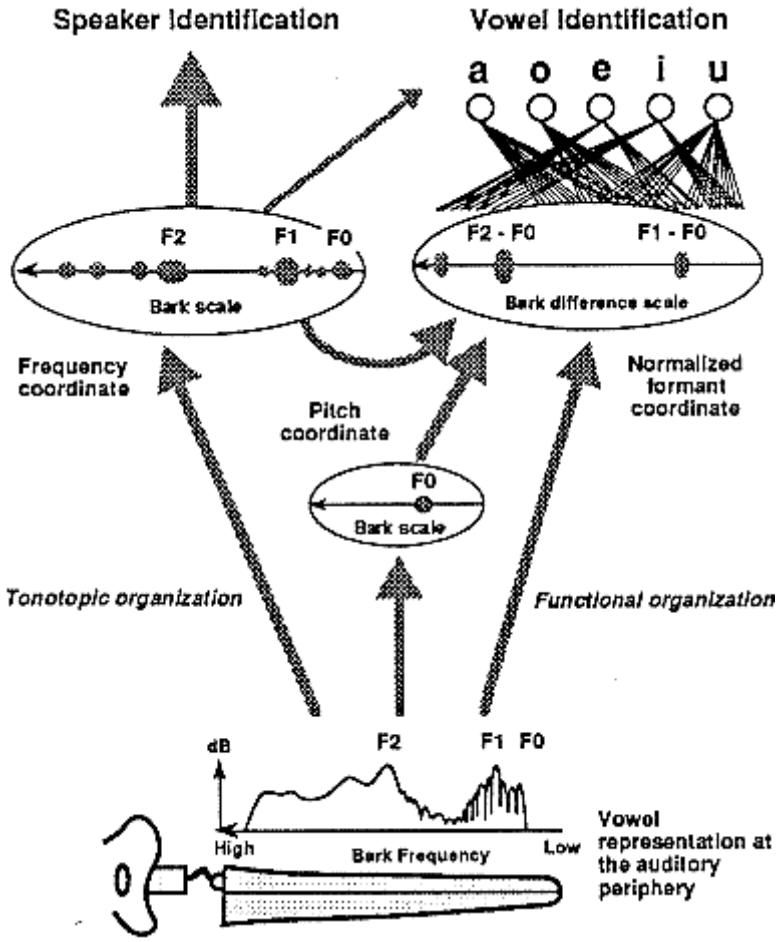


Figure 2. Representation of auditory perception of both indexical and linguistic properties of speech. The absolute frequencies (left side) provide speaker identification, while the relative frequencies (right side) provide vowel identification (Hirahara & Kato, 1992).

The remainder of this chapter is organized into three sections. In the first, we review several lines of neurolinguistic and psycholinguistic research on the perception of indexical properties of speech. The findings discussed in this section confirm that two distinct channels of information are carried in the speech signal. Moreover, the results suggest that the processing of one set of properties affects the processing of the other. In the second section, we consider research from the advertising and marketing literature that examines which voices are the most suitable for voiceovers, whether speech rate compression is advisable, and in what contexts selecting the appropriate voice is most important. In the last section, we integrate these two separate bodies of literature in order to determine what kinds of voices should be used for the most effective advertising.

The Science of Voice Processing

Behavioral and neural studies on the perceptual processing of speech illustrate its bipartite nature. By asking listeners to attend to either the linguistic or the indexical (voice) properties of speech, neuroscientists have shown that these two aspects of speech are processed differently in the brain.

Despite this difference in neural processing, behavioral studies show that the two are in fact closely linked and that voice (indexical) characteristics affect linguistic processing of speech.

Neural Processing of Voices

Neural studies of voice identification and discrimination reveal that characteristics of the voice are processed in brain areas which are distinct from those that process the linguistic properties of the speech signal. In an early study of hemispheric specialization, Landis, Buttet, Assal, and Graves (1982) found that while both hemispheres can be utilized in voice recognition, there was a distinct advantage of the left hemisphere for linguistic tasks. Landis et al. played monosyllabic consonant-vowel words into either the right or the left ear. In the linguistic task, listeners were asked to press a button every time they heard a specific target word. Listeners' reaction times showed a clear right-ear advantage (REA), responding faster when the target word was presented to the right ear than the left. Because the two hemispheres control contralateral body functions, showing a preference for the right ear, indicates that the left hemisphere dominates in the linguistic task. In a second experiment, listeners were asked to push a button when they heard a particular male or female voice. In this study, female voices elicited a REA, but male voices a left-ear advantage (LEA). Landis et al. interpreted these results by remarking that higher frequencies have been shown to elicit a REA and that female voices, with their higher pitch and formants, may therefore also be processed with a REA. The major finding of this study was the demonstration that both hemispheres are involved in voice recognition, whereas word recognition displays left hemisphere dominance.

Kreiman and Van Lancker (1988) found similar results using a dichotic listening paradigm. In a dichotic listening task, listeners hear different words presented simultaneously in both ears and are asked to attend only to the stimuli that are played in either the right or the left ear. Using a set of 50 famous male voices, they asked listeners to write down both the word (linguistic task) and the person who said the word (indexical task). As expected, they found a clear REA in the word recognition task. The results of the voice identification task were less conclusive. Listeners showed no ear advantage for the voice recognition task, consistent with the earlier results of Landis et al. (1982). They did, however, find a *relative* left-ear advantage; that is, relative to the word recognition task, listeners showed a greater advantage for the left ear.

More recent studies have been able to isolate voice processing to more specific brain regions. Glisky, Polster, and Routhieaux (1995) tested elderly listeners' ability to recall either the content or the voice of previously heard sentences. They found that listeners with high frontal lobe function outperformed those with poor frontal lobe function on the voice task, but showed no difference in their performance on the sentence recall task. Conversely, listeners with high medial temporal lobe function outperformed listeners with low function in the sentence recall task, but did not differ on the voice task. These results confirm that the processing of voice information is independent of linguistic processing.

More recently, using functional magnetic resonance imaging (fMRI), Stevens (2004) reported distinct brain regions for voice- and word-discrimination tasks. Listeners were asked to determine whether two talkers were the same or whether two words were the same. Stevens found that attending to either the word or the voice altered the functional activity of the brain. In particular, the voice comparison task produced activation in the right fronto-parietal area, whereas lexical processing was associated with increased activation in the left frontal and bilateral parietal areas.

Other studies have shown that voice processing can be further subdivided; familiar voices are processed differently than unfamiliar voices. In these studies, familiarity refers to people who were

personally known to the listeners.[2] Using fMRI, Shah et al. (2001) found that familiarity of voices and faces resulted in increased activity in the posterior cingulate cortex as compared to unfamiliar voice and face processing. Nakamura et al. (2001) also found different brain areas involved in familiar versus unknown voice processing using positron emission tomography (PET).

Taken together, these studies of the neural processing of speech demonstrate that the indexical (source) properties are indeed distinct from the linguistic (symbolic) properties of speech, despite the fact that they are carried simultaneously in the same speech waveform. When listeners are asked to attend to voice characteristics of the speaker, they utilize different areas of the brain than when they process the linguistic information in the signal.

Interactions of Indexical (voice) and Linguistic Processing

Although the studies reviewed in the previous section revealed that distinct brain areas are involved in voice perception and linguistic processing, results of behavioral studies indicate that these two properties of the speech signal are closely coupled functionally. Properties of the voice affect the processing of linguistic information. Most important for the concerns of advertisers is the incidental or indirect effects of voice information on the processing of the content of the message which show that consistency and familiarity of the voice facilitates linguistic processing of the message.

Evidence from a variety of behavioral studies shows that consistency of the voice is an important aspect of linguistic processing. Using a speeded classification task, Mullennix and Pisoni (1990) asked listeners to categorize a set of spoken words that differed on two perceptual dimensions: the linguistic dimension in which the initial sound of the words varied between “p” and “b” and the indexical/gender dimension in which words were spoken by either a male or a female talker. In the control conditions, a single talker produced all words, thereby holding the indexical dimension constant. In the orthogonal conditions, the two dimensions varied randomly so there was no consistency between the two dimensions. Listeners were asked to classify words using each dimension separately, ignoring possible variation along the other dimension. Mullennix and Pisoni found that reaction times were slower in the orthogonal conditions than in the control conditions, indicating that listeners were not able to “filter out” the indexical variation while performing the linguistic task and that variation of the voice inhibits listeners’ performance. They also found that increasing the number of talkers from two to 16 had an even greater effect of slowing down classification times. This study revealed that the indexical properties of speech are not processed independently of linguistic content of the signal and that irrelevant variation in a non-attended perceptual dimension (in this case, the indexical dimension) is not discarded when performing such a task, but is instead processed in an integral manner.

Schacter and Church (1992) found a similar same-voice advantage in a stem completion task. In the study phase, listeners heard a series of words and rated either the pleasantness of the word or the pitch of the voice. In the test phase, listeners heard a series of syllables mixed with noise and were asked to write down the first word that came to mind. Schacter and Church found that when voices of the study words and the test syllables matched, a greater priming effect was observed than when the voices were switched. In other words, listeners were more likely to respond with a word they had heard during the

[2] In a several studies, Van Lancker, Kreiman, and colleagues (Van Lancker & Kreiman, 1987; Van Lancker, Cummings, Kreiman, & Dobkin, 1988; Van Lancker, Kreiman, & Cummings, 1989) showed that recognizing famous voices and discriminating between unfamiliar voices engaged different brain areas. It is not possible to conclude from these studies that famous and unfamiliar voices themselves are processed differently because the two tasks were fundamentally different. In the famous voice recognition task, listeners were asked to name the famous voice and to draw on long term memory. In the unknown voice discrimination task, listeners compared two unknown voices that were presented one following the other.

study phase of the experiment if that studied word was spoken in the same voice as the syllable heard during the test phase.

In another study, Goldinger (1996) showed that listeners exhibit a same-voice advantage in recognition memory when performing a linguistic task. Listeners were asked to type the word they heard when it was presented in noise. Test words that were spoken by the same talker were recognized more often than words spoken by a different talker. Perhaps even more striking was the finding that the same-voice advantage did not decline significantly across different delays between study and test. Listeners who returned after a week showed the same voice advantage as those who returned after only a five-minute delay, indicating that listeners encode and store information about a voice for an extended period of time, even when the demands of the task do not consciously ask listeners do so. The lack of an effect of delay suggests that the voice effect does not disappear rapidly but is available and stored in memory for an extended period of time. In a separate voice-recognition task, Goldinger (1996) found that listeners' ability to explicitly remember voices did decline with an increased delay. Together these two sets of results suggest that while listeners may lose their ability to explicitly remember the voice, attributes of a voice remain in memory and have effects on language processing for an extended period of time.

Using a list recall task, Goldinger, Pisoni, and Logan (1991) also found an advantage for voice consistency in learning and memory. In this study, listeners first heard 10 words and were subsequently asked to recall the list. Goldinger et al. varied the number of voices which were used to present the list of words and the rate at which the stimuli were presented. The authors found that at fast presentation rates, lists of words produced by multiple talkers were recalled less accurately than lists that were spoken by only a single talker. In contrast, at slow presentation rates, lists produced by multiple talkers were actually remembered more accurately than single-talker lists. Lightfoot (1989) conducted a follow-up study using this same methodology. The difference in Lightfoot's study was that lists were spoken by voices that were familiar to the listeners. Interestingly, voice familiarity caused the advantage of voice consistency to disappear.

In a continuous recognition memory experiment using spoken words, Palmeri, Goldinger, and Pisoni (1993) played long lists of words to listeners and asked them to determine whether each word was "old" (one that had been previously heard) or "new" (one that had not been previously heard). In order to assess the effects of voice on recognition memory, half of the old words were repeated in the same voice and half were repeated in a different voice. As in the previous studies, listeners responded more quickly and more accurately when old words were repeated in the same voice. Palmeri et al. also found that the lag (i.e. the number of words intervening between the first and second presentation of a word) did not interact with the same-voice advantage, indicating that the facilatory effect of maintaining the same voice is robust over time.

In addition to consistency of voices, familiarity with voices facilitates recall and recognition of spoken language. Several studies have shown that familiarity with a set of talkers allows for faster and more accurate linguistic processing. For example, Nygaard, Sommers, and Pisoni (1994) trained listeners to identify ten unfamiliar talkers by name over a period of ten days. During the test phase on the last day, listeners were presented with novel words mixed in noise that were spoken either by the now familiar talkers or by unknown talkers. Subjects were simply asked to identify words and were not required to respond to the voice of the talker. The results indicated that listeners identified novel words in noise better when the words were spoken by familiar talkers, than when the words were spoken by unfamiliar talkers. In a follow-up study, Nygaard and Pisoni (1998) showed that the advantage of talker familiarity extends to sentence-length utterances as well.

The behavioral studies reviewed in this section suggest that the linguistic and indexical channels of speech are closely coupled. In linguistic tasks (e.g. word recognition and phoneme discrimination) that on the surface do not appear to rely on indexical or voice properties, a strong effect of voice is reliably observed. Both familiarity with the voice and consistency of the voice facilitate processing of the linguistic (symbolic) content of the message.

Advertising/Marketing

Advertising messages using spoken language contain meaningful information (the intended message text), visual information (in the case of television advertising), and voice information. It has been shown that when both audio and visual information are present, the auditory information has attentional priority over the visual modality and can mask otherwise distracting information in the visual signal. Drew and Cadwell (1985) varied the angle and zoom of jump cuts in an informational video. They found that when an audio signal accompanied the video there were no negative effects on viewers' attitudes towards the video, showing the importance of an audio stream for maintaining coherence and sufficiently masking distracting visual cues. Since audio information is clearly relevant in both radio and television advertising, selecting an appropriate voice to accompany the product of the advertisement is important and may have significant effects on a wide range of outcome measures. In this section, we consider some factors that are relevant for selecting an appropriate voice for an advertising campaign. We will also discuss under what conditions voice characteristics are likely to affect listeners' attitude towards and memory for the product.

Picking the Right Voice

In selecting the right voice to accompany an ad, several considerations must be made. For instance, should the voice of a famous person be used? What gender voice is appropriate for a given product? Does the accent or nativeness of the talker's voice play a role in listeners' understanding, attitude, and memory for the product?

A first question that an advertiser might consider is whether the spokesperson for a product should be famous. It may be the case that famous actors are better able to read the script of an ad (Alsop, 1987). Not surprisingly, it is also important that if a celebrity is used in an ad, that he/she match the product in such a way that credibility of the product is enhanced (Plapler, 1974; Misra & Beatty, 1990).

While using a celebrity voice in advertising is more expensive, it may be the case that a celebrity is actually better at selling a product than an unknown person. Leung and Kee (1999) conducted an experiment to test whether celebrity spokespeople were better than unknown actors in selling a product. They took a recent television commercial which used two well-known DJs in Hong Kong as the voiceovers for the ad and recorded the same ad with two trained but not well-known actors. Viewers who saw the ad with the celebrity voiceover had higher brand recall and encoded more product brand information, although there was no significant difference in viewers' intent to buy the product.

Finding the right talker for a voiceover also includes deciding the appropriate gender of the speaker. While male voices dominate the world of voiceovers (Bartsch, Burnett, Diller, & Rankin-Williams, 2000), several studies indicate that female voices may be a better choice under some circumstances and that the gender of the voice interacts with the product. Whipple and McManamon (2002) tested listeners' attitudes toward male-gendered, female-gendered, and neutral-gendered products that differed in the voice of the spokesperson. Their results indicated that the gender of the spokesperson

does not have an effect on gender-neutral or male-gendered products. However, for female-gendered products, a female voice elicited a more positive attitude toward the ad. The only scenario where a male voice was preferred was for the female-gendered product when men were the target audience (e.g. for men purchasing the product as a gift). Thus, Whipple and McManamon conclude that female voices have at least the same effectiveness as male voices, if not more.

In examining the gender of spokes-characters (non-human animated characters), Peirce (2001) found that the likelihood that a viewer would buy the product was increased when the gender of the spokes-character matched that of the product (golf balls vs. vacuum cleaners, in the case of this study). Conversely, the gender-neutral spokes-character was not the most effective for the gender-neutral product; instead, the female spokes-character was preferred for the gender-neutral product (coffee). These studies demonstrate that there is little basis to continue to prefer male-gendered voices or spokes-people in advertising.

A third consideration in selecting the voice for an ad is the nativeness of the talker. Although there may be other considerations such as the intended audience or product congruity (e.g., using an Italian-accented voice for pasta), several studies have shown that foreign-accented voices are less intelligible than native voices. In a study examining the effects of voice on listeners' ability to comprehend and retain information from a short narrative, Mayer, Sobko, and Mautone (2003) found that listeners performed better on both a retention task and a transfer task when the speaker was a native speaker of English compared to when the speaker was a second-language learner of English with a Russian accent. They also found that the native speaker received higher positive ratings scores than the nonnative speaker. Foreign accented speech has also been found to be less intelligible when mixed with noise (Lane, 1963; Munro, 1998) and requires more effort to process (Munro & Derwing, 1995).

In the advertising literature, foreign-accented voices have also been shown to elicit less favorable responses and lower purchase intentions. Tsalikis, DeShields, and LaTour (1991) found that Greek-accented English voices received lower scores on 15 bipolar adjectives than native-English voices for a hypothetical commercial for a VCR. In a similar study, DeShields, Kara and Kaynak (1996) tested listeners' attitudes towards native English and Spanish-accented speech when presented with an ad for car insurance. They found that the intent to buy was significantly higher when the speaker was native than for the Spanish-accented speakers. DeShields and de los Santos (2000) found that the impact of accent depends on the relationship between the source of the accent and the listeners. In accordance with previous work, they found that US listeners perceived the native English speaker more positively than the Spanish-accented speaker in an ad for car insurance. Mexican listeners, however, did not rate the native Spanish salesperson differently than the English-accented Spanish-speaking salesperson. DeShields and de los Santos hypothesized that this may be due to the influence the US has on Mexican culture.

Time is Money

Speaking rate is another indexical property that can be manipulated and controlled by advertisers. Because advertising time is expensive, a reasonable question to ask is whether the fast presentation of information, which allows more information to be transmitted in a shorter period of time, has any deleterious effects on listeners' attitudes toward the message, their ability to remember the product, or their intention to buy. Unfortunately, the studies which have examined the effects of speech rate are not conclusive, although the majority suggests that a faster rate is not problematic.

In some cases, faster rates of speech have been shown to be preferred by listeners. Miller, Maruyama, Beaber, and Valone (1976) conducted two experiments in order to test the effects of speech

rate on listeners' attitudes toward the speaker. In the first experiment, groups of listeners heard a passage about the dangers of coffee at two different speaking rates. In addition to varying rate, they also varied the credibility of the speaker by telling listeners that the speaker was either a locksmith or a biochemist. In a second experiment, listeners heard a passage about hydroponically grown vegetables at two speaking rates and at two levels of message complexity. Listeners answered a series of questions designed to determine their attitude toward the speaker. The results showed that listeners judged the speaker of the faster rate to be more knowledgeable, more persuasive, more objective, and also to have greater intelligence. The effects of speech rate were robust; the faster rate elicited more positive responses in all conditions, regardless of the credibility of the speaker or the complexity of the message. One limitation of this study, however, was that the speaker was asked to vary his speech rate, thus it is very likely that other aspects of the voice were altered as well, such as pitch and amplitude.

LaBarbera and MacLachlan (1979), however, avoided these possible confounds by electronically compressing the speech rate. First, they conducted a series of experiments to test listeners' preference for different speech rates. They compressed and expanded the speech rate without altering the pitch of the voice and asked listeners in a paired-comparison task to select which speech sample they preferred. The results of these studies indicated that listeners preferred a faster than normal speaking rate. In a follow-up study, LaBarbera and MacLachlan tested listeners' attitudes and recall of six radio commercials at both normal and fast speech rates. They found that in all cases, the faster commercial was rated as more interesting and elicited higher brand recall after a two-hour delay. Thus, the faster rate was both preferred by listeners and also resulted in higher retention.

MacLachlan (1982) also reported positive effects of faster speech rates. Four radio commercials were used either in their normal or compressed versions, and listeners rated the speaker along four dimensions: friendliness, knowledge, enthusiasm, and energy. The fast commercials were either rated the same as the normal version or more positively. In this study, then, increasing the speech rate had no negative effects on listeners' attitudes about the speaker.

Other studies have shown mixed effects of altering the speech rate. Schlinger, Alwitt, McCarthy, and Green (1983) found that time compression can sometimes interfere with encoding the content in television commercials. Viewers in this study expressed fewer ideas about one of the two commercials in their study when it was presented at the faster rate, but no significant difference was found for the second commercial. As for listeners' attitudes, six of 52 response statements showed the non-compressed version as receiving more positive responses, although the remaining 46 statements showed no difference. Furthermore, the results showed no significant difference in buying intentions for the normal and time-compressed versions of the commercials. Thus, listeners may encode less information and may have fewer positive responses for some response statements, but this does not seem to affect the likelihood that they will actually purchase the product.

More recently, Megehee, Dobie, and Grant (2003) found mixed results for faster rates of speech. They created five versions of a message about the benefits of using a "SmartCard" (an identification card that also functions as a debit card): normal, time-compressed, pause-compressed, time-expanded, and pause-expanded. Thus, three rates (normal, fast, slow) and two methods of rate alteration were studied. Time-adjusted speech changes the overall rate of the message by compressing or expanding all portions equally, but the tempo remains the same. In pause-adjusted speech, the pauses themselves are either shortened or lengthened; thus the actual presentation rate of the words remains the same, but the tempo of the utterance is altered. When comparing the main effect of rate, Megehee et al. found no difference in the attitude toward the product, message, or speaker, though the faster rate did have more affective responses, while the slower rate had more cognitive responses. The authors also found that at faster rates,

time-compression produced more affective responses and a more favorable attitude towards the speaker than did the pause-compressed version.

Chattopadhyay, Dahl, Ritchie, and Shahin (2003) found different results for the method of rate adjustment. They varied both syllable speed (compression of the actual speech forms) and interphrase pause duration and found that reducing the interphrase pause time had little effect on the way listeners processed the message, suggesting that this might be the preferred method of time compression. Increasing syllable speed, on the other hand, did affect the way listeners processed the message, as revealed by measures of attention and recall. They found, however, that increasing syllable rate can increase persuasion, implying that this might be the preferred method of rate compression.

Although a few results from these studies show that increasing the speech rate has some negative consequences (e.g. fewer cognitive responses), the overwhelming conclusion is that faster rates are not problematic and are in some cases preferred. The best method of compression, however, is less obvious. Megehee et al. showed a clear advantage of overall time-compression, whereas Chattopadhyay et al. showed some superiority for pause-compression. Whatever the method of compression, increases in speech rate appear to be well-tolerated by listeners.

When Voice Characteristics Matter Most

The impact of voice characteristics varies depending on how much involvement and interest the listener has with the message. Gelinas-Chebat and Chebat (2001) conducted a study to examine the contribution of voice characteristics on listeners' attitudes toward an ad by varying the level of involvement. Listeners, who were all university students, heard either a low involvement ad which invited them to visit the local bank to acquire an ATM card or a high involvement ad which invited them to visit the local bank to learn about student loans. The assumption was that students would be more interested in learning about student loans since it could directly affect their financial situation. In addition to varying the level of involvement, voice characteristics were varied orthogonally along two dimensions (intensity and intonation) with two levels each, creating four versions of each message. As predicted, the high involvement message increased the acceptance of the arguments of the message. Additionally, changes in voice characteristics did not have an impact on listeners' attitudes in the high-involvement message. However, in the low-involvement message where listeners did not have an a priori interest in the message, the peripheral characteristics of the message (i.e. the changes in voice characteristics) did have an effect on their attitude toward the message. In other words, when listeners do not have a particular interest in the product or message, the quality of the voice that is used has an effect on listeners' attitudes.

Further support for effects of voice on processing and memory comes from another study by Goldinger (1996). He varied the level of processing (LOP) in order to determine whether the focus of listeners' attention would interact with changes in a speaker's voice. In the study phase, listeners encoded 150 words in terms of the gender of the speaker (shallowest LOP), their initial sound, or their syntactic class, namely noun, verb, or adjective (deepest LOP). In the test phase, listeners were given a set of 300 words and were asked to classify words as old or new depending on whether they had been heard in the initial part of the study. Half of the old words were repeated in the same voice and half in a different voice. The strongest effect of voice change was found at the shallowest LOP where words repeated in the same voice received more accurate responses. This result suggests that when listeners' attention is directed toward the deeper symbolic content of the message, they are less disrupted by changes in voice in later recognition tasks. On the other hand, if listeners are not encoding the meaning of the words, but instead are processing them in a shallower manner, inconsistencies in the voice have a significant effect

on their recognition accuracy. Thus, the initial level of encoding of the spoken words determines how much of an impact the voice will have on memory tasks following acquisition.

Integrating Psycholinguistic and Advertising Research

The ultimate goals of advertising are to increase brand recall, instill confidence in the brand, and finally to sell a product. Because much advertising relies on auditory input using spoken language to transfer information about the product to the target audience, the effects of the speech input must be carefully considered. Based on a number of studies in psycholinguistics, speech perception, and marketing research, several general conclusions can be drawn as to how to best control for and manipulate the effects of voice on listeners' attitudes toward and memory for a product.

A natural first question to ask is whether it is important to be selective when choosing a voice for an ad. Two factors related to the encoding of speech make it clear that voice characteristics are crucial for advertising. First, advertising frequently targets a listener's implicit memory for voices since potential consumers are not generally asked to make explicit, direct judgments about the speaker when confronted with a television or radio commercial. Psycholinguistic research demonstrates that voice information that is encoded implicitly lasts at least up to a week in memory. Since advertising tends to target a listener's implicit memory for voices, voice changes and voice characteristics may have both short-term and long-term effects on the success or failure of a marketing campaign.

The second factor which illustrates the importance of voice characteristics in advertising relates to the level of processing. If listeners already have a vested interest in the product, differences in the voice may not affect listeners' perceptions very much. However, when listeners are not personally invested in the content of the message, the vocal characteristics of the talker have significant effects on their attitudes toward the message. Similarly, when listeners encode stimuli in a shallow manner, voice effects are most apparent. Since advertisers are interested in both retaining current consumers and gaining new ones, they cannot guarantee that the listener will have a prior interest in the product. Therefore, voice characteristics are likely to influence the initial encoding of the message and carryover to the buying intentions of potential consumers.

The research reviewed in this chapter establishes a reliable benefit of voice consistency, revealed by a same-voice advantage. Listeners are faster and more accurate when performing linguistic and memory tasks if the voice of the speaker remains constant. Thus, in advertising, it would be advantageous to use a consistent mapping between a voice and a set of ads for a given product. In addition to consistency, familiarity with voices provides a facilatory effect on a range of language processing tasks. Psycholinguistic research reveals that the intelligibility of a talker's voice in noise is better when listeners are familiar with the speaker. This finding is directly relevant for advertising because many commercials are likely to contain music or may be heard in noisy environments (e.g., in a car). Thus, if the listening environment is not ideal and contains conditions that make perceiving the speech more difficult, having a familiar voice can mitigate these factors. Advertising research shows that brand recall is higher when the voice is a celebrity, and therefore familiar to the listener. Finally, the voice of the spokesperson should of course be highly intelligible. Research has shown that nonnative speakers are less intelligible than native speakers and are thus likely to make less ideal candidates for voice advertising, unless other factors, such as product congruity, are relevant.

A practical concern for advertisers is the cost of air time. A realistic concern is whether speech rate can be increased without causing negative effects on listeners' attitudes and memory for the product. In this area, the evidence is promising. Studies of the effects of speech rate on listeners' attitudes and

memory suggest that in general, increasing the rate has no negative effects, and may in fact be preferred. However, considerations of speech rate are not independent of other concerns of voice and linguistic processing. If advertisers elect to use a faster rate of speech, they must be aware of the possible consequences on the behavior of the intended audience. The psycholinguistic research reviewed here shows that at fast presentation rates, consistency of the voice is more important than at slower rates. Thus, if advertisers use a fast rate, they should be sure to use only a small number of voices. Additional research shows that consistency of the voice is less important if listeners are familiar with the speakers. Thus, if advertisers use well-known "celebrity voices" then it may be possible to use more or varied voices in the ad.

Conclusion

Advertisers have a great deal of control over both the linguistic information of an advertising campaign, as well as the indexical information encoded in the speech signal. Evidence from psycholinguistic studies indicates that voice characteristics have an effect on the processing (encoding, storage, retrieval, and transfer) of linguistic information in the message. Thus, it is not only important that advertisers display care when selecting the particular words and content of the message, but also when choosing a voice to represent a specific marketing campaign. If possible, the voice should remain constant across repetitions of an ad, be familiar (either famous or familiar as the result of repetition), and be produced by a native speaker of the language. It is not surprising that these aspects of a speaker's voice affect language processing; these findings are consistent with psychological research on human factors and ergonomics which shows that response consistency, repetition, and familiarity are important for learning and retention. The rate of speech of an ad may be increased without deleterious effects, although in this case, it is even more important that the voice remain consistent. Because consumers of advertising may only be passively attending to a particular ad, selection of the right voice is even more important in these cases where the effects of voice quality have been found to be most apparent.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Chicago: Aldine Publishing Company.

Alsop, R. (1987). Listen closely: These TV ads might have a familiar ring. *The Wall Street Journal, 22* October.

Bartsch, R.A., Burnett, T., Diller, T.R., & Rankin-Williams, E. (2000). Gender representation in television commercials: updating an update. *Sex Roles, 43*, 735-743.

Chattopadhyay, A., Dahl, D.W., Ritchie, R.J.B., & Shahin, K.N. (2003). Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. *Journal of Consumer Psychology, 13,* 198-204.

DeShields, O.W. Jr., & de los Santos, G. (2000). Salesperson's accent as a globalization issue. *Thunderbird International Business Review, 42*, 29-46.

DeShields, O.W. Jr., Kara, A., & Kaynak, E. (1996). Source effects in purchase decisions: The impact of physical attractiveness and accent of salesperson. *International Journal of Research in Marketing, 13,* 89-101.

Drew, D.G., & Cadwell, R. (1985). Some effects of video editing on perceptions of television news. *Journalism Quarterly, 62*, 828-31, 849.

Gelinas-Chebat, C., & Chebat, J-C. (2001). Effects of two voice characteristics on the attitudes toward advertising messages. *The Journal of Social Psychology, 132,* 447-459.

Glisky, E.L., Polster, M.R., & Routhieaux, B.C. (1995). Double dissociation between item and source memory. *Neuropsychology, 9,* 229-235.

Goldinger, S.D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166-1183.

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.

Hilford, A., Glanzer, M., Kim, K., & DeCarlo, L.T. (2002). Regularities of Source Recognition: ROC Analysis. *Journal of Experimental Psychology: General, 131*, 494-510.

Hirahara, T., & Kato, H. (1992). The effect of F0 on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 89-112). Tokyo: Ohmsha Publishing.

Kreiman, J., & Van Lancker, D. (1988). Hemispheric specialization for voice recognition: Evidence from dichotic listening. *Brain and Language, 34*, 246-252.

LaBarbera, P., & MacLachlan, J. (1979). Time-compressed speech in radio advertising. *Journal of Marketing, 43*, 30-36.

Landis, T., Buttet, J., Assal, G., & Graves, R. (1982). Dissociation of ear preference in monaural word and voice recognition. *Neuropsychologia, 20*, 501-504.

Lane, H. (1963). Foreign accent and speech distortion. *Journal of the Acoustical Society of America, 35*, 451-453.

Leung, L., & Kee, O.K. (1999). The effects of male celebrity voice-over and gender on product brand name recall, comprehension, and purchase intention. *The New Jersey Journal of Communication, 7*, 81-92.

Lightfoot, N. (1989). Effects of talker familiarity on serial recall of spoken word lists. In *Research on Speech Perception Progress Report No. 15* (pp. 419-443). Bloomington, IN: Speech Research Laboratory, Indiana University

MacLachlan, J. (1982). Listener perception of time-compressed spokespersons. *Journal of Advertising Research, 22*, 47-51.

Mayer, R.E., Sobko, K., & Mautone, P.D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*, 419-425.

Megehee, C.M., Dobie, K., & Grant, J. (-2003). Time versus pause manipulation in communications directed to the young adult population: Does it matter? *Journal of Advertising Research, 43*, 281-292.

Miller, N., Maruyama, G., Beaber, R.J., & Valone, K. (1976). Speed of speech and persuasion. *Journal of Personality and Social Psychology, 34*, 615-624.

Misra, S., & Beatty, S.E. (1990). Celebrity spokesperson and brand congruence: An assessment of recall and affect. *Journal of Business Research, 21*, 159-173.

Munro, M.J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition, 20*, 139-154.

Munro, M.J., & Derwing, T.M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

Mullennix, J.W., & Pisoni, D.B. (1990) Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics, 47*, 379-390.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., & Kojima, S. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia, 39*, 1047-1054.

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

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

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 309-328.

Peirce, K. (2001). What if the Energizer *Bunny* were female?: Importance of gender in perceptions of advertising spokes-character effectiveness. *Sex Roles, 45*, 845-858.

Plapler, L. (1974). Some famous spokesmen—and how to use them in ads. *Advertising Age*, April 15, 38.

Schacter, D.L. & Church, B.A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 915-930.

Schlinger, M.J.R., Alwitt, L.F., McCarthy, K.E., & Green, L. (1983). Effects of time compression on attitudes and information processing. *Journal of Marketing, 47*, 79-85.

Shah, N.J., Marshall, J.C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H.J., & Fink, G.R. (2001). The neural correlates of person familiarity: A functional magnetic resonance imaging study with clinical implications. *Brain, 124*, 804-815.

Stevens, A.A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research, 18*, 162-171.

Tsalikis, J., DeShields, O.W. Jr., & LaTour, M.S. (1991). The role of accent on the credibility and effectiveness of the salesperson. *Journal of Personal Selling & Sales Management, 11*, 31-41.

Van Lancker, D.R., Cummings, J.L., Kreiman, J., & Dobkin, B.H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195-209.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia, 25*, 829-834.

Van Lancker, D.R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11*, 665-674.

Whipple, T.W., & McManamon, M.K. (2002). Implications of using male and female voices in commercials: An exploratory study. *Journal of Advertising, 31*, 79-91.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Spoken Word Recognition Development in Children with Cochlear Implants: Effects of Residual Hearing and Hearing Aid use in the Opposite Ear [1]

Rachael Frush Holt[2] **and Karen Iler Kirk**[2]

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

[1]This work was supported by NIH-NIDCD Training Grant T32 DC00012 to Indiana University, NIH-NIDCD Research Grant R01 DC00064 to Indiana University School of Medicine, and Psi Iota Xi National Philanthropic Organization. We thank our collaborators on this work: Laurie Eisenberg and Amy Martinez, House Ear Institute, Los Angeles, CA, and Wenonah Campbell, DeVault Research Laboratory, Department of Otolaryngology – Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN. Portions of this paper appear in: Holt, Kirk, Eisenberg, Martinez, & Campbell (2005). Spoken word recognition development in children with residual hearing using cochlear implants and hearing aids in opposite ears. *Ear and Hearing, 26* (Suppl.), 82S-91S.

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

Spoken Word Recognition Development in Children with Cochlear Implants: Effects of Residual Hearing and Hearing Aid use in the Opposite Ear

Abstract. With broadening candidacy criteria for cochlear implantation, a greater number of pediatric candidates have usable residual hearing in their nonimplanted ears. This population potentially stands to benefit from continued use of conventional amplification in their nonimplanted ears. The purposes of this investigation were to examine the speech and language development of pediatric cochlear implant recipients with either profound or severe hearing loss in their nonimplanted ears, including a subset with severe hearing loss who continued wearing hearing aids in their nonimplanted ears; to evaluate whether children benefit from binaural use of cochlear implants and hearing aids; and to investigate the time course of adaptation to combined use of the devices together. Children were tested on a battery of speech recognition measures in quiet and background noise and language measures in quiet. The results suggest that, although children with different degrees of residual hearing have improved speech recognition and language skills after cochlear implantation, the developmental time course differs for the two groups. Children with severe hearing loss required more than 1 year of cochlear implant experience to demonstrate spoken word recognition gains, whereas children with profound hearing loss showed more benefit during the first year after cochlear implantation. For measures in which group performance differed, children with severe hearing loss had better speech recognition and language skills than the children with profound hearing loss. Furthermore, children with severe hearing loss who continued using hearing aids in their nonimplanted ears benefited from combining the acoustic input received from a hearing aid with the input received from a cochlear implant, particularly in background noise. However, this benefit emerged with experience. Our findings suggest that it is appropriate to encourage pediatric cochlear implant recipients with severe hearing loss to continue wearing an appropriately fitted hearing aid in the nonimplanted ear to maximally benefit from bilateral stimulation. They also suggest that speech and language gains for children with nonimplanted-ear residual hearing occur after children have at least one year of cochlear implant listening experience.

Introduction

Criteria for cochlear implantation in children have changed dramatically since the first individual under 18 years of age received a cochlear implant (CI) in 1980 (Eisenberg & House, 1982). When the U.S. Food and Drug Administration first approved cochlear implantation in children in 1990, criteria for implantation included bilateral profound deafness, age 2 years or older, and demonstration of little or no benefit from amplification (Staller, Beiter, & Brimacombe, 1991). Since that time, candidacy criteria have broadened to include children as young as 1 year of age with profound hearing loss and children at least 2 years of age with severe-to-profound hearing loss. These changes in candidacy criteria are due to improvements in CI technology and increasingly positive speech and language outcomes after cochlear implantation in many users (e.g., Skinner, Fourakis, Holden, Holden, & Demorest, 1996). These changes also have resulted in an increased number of children with CIs who have some degree of residual hearing in their nonimplanted ears. Some of these children have enough residual hearing that they might receive some benefit from using a hearing aid (HA) in their nonimplanted ears. This is a relatively new population at CI centers and a number of investigators have begun to examine whether continued use of a

HA in the nonimplanted ear is beneficial for pediatric CI recipients (Ching, Psarros, & Hill, 2000; Ching, Psarros, Hill, Dillon, & Incerti, 2001).

There are a number of reasons why individuals with CIs might benefit from continued HA use in their nonimplanted ears. First, providing auditory input to the nonimplanted ear might help prevent neural degeneration that is associated with auditory deprivation. Chronic stimulation is known to influence spiral ganglion cell survival in animals (e.g., Miller, 2001). The importance of continued auditory stimulation also has been demonstrated in individuals with CIs and in HA users. In CI recipients, longer periods of profound deafness routinely are associated with poorer speech and language outcomes (Blamey et al., 1992; Cohen, Waltzman, & Fisher, 1993; Gantz et al., 1988). Similarly, word recognition skills in the nonstimulated ear of individuals with bilateral hearing loss fitted with monaural amplification have been shown to worsen over time (Gatehouse, 1992; Hattori, 1993). Thus, the stimulation provided by a HA might help maintain spiral ganglion cell survival in the nonimplanted ear for future advances in hearing restoration or future cochlear implantation.

A second reason why continued HA use might be beneficial to CI users is that monaural listeners (whether it be due to unilateral hearing loss or monaural CI or HA use in listeners with bilateral hearing loss) are unable to benefit from the advantages of bilateral listening, such as binaural summation, localization, squelch effects, head shadow, and aspects of precedence effects. Unable to take advantage of binaural benefits, monaural listeners achieve lower levels of spoken word recognition than binaural listeners, especially in noise (e.g., Giolas & Wark, 1967; Konkle & Schwartz, 1981). Bilateral input might be particularly important for children, because they tend to spend much of the day in school classrooms with high noise levels and long reverberation times (Knecht, Nelson, Whitelaw, & Feth, 2002).

A final reason for continued contralateral HA use in CI users is that the acoustic stimulation provided by a HA might provide the user access to finer spectral and temporal pitch cues in the speech signal that are not resolved well by CIs. A similar argument has been made by Henry and Turner (2003) in discussing the potential benefits of using a HA in an ear implanted with a short electrode array. They suggested that preserving low-frequency hearing in the implanted ear by using a short electrode array and stimulating the apical areas of that cochlea with acoustic amplification (from a HA) together might allow listeners better spectral resolution of the speech signal relative to using a long electrode array alone. Although sensorineural hearing loss in and of itself significantly reduces spectral resolution, Henry and Turner (2003) demonstrated that individuals with sensorineural hearing loss using acoustic stimulation had better spectral resolution than that which is provided by a typical CI. Therefore, it is possible that providing acoustic amplification to the nonimplanted ear with residual hearing might provide additional spectral resolution that could aid in spoken word recognition. On the other hand, due to the severity of the sensorineural hearing loss in the nonimplanted ear of typical CI recipients, the benefit provided by acoustic amplification might be negligible.

Despite all of the potential benefits of HA use in the nonimplanted ear of CI recipients, there is a concern that balancing the two discrepant signals between ears poses some challenges to the listener (Ching, Psarros, et al., 2001). Further, while they learn to use these two discrepant modes of stimulation, listeners must adapt to the novel sensory input provided by a CI. There also is concern that the stimulation received from the nonimplanted ear via a HA might not only result in no further benefit beyond that received from the CI alone, but could in fact cause interference. This interference might result in poorer spoken word recognition when both devices are used simultaneously than when the CI is used alone. In response, many audiologists recommend that children remove the HA from their nonimplanted ears for several months following the initial CI stimulation while they learn to use the new

auditory input. However, this may not be in the best interest of all children. Evidence is accumulating to suggest that continued use of a HA in the nonimplanted ears of children with CIs does in fact aid in speech perception.

A number of investigators have reported higher auditory-only speech perception scores in adults when they used CIs and HAs bilaterally, especially in the presence of competing noise, than when they used either device alone (Armstrong, Pegg, James, & Blamey, 1997; Blamey, Armstrong, & James, 1997; Ching, Incerti, & Hill, 2001; Dooley et al., 1993; Hamzavi, Pok, Gstoettner, & Baumgartner, 2004; Shallop, Arndt, & Turnacliff, 1992; Tyler et al., 2002). Further, Tyler et al. reported that two of their three participants had improved localization ability and Ching, Incerti, et al. (2001) found overall improved localization with combined bilateral CI+HA use relative to monaural CI-only listening.

Similar results have been found in children. Ching et al. (2000) examined speech perception performance in five children (ages 6 to 18 years) with CIs who wore HAs in their nonimplanted ears. Participants had used their CIs for at least 6 months (mean length of CI use was approximately 1 year) and continued to wear HAs in their nonimplanted ears immediately after cochlear implantation. All of the children had profound hearing losses in their nonimplanted ears. Open-set sentence and closed-set consonant recognition (12 alternatives) in 4-talker babble (+10 dB signal-to-noise ratio [SNR]) were significantly better with combined CI+HA use than with CI alone. These differences were primarily due to significantly improved transmission of voicing and manner cues, but not place of articulation cues, in the CI+HA condition relative to the CI-alone condition. Further, 4 of the 5 children had improved horizontal localization abilities in the CI+HA condition relative to the CI-only condition. Similar findings were reported in a larger sample of 11 children in the same age range (Ching, Psarros, et al., 2001). These children also had used their CIs for at least 6 months (mean and individual length of CI use were not provided), continued HA use in their nonimplanted ears immediately following cochlear implantation, and all but one had profound hearing loss in their nonimplanted ears. Children were tested in quiet and in 4-talker babble (+10 dB SNR) on open-set sentence recognition and closed-set consonant recognition (12 alternatives). In the background noise condition, both speech and babble were presented from 0 degrees azimuth in order to minimize head shadow and bilateral squelch effects, thereby underestimating bilateral advantage. Despite this, sentence recognition was significantly better in both quiet and background noise when using a CI combined with a HA in the nonimplanted ear than when using either device alone. However, consonant recognition was significantly better only in the combined CI+HA condition when compared to HA-only performance, not when compared to CI-only performance. The advantage of combining the acoustic and electric stimulation bilaterally was due to better transmission of manner, but not voicing or place of articulation, cues.

At least one investigation did not find an advantage for combining acoustic amplification in the nonimplanted ear with a CI over using a CI alone in children, although such an advantage was noted for a group of postlingually deafened adults. In this early study, Waltzman, Cohen and Shapiro (1992) reported that children who were deaf before age 5 years did not show better spoken word recognition performance when using their CIs with FM systems in the nonimplanted ear than with their CIs alone. Conversely, they found that postlingually deafened adults did show improved spoken word recognition when using both a CI and a HA in the nonimplanted ear than when using either the CI or the HA alone. Despite being fitted with FM systems that can provide more gain, higher output, and an improved SNR relative to HAs, the children failed to improve in the bilateral condition over CI-alone. This developmental difference might stem from language delays typically experienced by the children with severe to profound hearing loss, a concern which would not be expected in postlingually deafened adults. Another particularly important difference was that the children had much less residual hearing in their nonimplanted ears than

the adults. This likely influenced the amount of benefit they received from using an FM system in that ear.

Indeed, in a study of adult combined bilateral CI+HA users, Tyler et al. (2002) suggested that the amount of residual hearing in the nonimplanted ear likely influences the ability of listeners to integrate and capitalize on the input to both ears together. Conversely, Ching, Psarros, et al. (2001) did not find a relationship between amount of residual hearing in the nonimplanted ear and amount of benefit received by children wearing their CIs and HAs together. However, the children who participated in Ching, Psarros et al.'s investigation had at least borderline profound hearing losses in their nonimplanted ears (pure tone averages [PTAs] ranged from 88.3 to 118.3 dB HL). Therefore, amount of residual hearing could be an important factor in determining the benefits of bilateral acoustic-electric hearing in children. Specifically, if children with even more residual hearing were included in such studies, they might demonstrate more benefit from acoustic stimulation of the nonimplanted ear than children with profound hearing loss. With changes in CI candidacy criteria, there are now more children than ever with "aidable" residual hearing in their nonimplanted ears who could benefit from investigating these issues.

One reason why CI candidacy criteria have broadened is that children with some degree of open-set word recognition prior to cochlear implantation demonstrate better spoken word recognition after implantation than do children with very little or no pre-implantation open-set word recognition (Osberger & Fisher, 2000; Staller, Arcaroli, Parkinson, & Arndt, 2002; Zwolan et al., 1997). Furthermore, children with profound hearing loss who use CIs now have word recognition skills that are similar to children with severe hearing loss who use HAs (Boothroyd & Boothroyd-Turner, 2002; Eisenberg, Kirk, Martinez, Ying, & Miyamoto, 2004). Pediatric CI recipients with residual nonimplanted-ear hearing represent a different population than has been studied in the past. The children we studied, including a subset who continued wearing hearing aids in their nonimplanted ears, have more residual hearing in their nonimplanted ears than those children studied by either Ching and colleagues or Waltzman and colleagues, and thus potentially stand to gain more from acoustic input to their nonimplanted ears. Moreover, these children have been tested longitudinally. Although Tyler and Ching and their respective colleagues have suggested that children show benefit from combined CI+HA use, the greatest benefits of combined CI+HA use may emerge over time as the child learns to integrate the two different signals from each ear. The work done on binaural acoustic-electric hearing typically has assessed performance at a single point in time, and therefore the time course of this development is not known. In the current investigation, we followed children longitudinally over the course of 1 to 3 years post-CI activation to examine: 1) the effects of residual hearing on spoken language development in cochlear implanted children; 2) whether pediatric CI recipients with residual hearing in their nonimplanted ears benefit from the bilateral input received by using a HA on their nonimplanted ears; and 3) the time course over which this benefit might emerge.

Methods

Participants

Inclusion criteria included onset of severe-to-profound bilateral sensorineural hearing loss by age 3 years, no other identified disability (such as, physical, visual, or cognitive impairment), etiology of hearing loss other than auditory neuropathy/dysynchrony, and implanted with a current device and fitted with a current speech processing strategy. Based on these criteria, two groups of CI recipients were identified for inclusion in the portion of the investigation designed to examine the influence of amount of residual hearing (in the nonimplanted ear) on spoken language outcomes following cochlear implantation. The first group (Profound) consisted of 124 children with profound sensorineural hearing

loss in their nonimplanted ears. The second group (Severe) consisted of 22 children with severe sensorineural hearing loss in their nonimplanted ears. Ten children in the Severe group continued wearing HAs in their nonimplanted ears following cochlear implantation (NiEHA), whereas the remaining 12 children used their CIs exclusively (No-NiEHA). These two Severe subgroups were included in the portion of the investigation examining the influence of nonimplanted-ear hearing use. Demographic information for the participants is displayed in Table 1. Mean PTAs in the implanted and nonimplanted ears, age at onset of deafness, age at initial CI stimulation, proportion using oral communication, and proportion of females are shown. Standard deviations are displayed in parentheses where indicated.

Participant Group	N	PTA, implanted ear (dB HL)	PTA, nonimplanted ear (dB HL)	Age at onset of deafness (mo)	Age at initial cochlear implant stimulation (mo)	Proportion using oral communication (percent)	Proportion female (percent)
Profound	124	114.5 (6.5)	113.4 (7.9)	1.6 (4.9)	36.9 (20.1)	61%	51%
Severe	22	92.1 (13.2)	80.0 (8.0)	2.2 (6.5)	64.7 (35.2)	55%	41%
NiEHA	10	95.0 (13.6)	81.1 (6.6)	4.4 (9.3)	83.5 (37.0)	80%	40%
No-NiEHA	12	89.5 (12.8)	78.4 (9.0)	0.3 (1.2)	49.1 (25.6)	33%	42%

PTA = pure-tone average, NiEHA = nonimplanted-ear hearing aids.

Table 1. Demographic information for participants.

The primary differences among these groups, other than degree of residual hearing, were amount of hearing loss in the implanted ear prior to cochlear implantation, age at implantation, and the relative proportion of children using oral communication and those using total communication. Although all of the groups on average had profound hearing losses in their implanted ears, hearing losses were approximately 20 dB worse in the Profound group than the other groups of participants. On average, the Severe children were implanted about 2.5 years later than the Profound children and the Severe NiEHA children were implanted approximately 4 years later than the Profound children. This likely reflects recent changes in candidacy criteria to include children with severe hearing loss over the age of 2 years and children with profound hearing loss age 1 year or older. Therefore, under FDA guidelines, a greater number of younger children are eligible for implantation with profound hearing loss than with severe hearing loss. Furthermore, until recently, most cochlear implant teams have been reluctant to implant children who demonstrated some speech understanding with a HA.

There was a slightly larger proportion of children in the Profound group using oral communication (61%) than in the Severe group (55%), although this trend did not extend to the subset of children in the Severe group who used HAs on their non-implanted ears, who were mainly oral communicators (80%). Total communication (TC) combines oral speech with signing in English word order (also known as Signed Exact English). Oral communication (OC) does not use any signing. One potential explanation for a greater proportion of children in the Profound group using OC than in the Severe group is that the children in the Severe group received their CIs much later in life than the children in the Profound group. Therefore, it is possible that children in the Severe group were encouraged to enroll in TC programs early on before they were considered CI candidates, as opposed to the children in the Profound group who received their devices much earlier and may have been

encouraged to enroll in OC programs that promote the use of the auditory signal received from their CIs. We are unable to test this conjecture with this sample, but it is one explanation for the difference in proportion of OC users between the groups.

Sensory Aids

Table 2 displays the number of children implanted with each type of CI system and the speech processing strategies employed. The children who continued wearing HAs in their nonimplanted ears were fitted with a variety of current HAs by each child's clinical audiologist. All of the hearing aids were behind-the-ear styles. The majority of the HAs were digitally programmable, with only a few being fully digital. Both the CIs and the HAs were set at their regular-use settings during testing.

		Cochlear implant				Processing strategy					
Participant group	N	Nucleus 24	Nucleus 22	Clarion	Med-El Combi 40+	MPS*	SPEAK	CIS*	SAS*	ACE*	HiRes
Profound	124	58	40	23	3	4	51	18	5	45	1
Severe											
NiEHA	10	4	0	5	1	0	0	1	5	4	0
No-NiEHA	12	5	0	2	5	0	0	7	2	3	0

**MPS = Multiple Pulsatile Sample, CIS = Continuous Interleaved Sampling, SAS = Simultaneous Analog Stimulation, ACE = Advanced Combination Encoder*

Table 2. Cochlear implant devices and processing strategies used by the participants.

Test Battery

We used a battery of speech recognition and receptive and expressive language tests to evaluate the children's speech and language processing skills. A test battery approach was used for two primary reasons. First, some traditional clinical measures of speech and language processing skills may be insensitive to differences in speech and language abilities among pediatric CI users and within a single child over time (Kirk, Diefendorf, Pisoni, & Robbins, 1997). It is likely that these measures employ vocabulary too advanced for children with severe to profound hearing loss (Boothroyd, 1993; Carney et al., 1993; Moeller, Osberger, & Eccarius, 1986). Second, speech and language processing is hierarchical in nature and therefore, is best examined by combining results from different speech and language measures (Mendel & Danhauer, 1997).

The test battery that was used to examine the influence of amount of residual hearing on CI users' speech and language performance included open- and closed-set word and sentence recognition tests presented with auditory cues only (e.g., no visual cues via speechreading were provided), one test of receptive and expressive language, and one test of receptive vocabulary. A smaller set of spoken word recognition measures was selected for examining the influence of combining acoustic and electric stimulation across ears, because we primarily were interested in examining whether longitudinal changes were evident in listeners' spoken word recognition.

Tests of Spoken Word Recognition. The Grammatical Analysis of Elicited Language – Pre-Sentence Level Test (GAEL-P; Moog, Kozak, & Geers, 1983) was adapted for use as a closed-set word

recognition measure. Before testing, the examiner familiarized each child with the test objects using auditory and visual cues. However, testing was conducted in the auditory-only modality via live-voice. During each trial, the child was presented with four objects: the target and three foils. After the examiner presented the word corresponding to the target item, the child was to respond by pointing to the target object. Performance was scored by percent correct, with chance equaling 25% correct.

The Mr. Potato Head Task (Robbins, 1994) employs a relatively familiar children's toy that consists of a "potato" body along with approximately 20 body parts and accessories that can be attached to the potato body. For some body parts and accessories, there is more than a single exemplar (e.g., several colors and styles of shoes). Children were given a list of 20 auditory-only sentence-length instructions on how to assemble the toy. Their responses were scored for sentence and word correct in percent. An example of one test item is, "He wants *green shoes*," in which "green" and "shoes" are the key words in the sentence. If the child picked up or pointed to any pair of shoes belonging to Mr. Potato Head or if the child picked up a green object, she/he would get 1 out of 2 possible key words correct, but not the sentence correct. If the child picked up or pointed to Mr. Potato Head's green shoes, she/he would get 2 out of 2 key words correct, but not the sentence correct. Finally, if the child put the green shoes on Mr. Potato Head, she/he would get both key words correct and the sentence correct. The word recognition task is considered closed-set because, by chance, the child could select 1 of the 20 body parts or accessories (chance performance = 5%). However, the sentence recognition task is open-set because the child could not carry out the instructions simply by chance.

The Phonetically Balanced–Kindergarten Word Lists (PB-K; Haskins, 1949) is an open-set word recognition test that consists of four lists of 50 phonetically balanced monosyllabic words. However, only three lists are used because the fourth was shown not to be equivalent to the others in Haskins' thesis. For this test, the child is asked to repeat each word after it is presented. Both word correct and phoneme correct scores were calculated in percent correct. Due to a slight protocol difference between testing sites, the PB-K was administered live-voice without lipreading cues in one laboratory (Indiana University School of Medicine) and via recorded compact disc in the other laboratory (House Ear Institute). Just four children in the Severe group (all of whom were also in the subgroup NiEHA) were tested in the laboratory that used recorded PB-K materials; the remaining children in the Severe group and all of the children in the Profound group tested on the PB-K were administered the materials in the laboratory that used live-voice presentation. Using a One-way ANOVA with type of PB-K presentation format (recorded and live-voice) as the between-participant factor and CI-only score at each testing interval as the dependent measure, we found no significant differences in performance on either phoneme or word correct between the four children tested using recorded materials and the other children in the Severe group who received the materials live-voice at any testing interval. Furthermore, using a separate One-way ANOVA, no significant performance differences at any testing interval using any sensory aid condition (CI-only, HA-only and HA combined with CI) were found between the children tested using the recorded materials and the other children in the NiEHA group who received the materials live-voice. Therefore, the data were collapsed across test administration format in analyzing and reporting the results.

The Lexical Neighborhood Test (LNT; Kirk, Pisoni, & Osberger, 1995) is a recorded open-set word recognition test. The LNT consists of two lists of 50 monosyllabic words. Within each list, half of the test items are lexically "easy" (e.g., they occur often in English and have few phonemically similar words, or lexical neighbors, with which they can be confused); the remaining items are "lexically hard" (e.g., they occur rarely in English and have many phonemically similar words with which they can be confused). The recorded version used in our laboratory has five different talkers (two male and three female) within each list (Kirk, Eisenberg, Martinez, & Hay-McCutcheon, 1999). The use of multiple

talkers allows us to assess the child's ability to process speech in the presence of one source of variability encountered in real-world listening situations. Children respond by repeating each word they heard. Their responses were scored as the percent of lexically easy and lexically hard words correctly identified.

The Hearing-In-Noise Test–Children's Version (HINT-C; Nilsson, Soli, & Gelnett, 1996) was modified for use as a test of spoken word recognition in which the percent of words in a sentence correctly repeated at a fixed signal-to-noise ratio (SNR) was used as the dependent measure. The test is composed of 13 lists of 10 sentences that are identifiable to normal-hearing children as young as 5- and 6-years-old. One list was presented in each testing condition. Performance was scored by the percent of words correctly repeated in each sentence.

Tests of Receptive and Expressive Language. The Peabody Picture Vocabulary Test, Third Edition (PPVT-III; Dunn & Dunn, 1997) measures receptive vocabulary development. During testing, the child is presented with a word and is asked to identify it from four line drawings. The presentation format of each word differs depending on the child's primary mode of communication (OC or TC): for children who use OC, the stimuli are presented with auditory and visual cues; for children who use TC, the stimuli are presented with auditory, visual, and sign cues. A receptive vocabulary age is derived and then is converted into a receptive language quotient (receptive vocabulary age divided by chronological age). Language quotients of 1.0 indicate that language and chronological age are equal. In other words, children with language quotients of 1.0 have age-appropriate receptive vocabulary skills. Children with better receptive vocabulary skills than children their age have language quotients above 1.0; children whose receptive vocabulary lags behind their chronologically age-matched peers have language quotients below 1.0.

The Reynell Developmental Language Scales (RDLS; Reynell & Huntley, 1985) assess both receptive and expressive language abilities. Both the receptive and expressive portions of the RDLS use 62 items arranged into 10 sections to assess language development. The receptive portion assesses children's comprehension of a hierarchy of language structures ranging from identifying named objects to inferencing and vocabulary/grammar; the expressive portion assesses children's ability to express a hierarchy of language structures ranging from object labeling to complex instructions. As with the PPVT-III, the RDLS is administered in each child's primary mode of communication and receptive and expressive language quotients were derived from each child's receptive and expressive vocabulary ages.

Procedure

Children were administered the test battery prior to cochlear implantation and at approximately regular 6-month intervals after the CI was first activated. Due to the longitudinal nature of this investigation, not all children were tested on every test administered at each interval due to time constraints, lack of attention for the full test battery, age of the child, or missed appointments. However, all of the participants were administered at least one of the measures in the test battery. Therefore, the reader should note that the number of participants tested in each group varied across tests and testing intervals. The number of participants tested at each interval is noted in the figures.

Licensed speech-language pathologists with training in working with children with CIs administered and scored all of the test measures. The spoken word recognition measures were administered in an auditory-only format, whereas the language measures were presented in the child's primary mode of communication. In contrast to test administration, test instruction for all measures was carried out in the child's primary mode of communication. Both spoken and/or signed responses were acceptable responses for all test measures. Testing was conducted in a quiet room using live-voice

presented at approximately 70 dB SPL, with three exceptions: the PB-K at House Ear Institute, HINT-C sentences and the LNT at both laboratories were presented in a double-walled sound booth using recorded stimuli played through a clinical audiometer. The speech was presented at an average long-term rms of 70 dB SPL. The PB-K and LNT were always administered in quiet, whereas the HINT-C sentences were presented in quiet and +5 dB SNR. For the latter condition, the noise was presented at an average level of 65 dB SPL. Both the speech and noise, where appropriate, were presented from a single speaker placed at 0 degrees azimuth from the listener.

The subgroup of children who continued to wear HAs on their nonimplanted ears were tested in three additional conditions on the PB-K and the HINT-C sentences: 1) HA-only, in which each listener's CI was turned off and the child wore her/his HA at her/his everyday setting; 2) CI-only, in which each listener's HA was removed and the child wore her/his CI at her/his everyday setting; and 3) CI+HA, in which both the CI and HA were activated and worn at everyday settings.

Results

The data were collapsed from blocks of two consecutive 6-month intervals and mean scores by year will be reported to increase statistical power. If a child were tested once during two consecutive 6-month intervals, that score was used in our calculation; if a child were tested twice, the score from the later test interval was used in our calculations. This allowed us to include more data points per 1-year testing interval.

Effects of Residual Hearing

Figures 1 through 6 display results from the speech and language measures for the children in the Profound and Severe groups in order to examine performance differences between children with differing amounts of residual hearing. The top panels show mean group data and +1 standard deviation for the Profound (unfilled bars) and the Severe (black-filled bars) groups. The numbers on the bars in the top panels represent the number of children tested in that specific group for that particular interval. The lower panels display either individual data for the Severe group and average group data for the Profound/TC and Profound/OC groups or average group data for Profound/TC, Profound/OC, Severe/TC, and Severe/OC groups. For test measures that were given to only a few children in the Severe group, individual data are shown; otherwise group data are displayed. Individual data from children in the Severe group who used TC (Severe/TC, light gray-filled bars) and Severe children who used OC (Severe/OC, dark-gray striped bars) appear in the lower panels of Figures 1, 2, and 5, along with mean group data and +1 standard deviation for the children in the Profound group who used TC (Profound/TC, unfilled bars) and who used OC (Profound/OC, black-filled bars). The numbers on each bar in the lower panels of Figures 1, 2, and 5 represent the participant identification (ID) number. Participant ID numbers are consistent, such that participant ID 3 represents the same participant across figures. The lower panels in Figures 3, 4, and 6 display mean group data for the Profound/TC, Profound/OC, Severe/TC, and Severe/OC groups rather than individual data. The numbers on the bars in the lower panels of Figures 3, 4, and 6 represent the number of children tested in that group for that particular interval.

Data from each test measure were entered into a two-way Analysis of Variance (ANOVA) with one repeated measure to determine whether performance changed over time. The between-participant factor was participant group and the within-participant factor was years of CI use. Because participants with missing data cannot be included in the ANOVA, there are instances in which the number of children included in the statistical analysis is smaller than the number of children administered in a given test measure. In some cases the number of children tested in multiple intervals in the Severe group was so

small (e.g., one or two participants) that we selected to also analyze the data using multiple one-way ANOVAs to compare performance differences between the groups.

Figure 1 displays results from the GAEL-P across time. Recall that the GAEL-P is a closed-set measure of spoken word recognition in which chance performance is 25% (indicated by the dashed line in both panels of Figure 1). After 1 year of CI use, both groups demonstrated significant improvements in their spoken word recognition on the GAEL-P, $F(1, 79) = 7.724$, $p = .007$. Overall, the Severe group scores declined during the first year of CI use. However, the Severe children tested at both intervals (IDs 2, 3, and 8) demonstrated gains over time. Prior to cochlear implantation, the Severe group had significantly higher scores than the Profound group, $F(1, 88) = 28.015$, $p < .001$. After one year of device use, the significant difference between the groups disappeared. The bottom panel of Figure 1 shows that, prior to cochlear implantation, the Severe/OC children were performing at ceiling on the GAEL-P, whereas only one Severe/TC child performed above chance. The two children in the Severe/TC group who were tested both before implantation and 1 year post-operatively (IDs 2 and 3) showed improved closed-set word recognition after 1 year of device use, particularly participant ID 3.

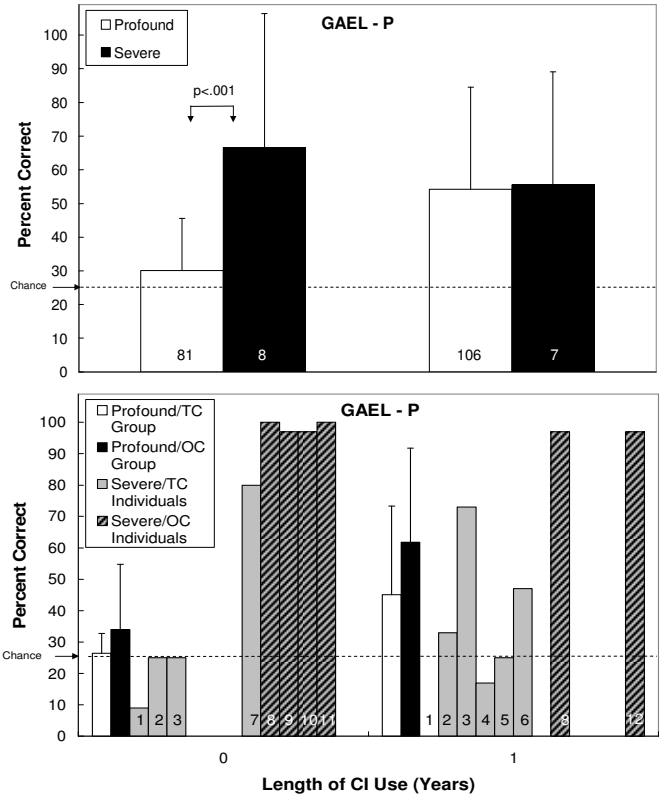


Figure 1. Test results from the GAEL-P. Note that participant 1 had a score of 0 after 1 year of CI use.

The results from the word and sentence recognition portions of the Mr. Potato Head Task are displayed in the first and second columns of Figure 2, respectively. Similar to the GAEL-P results, word recognition performance improved significantly after 1 year of device use for both groups, $F(1, 70) = 6.031$, $p = .017$, as did sentence recognition, $F(1, 68) = 4.957$, $p = .029$. Overall, the Severe group scores

declined during the first year of CI use. However, the Severe children tested at both intervals (IDs 3 and 10) demonstrated gains in word and sentence recognition performance over time. Prior to cochlear implantation, the Severe group had significantly higher word and sentence recognition scores than the Profound group, $F(1, 80) = 66.270$, $p < .001$ and $F(1, 78) = 32.913$, $p < .001$, respectively. After one year of device use, the difference between the groups disappeared. With one exception (ID 7), individual Severe/OC children had higher word and sentence recognition scores than Severe/TC children at both testing intervals.

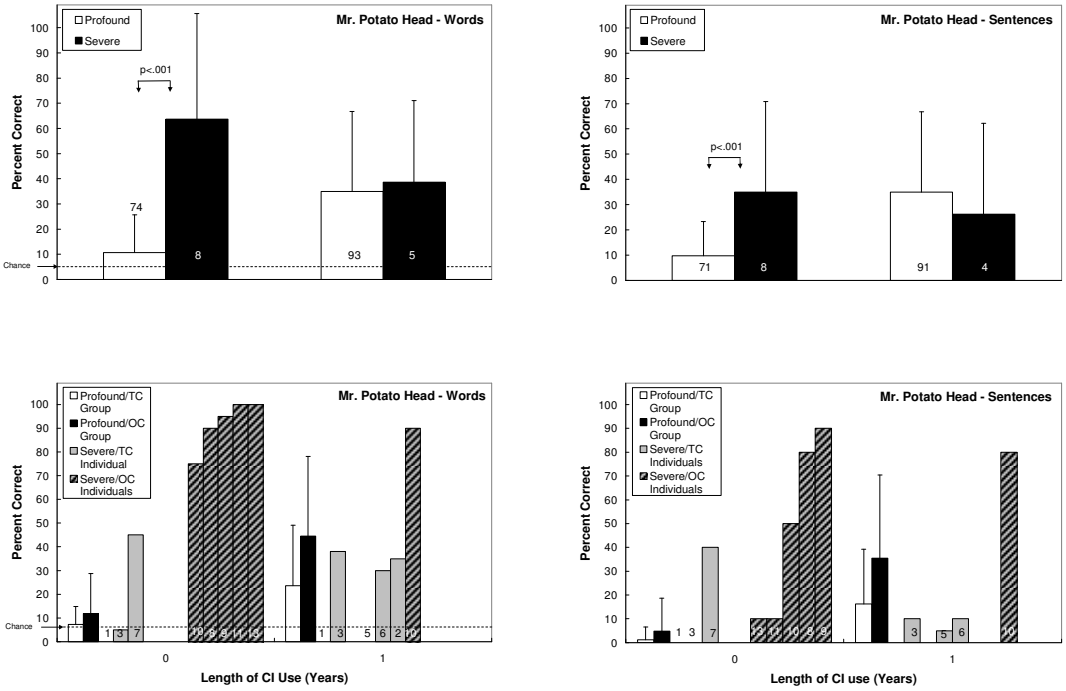


Figure 2. Test results from the Mr. Potato Head Task. Note that participants 1 and 5 had scores of 0 on the word recognition portion and participants 1 and 3 had scores of 0 prior to cochlear implantation on the sentence recognition portion.

Figure 3 displays the results from the phoneme recognition (first column) and word recognition (second column) portions of the PB-K. Unlike the previous word recognition measures, enough data were collected through 2 years of CI use to include this later interval. Including more longitudinal data revealed some interesting effects. Both groups had significant improvements in both phoneme and word recognition over time, $F(2, 16) = 10.488$, $p = .001$ and $F(2, 18) = 6.813$, $p = .006$, respectively. As with the previous word recognition measures, the Severe group performed significantly better than the Profound group prior to cochlear implantation for both phoneme and word recognition on the PB-K, $F(1, 14) = 17.850$, $p = .001$ and $F(1, 14) = 10.368$, $p = .006$, respectively, and the two groups performed similarly after 1 year of CI experience. However, after 2 years of CI use, the Severe group had significantly higher scores than the Profound group on both phoneme and word recognition, $F(1, 27) = 5.943$, $p = .022$ and $F(1, 29) = 12.490$, $p = .001$, respectively. For phoneme recognition, but not for word recognition, there was an interaction between length of device use and degree of residual hearing, $F(2, 16) = 4.978$, $p = .021$. That is, the Profound group made great gains during their first year of CI use, particularly in their phoneme recognition skills, and then plateaued between post-implantation years 1

and 2; whereas the Severe group's gains, although more moderate than those seen in the Profound group, were made between post-implantation years 1 and 2. Finally, there was a trend for OC users to have higher scores than TC users, regardless of degree of residual hearing.

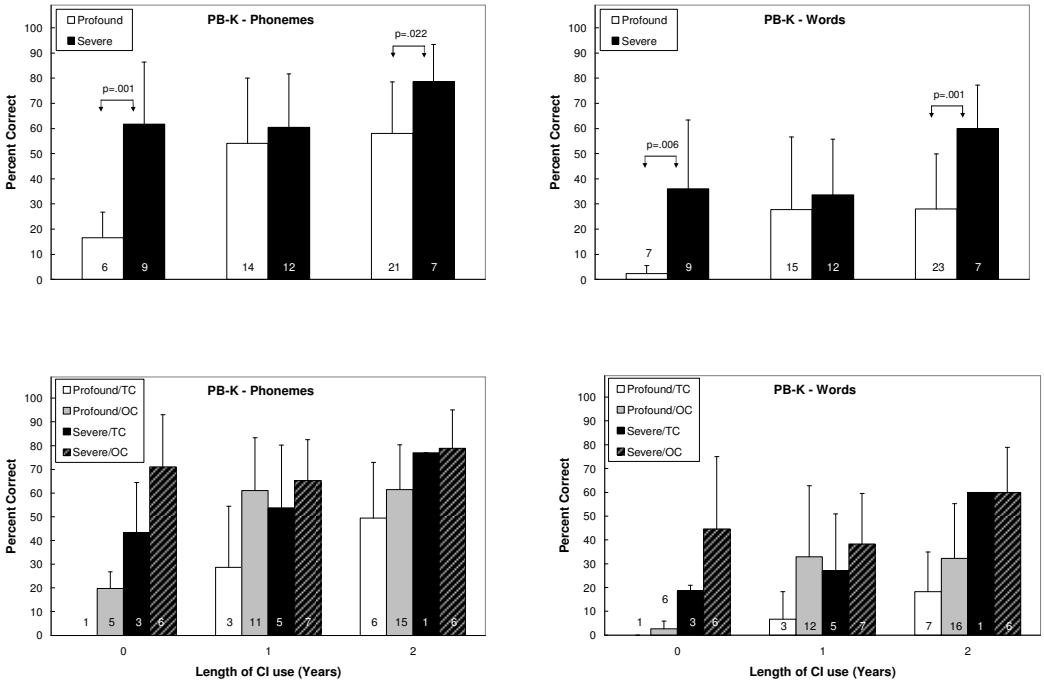


Figure 3. Test results from the PB-K. The phoneme recognition results are shown in the first column and the word recognition results are shown in the second column.

Figure 4 displays the results from the LNT. Too few children were tested on the LNT before cochlear implantation, particularly in the Profound group, to complete a repeated measures ANOVA. There are two reasons that few Profound children were tested near the time of their implantation: the recorded version of the LNT was not yet developed when some of these children received their CIs and many children in the Profound group were too young to be administered the LNT at early testing intervals. Therefore, we will concentrate most of the discussion on the results from the later testing intervals. After 3 years of CI experience, the average Profound participant correctly identified about 50% of the easy and nearly 40% of the hard words, whereas the average Severe participant correctly identified nearly 80% of the easy and greater than 60% of the hard words. Multiple one-way ANOVAs were carried out to examine differences between groups. The Severe group had higher word recognition scores for both lexically easy and hard words than the Profound group at every test interval except on hard words at 2 years post-operative interval (see Figure 4 for *p*-values). Although there were few Severe/TC children tested at any given interval, most of the difference between the Severe and Profound groups primarily was due to the higher average word recognition scores of the Severe/OC children. As shown in the bottom panels of Figure 4, the Severe/OC group had average scores that were at least 20% better after cochlear implantation than those for children in the Severe/TC, Profound/TC and Profound/OC groups.

The Wilcoxon Signed Rank Test was used to examine the effects of lexical difficulty for the Severe and Profound groups separately. The Severe group had significantly higher word recognition scores for easy words than hard words at 1 and 2 years after cochlear implantation, $z = -2.673$, $p = 0.008$ (2-tailed) and $z = -2.375$, $p = 0.018$ (2-tailed), respectively. Only after 3 years of CI experience were the

Profound group's word recognition scores significantly different for easy versus hard words, $z = -5.351$, $p < 0.001$ (2-tailed), with easy words identified with greater accuracy than hard words.

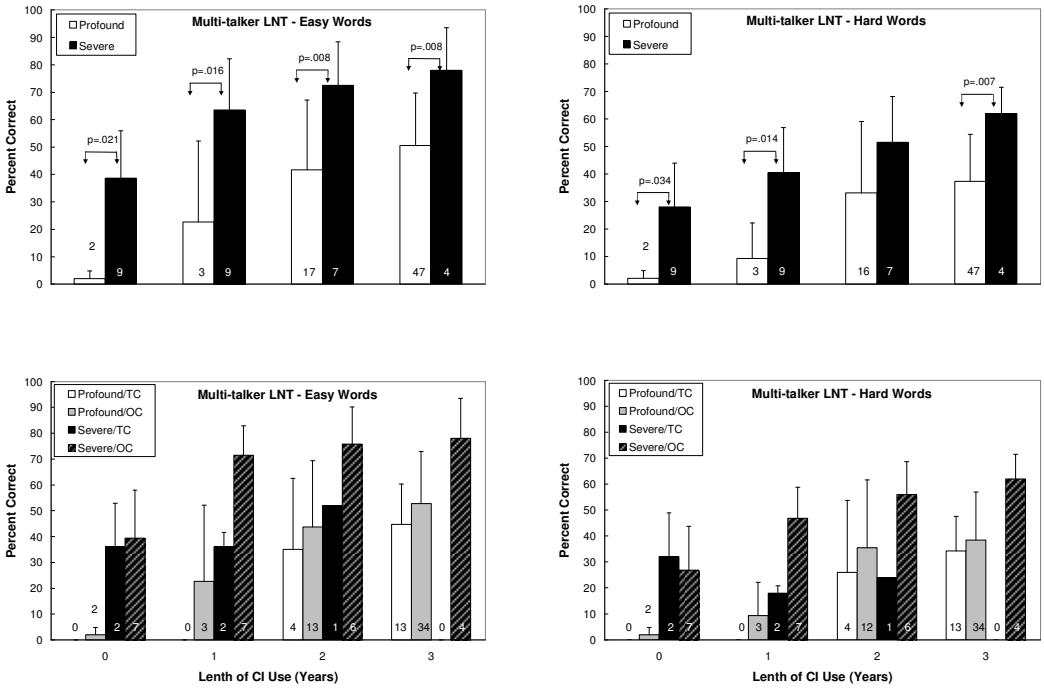


Figure 4. Test results from the LNT. The lexically easy and hard word recognition results are shown in the first and second columns, respectively. Note that no children from the Profound/TC group were tested before cochlear implantation and after 1 year of CI use.

Figure 5 displays the results from the receptive (left panels) and expressive (right panels) portions of the RDLS. Significant improvements in receptive, but not expressive, language beyond those expected by typical development were made by both groups over time, $F(1, 54) = 11.909$, $p = .001$. No significant differences were found between the Severe and Profound groups for both receptive and expressive language at the two intervals tested. Further, no child in the Severe/TC group scored higher at any interval than any Severe/OC child on either the receptive or expressive portion of the RDLS. All but two Severe children (IDs 5 and 8) tested at both intervals made receptive language gains beyond those expected by typical development. Similarly, all but two Severe children (IDs 5 and 10) made expressive language gains beyond those expected by typical development.

Finally, the results from the PPVT are shown in Figure 6. Both groups made significant gains in their receptive vocabulary beyond that expected by typical development through 2 years of CI experience, $F(2, 64) = 7.305$, $p = .001$. The Severe group had higher language quotients than the Profound group prior to cochlear implantation, $F(1, 75) = 15.178$, $p < .001$, at 1 year post-operatively, $F(1, 73) = 6.641$, $p = .012$, and at 2 years post-operatively, $F(1, 74) = 11.450$, $p = .001$. Further, the children in the Severe/OC group appear to be primarily responsible for the Severe group's high performance as a whole, because the Severe/TC group performed similarly to the Profound group (as is evident in the lower panel of Figure 6). Of note is that the Severe/OC group had average language quotients commensurate with their chronological ages after 2 years of CI experience. However, the language quotients at this interval for the Severe/OC group varied greatly from 0.57 to 1.65.

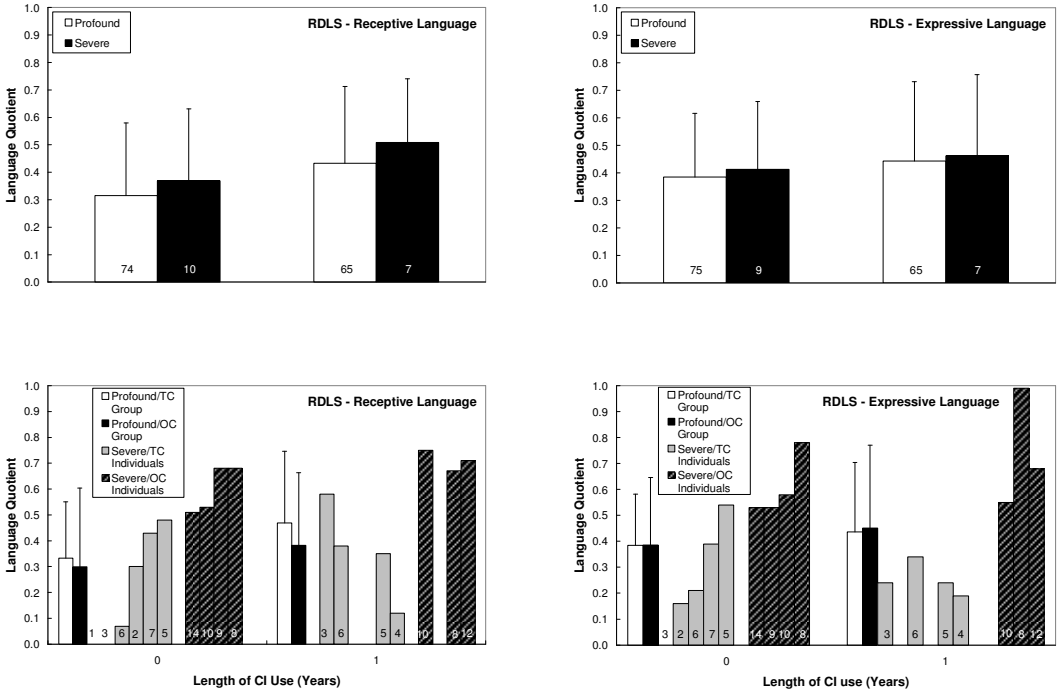


Figure 5. Test results from the RDLS. The receptive and expressive language results are shown in the first and second columns, respectively. Note that participants 1 and 3 had receptive language quotients of 0 and participant 3 had an expressive language quotient of 0 prior to cochlear implantation.

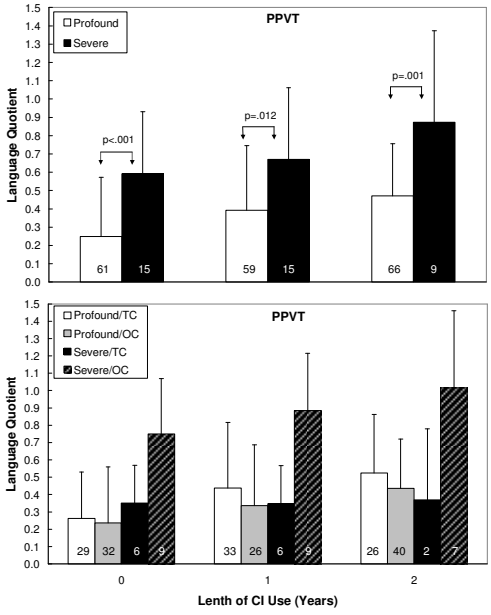


Figure 6. Test results from the PPVT.

Effects of Combined Cochlear Implant and Hearing Aid Use

The subset of participants with severe hearing loss who wore HAs on their nonimplanted ears (NiEHA group) underwent further evaluation of their individual-ear and binaural word recognition skills in both quiet and noise. The PB-K (scored by word correct) was administered in quiet, whereas the sentences from the HINT-C were given in both quiet and at +5 dB SNR. Mean group data and +1 standard deviation on the PB-K are shown in Figure 7. Performance of the children who continued HA use (NiEHA) is shown by unfilled bars (HA-only condition), gray-filled bars (CI-only condition), and black-filled bars (CI+HA condition). For comparison purposes, the striped bars indicate performance of the children who did not continue HA use (No-NiEHA). Note that the data from the No-NiEHA group reflect performance with a HA prior to cochlear implantation (0 years of CI use) and with their CI-only at 1-year intervals after cochlear implantation. The numbers on each bar indicate the number of participants tested from that particular group for the given 1-year interval. No data for the CI-only or CI+HA conditions are displayed at 0 years of CI use because, by definition, participants had not yet received their CIs at this interval. The data from the NiEHA group were analyzed using the Wilcoxon Signed Ranks Test to evaluate differences among device testing conditions. After 2 years of CI use, the children had significantly higher PB-K word identification scores using their CIs and HAs simultaneously than using their HAs alone, $z = -2.023$, $p = 0.043$ (2-tailed). In fact, all five children tested after 2 years of CI use showed this effect. The difference between using CIs and HAs together and using a HA alone 1 year after cochlear implantation approached significance, $z = -1.897$, $p = 0.058$ (2-tailed). At this interval, 7 of the 8 children had higher word recognition scores using both devices simultaneously than using their HAs alone.

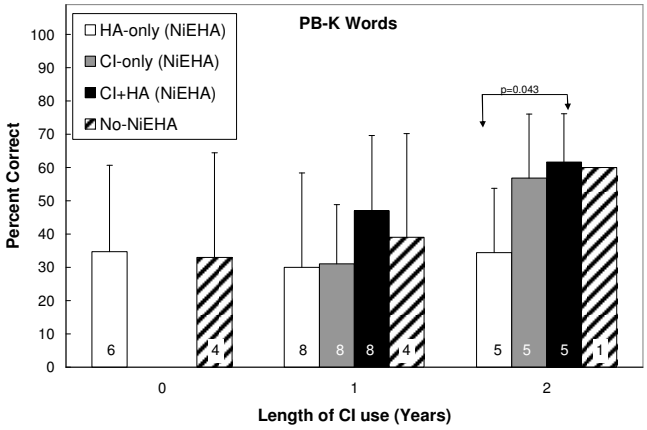


Figure 7. Mean group data and +1 standard deviation on the PB-K in quiet. Performance of the children who used nonimplanted-ear HAs (NiEHA) is indicated by the unfilled bars (HA-only condition), gray-filled bars (CI-only condition), and the black-filled bars (CI+HA condition). Group data for the children who did not continue to wear HAs in their nonimplanted ears (No-NiEHA) are indicated by the striped bars. The numbers on each bar indicate the number of participants tested in a particular group and time interval. No data for the CI or CI+HA conditions are displayed at 0 years of CI use because, by definition, participants had not yet received their CIs at this interval.

Before cochlear implantation, word recognition performance on the PB-K of the children who would later continue nonimplanted-ear HA use and those children who would not continue HA use was very similar (mean scores differed by 2%). Further, the variability across participants within each group

was similar. This indicates that there were no gross pre-implant differences in spoken word recognition in quiet between children who continued HA use and children who stopped wearing HAs after cochlear implantation. Few children in the No-NiEHA group were tested on the PB-K after cochlear implantation, so we were unable to evaluate performance differences statistically.

Figure 8 displays mean group data and +1 standard deviation on HINT-C for the children who continued wearing HAs after cochlear implantation. The top panel shows the results in quiet and the bottom panel shows the results in +5 dB SNR. Similar to the results for the PB-K, the only significant difference between sensory aid conditions in the quiet condition was at 2 years after cochlear implantation between HA+CI and HA-only, $z = -2.023$, $p = 0.043$ (2-tailed). All 5 participants had higher HINT-C word recognition scores in quiet using both sensory aids together than using their HAs alone. In contrast to the results in quiet, word recognition in noise was significantly better after 2 years of CI use in the combined CI+HA condition than in either CI-alone, $z = -2.023$, $p = 0.042$ (2-tailed), or HA-alone, $z = -2.023$, $p = 0.043$ (2-tailed). In both cases, all 5 participants demonstrated this effect.

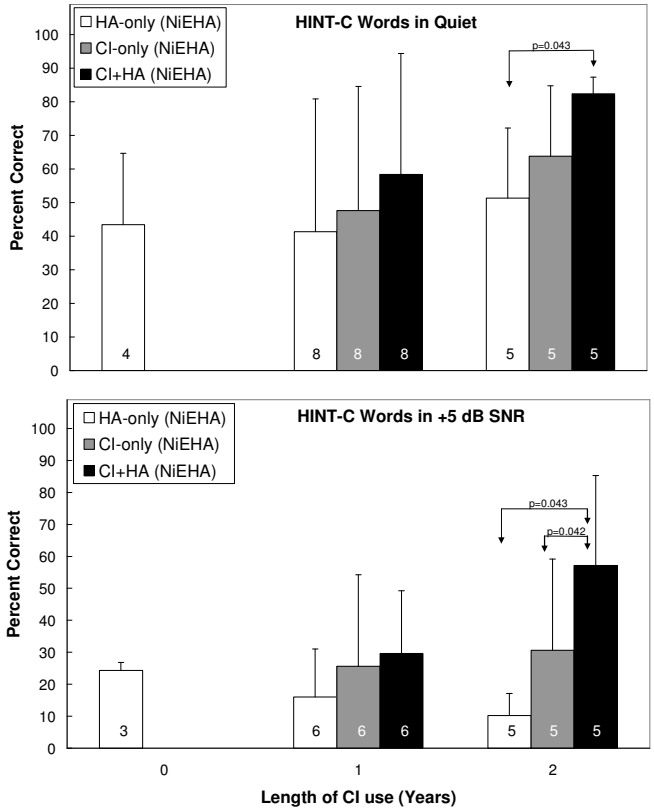


Figure 8. Mean group data and +1 standard deviation on the HINT-C sentences scored by words correctly repeated in quiet (top panel) and in +5 dB SNR (bottom panel) for the children who used nonimplanted-ear HAs (NiEHA). The unfilled bars indicate performance in the HA-only condition; the gray-filled bars indicate performance in the CI-only condition; and the black-filled bars indicate the performance in the CI+HA condition. The numbers on each bar indicate the number of participants tested in a particular group and time interval. No data for the CI or CI+HA conditions are displayed at 0 years of CI use because, by definition, participants had not yet received their CIs at this interval.

Using a repeated measures ANOVA (within factor was years of CI use [1 and 2 years post-operatively] and the between factors were noise condition [quiet and +5 dB SNR] and the sensory aid configuration [HA, CI, and CI+HA]), performance was significantly better in quiet than in noise, F (1, 19) = 9.908, p = 0.005. There was no interaction between device and noise condition. However, there was a significant effect for length of device use, F (1, 19) = 5.857, p = 0.026, and an interaction between length of device use and sensory aid configuration, F (2, 19) = 5.578, p = 0.012. In light of the results from the Wilcoxon Signed Rank Test, performance increased more between 1 and 2 years of experience using a CI and HA simultaneously than using either a HA or a CI alone.

Discussion

Effects of Amount of Residual Hearing

We hypothesized that children with more residual hearing (those in the Severe group) would demonstrate better spoken word recognition and language skills than children with less residual hearing (those in the Profound group) due to their greater number of residual functioning auditory neurons and presumably shorter periods of auditory deprivation. The results from the GAEL-P, Mr. Potato Head Task, and PB-K through 1 year of CI experience suggest that, although children with severe hearing loss in the nonimplanted ear start out with better spoken word and sentence recognition skills prior to cochlear implantation than children with profound hearing loss in the nonimplanted ear, both groups have similar skills after 1 year of CI use. An explanation for this finding is that often individuals in the Severe group who score near ceiling performance before cochlear implantation were not tested again on the same measure, leaving primarily lower performers being tested at the 1-year interval. For example, three of the high Severe/OC performers who were tested on the GAEL-P pre-operatively were not re-tested post-operatively because they had already reached ceiling performance. None of the children in the Severe group who scored above 75% prior to cochlear implantation on the Mr. Potato Head Task were tested 1 year after device use because they were already at ceiling performance. Another reason for the similar group performance after 1 year of CI experience is that the Profound group made gains during the first year of CI use that were significant from both statistical and practical points of view. As a group, the Profound children's word and sentence recognition scores increased by over 20% after 1 year of CI experience, thereby closing the performance gap considerably between the two groups.

More data were collected at later intervals on the PB-K and LNT than the other spoken word recognition measures, allowing for a longitudinal examination of performance changes over a greater period of time. This analysis revealed some important findings. First, similar to the GAEL-P and Mr. Potato Head Task, the Severe group had better spoken word and phoneme recognition skills on the PB-K than the Profound group before cochlear implantation, but these differences disappeared after using a CI for 1 year. However, between post-implantation years 1 and 2, the Severe group showed significant improvements in word and phoneme recognition, whereas the Profound group's performance plateaued between these two intervals. In other words, the children with profound hearing loss primarily made their open-set word recognition gains in the first year of device use, whereas the children with severe hearing loss started out with better spoken word recognition skills before cochlear implantation and, as a group, required over 1 year of device use before they began to demonstrate improvements in these skills. On the LNT, the Severe group's recognition of lexically easy words was superior to that of the Profound group at all test intervals. Their recognition of lexically hard words was better than that of the Profound group after 1 and 3 years, but did not significantly differ after 2 years of CI use. Lexical difficulty effects were found for the Severe group at 1 and 2 years after cochlear implantation; that is, easier words were identified with greater accuracy than hard words. In contrast, the Profound group only showed this effect after 3 years of CI experience. Recall that lexically easy words have few phonemically similar words

(e.g., reside in a sparse lexical neighborhood) with which to compete for recognition. Words in sparse lexical neighborhoods often can be accurately retrieved from memory even if the listener can encode only gross spectral differences. In contrast, lexically hard words have many phonemically similar words competing for attention in a dense lexical neighborhood. Accurate retrieval of lexically hard words from memory requires that the listener encode greater spectral information. Children in the Severe group obtained greater benefit from a HA prior to cochlear implantation than children in the Profound group; thus, they could use gross spectral cues to identify lexically easy words better than hard words even after cochlear implantation. Because a CI is not particularly good at conveying fine spectral detail in the speech signal (such as place of articulation cues), both groups were less proficient at identifying lexically hard words. It should be noted also that the children in the Severe group were implanted at later ages and thus were older at each testing interval than the children in the Profound group. Therefore, the Severe group's word recognition results also may reflect the influence of a better-developed oral vocabulary.

Both groups demonstrated significant improvements in receptive, but not expressive, language and receptive vocabulary beyond those expected by typical development after 1 to 2 years of CI experience, based on their RDLS and PPVT scores. Further, there were no differences on the RDLS in either receptive or expressive language quotients between the Severe and Profound groups. In contrast, group differences were found for receptive vocabulary on the PPVT. The children with severe hearing loss had significantly higher language quotients than the children with profound hearing loss. However, this difference held only for the children with severe hearing loss who use OC, but not for those who use TC. In other words, the children in the Severe group who use OC tended to have higher language quotients than those who use TC. This is in contrast to the Profound group in which children who use TC tended to have higher average receptive language quotients, but not expressive language quotients, than children who use OC. The results from the Profound group are consistent with previous findings from this population (Holt & Kirk, 2005; Kirk, Miyamoto, Ying, Perdew, & Zuganelis, 2002). One commonly cited explanation for TC users scoring higher on language measures than OC users is that the measures are administered in the child's primary mode of communication. Because some of the vocabulary used, especially on the PPVT, has accompanying signs that "look" like the test words, TC users may have an advantage over OC users on these measures. Although we do not know why the Severe group failed to show this typical pattern, we speculate that it might be due to their more intact auditory systems. Specifically, the children in the Severe group using OC might have been able to capitalize on their usable hearing to acquire a larger oral vocabulary in a way that children with Profound hearing loss and children using TC might not have been able to do, perhaps due to limited auditory abilities in the case of the former or more reliance on supplemental signing as opposed to oral skills in the latter.

Despite the fact that we are limited in our interpretation of the results for children using different communication modes, because we cannot randomly assign children to OC or TC environments, there is a trend, especially for children with severe residual hearing loss, to achieve better spoken word recognition and language if they use OC rather than TC. Although our research design does not permit us to infer a cause-and-effect relationship, these results support those of other investigators who have found greater speech and language gains in pediatric CI users with profound hearing loss who use OC compared to those who use TC (Kirk et al., 2002; Osberger et al., 1991; Sommers, 1991). Our results expand upon these findings and suggest that children with more usable hearing might benefit from a rich oral environment that allows them to capitalize on their greater auditory potential. Further research is needed on this population of cochlear implant recipients with more residual hearing to determine if specific communication and educational environments result in better speech and language outcomes.

Effects of Combined Cochlear Implant and Hearing Aid Use

The results from this investigation suggest that after 2 years of CI use, cochlear implanted children with severe hearing loss in the nonimplanted ear demonstrate significantly better word recognition skills when combining a HA on their nonimplanted ears with their CI than when using their HAs alone in quiet listening environments. However, this word recognition benefit does not extend to quiet listening conditions in which they use their CI alone. In contrast, spoken word recognition in background noise is significantly improved by combining a HA with a CI than by using either device alone after 2 years of CI experience. These results were found despite the discrepant signals received by the two ears.

Keys (1947) and Pollack (1948) observed that binaural auditory thresholds are about 3 dB better than monaural thresholds. Based on a 3-dB shift on the performance-intensity functions for words, this improvement in auditory thresholds can result in an 18% improvement in word recognition (Konkle & Schwartz, 1981). CI+HA PB-K word recognition performance in quiet was between 5-16% higher than CI-only performance, somewhat less than the 18% bilateral improvement predicted by Konkle and Schwartz based solely on binaural summation. There are at least three reasons why the actual increase in performance was less than predicted. First, the signals presented to each ear were quite different – one being acoustic and one being electric. Konkle and Schwartz' predictions were based on both ears receiving similar acoustic signals. Second, Konkle and Schwartz' predictions were based on data from adults with normal hearing, whereas our data are from children with severe-to-profound hearing losses. Finally, the majority of children in this investigation displayed delays in their vocabulary development (based on their scores on the Peabody Picture Vocabulary Test [Dunn & Dunn, 1997]), which can negatively influence performance on word recognition measures (Boothroyd, 1993; Carney et al., 1993; Moeller et al., 1986). Despite the differences in mode of auditory stimulation and participant characteristics, these pediatric CI recipients achieved some degree of the bilateral benefit in spoken word recognition; however, it was less than that predicted by Konkle and Schwartz based solely on binaural summation, just one of the identified benefits of bilateral listening.

The individual data also support these group results. After 1 and 2 years of CI experience, 4 of the 8 children and 2 of the 5 children tested on the PB-K, respectively, had significantly higher scores in the bilateral condition than in the CI-only condition (based on the 95% confidence intervals for a 50-item list by Thornton and Raffin [1978]). No child had significantly lower scores on the PB-K in the bilateral condition than in either the CI- or HA-only conditions. The HINT-C scores cannot be directly analyzed using the confidence limits determined by Thornton and Raffin because the words scored are presented in sentences and are not independent of one another. However, descriptively, scores were equivalent to or higher in the bilateral condition than in the CI-only condition for 6 of the 8 children tested in quiet on the HINT-C after 1 year of CI experience. The bilateral scores for the two children who failed to show improvement after 1 year of CI use were 6% and 13% lower than CI-only scores, respectively. After 2 years of CI use, 4 of the 5 children tested had higher scores in the bilateral condition than in the CI-only condition. The fifth child already was scoring at ceiling in the CI-only condition and her/his CI+HA score was only 6% below her/his CI-only score after 2 years of CI use.

For the HINT-C in noise, performance was significantly better in the bilateral condition than in either the CI- or HA-only conditions after 2 years of CI use. Two of the 6 children tested after 1 year of CI use had substantially higher word recognition scores in the bilateral condition than the CI-only condition and two others had nearly equivalent CI+HA and CI-only scores. The bilateral scores for the two children who failed to show improvement after 1 year of CI use were 5% and 23% lower than CI-only scores, respectively. The child with the substantial drop in bilateral relative to CI-only performance

had much better (42%) bilateral than HA-only performance, however. All five children tested after 2 years of CI use had higher scores in the bilateral condition than in either the CI- or HA-only conditions. Moreover, the increase in spoken word recognition received from bilateral listening was larger in noise than it was in quiet, particularly after 2 years of CI experience (27% in noise versus 19% in quiet). These results suggest that the benefit derived from bilateral auditory input is greatest in the presence of background noise.

Large spoken word recognition gains did not appear until at least 2 years of CI use in the CI-only and the combined CI+HA condition. In other words, children with severe hearing loss in their nonimplanted ears require over 1 year of both CI experience and combined CI+HA experience to begin demonstrating gains in CI-only and combined CI+HA word recognition in both quiet and in the presence of background noise. This finding suggests that experience with both signals is needed before monaural CI-only and bilateral CI+HA benefit is evident.

These results support previous work carried out by Ching et al. (2000) and Ching, Psarros, et al. (2001) in which children who had used their CIs for at least 6 months demonstrated better spoken sentence and consonant recognition in quiet and noise when using their CIs and HAs simultaneously than when using their CIs alone. However, our results expand upon theirs by examining the performance of children with more residual hearing in their nonimplanted ears who stand to benefit more from acoustic amplification (e.g., Tyler et al., 2002). Specifically, children in our study had severe hearing loss, whereas those studied by Ching and colleagues had profound hearing loss. Additionally, the children who participated in the current study were followed longitudinally for up to 2 years of CI use, whereas those studied by Ching and colleagues were tested at a single time interval (approximately 1 year after cochlear implantation in Ching et al. [2000]). The longitudinal nature of our study is important because our results suggest that children with severe nonimplanted-ear hearing loss who continue to use HAs in their nonimplanted ears might require up to 2 years of experience before they demonstrate sufficient integration of both signals effectively enough to show significant gains from bilateral input relative to using either device alone.

Conclusions

In summary, these results suggest that children with different degrees of residual hearing in their nonimplanted ears (from severe through profound) benefit in their spoken word and sentence recognition and language skills from cochlear implantation. However, the time course of the changes might be different for the two groups. Children with severe hearing loss might require more than 1 year of CI experience to demonstrate gains in their spoken word recognition, whereas children with profound losses appear to show more of their benefit early on. One potential explanation for the different developmental time course is that prior to cochlear implantation children with severe hearing loss receive auditory input, albeit degraded, whereas children with profound hearing loss receive virtually no auditory stimulation. Therefore, the children with profound hearing loss most likely have not experienced any usable auditory stimulation and thus, must learn to use the input from a CI to both develop and access their mental lexicon. In contrast, children with severe hearing loss have experienced acoustic stimulation and likely achieved limited spoken word recognition with it. When they receive a CI, they need to re-map the perceptual categories they have already learned (however crude they may be). This perceptual re-mapping might take longer than simply learning to use auditory stimulation. In contrast, postlingually deafened adult CI recipients typically have much better spoken word recognition skills than prelingually deafened adult CI recipients, even those with newer CIs and speech processing strategies (e.g., Teoh, Pisoni, & Miyamoto, 2004). However, postlingually deafened adults already have well-established categories. Presumably, it takes experience with the new signal before children who have some

experience with oral language, but who have less well-developed perceptual categories than adults, to re-map these categories.

For measures in which the groups showed performance differences, the children with severe hearing loss had better speech perception and language skills than did the children with profound hearing loss. Furthermore, children with severe hearing loss in their nonimplanted ears benefit from combining the acoustic input received from a HA in the nonimplanted ear with the electric input received from a CI, particularly in background noise, a very common listening environment. However, the benefit emerges after the children adapt to the novel input from the CI and gain experience combining the two signals from the CI and the HA. Importantly, there was only one instance in which bilateral listening was related to a relatively large drop (23%) in word recognition performance relative to the CI-only condition. This occurred for one participant on the HINT-C sentences in noise after 1 year of CI experience. Because this participant was not tested again after 2 years of CI use, we are unable to determine whether the same pattern of performance was maintained with more experience combining the input from both devices.

Overall, our data do not support the concern that input from a HA in the contralateral ear of a cochlear implanted child will cause interference that results in poorer word recognition than when a CI is used alone, even early on when the child is learning to use the novel input from the CI. Indeed, our findings suggest that it is appropriate to encourage children receiving CIs with severe hearing loss in their nonimplanted ears to continue wearing an appropriately fitted hearing aid in their contralateral ears in order to maximally benefit from the input offered to both ears. If a child appears to be struggling to adapt to the novel input of the CI in combination with their HA, it might be prudent to arrange training to the novel CI stimulation without the input from her/his hearing aid during specified listening times. However, our data suggest that these children will likely learn to adapt to both signals over time and will benefit in their spoken word recognition ability from doing so.

This area of research would benefit from investigating whether the advantages of combining a CI with conventional amplification on the contralateral ear seen in a controlled laboratory setting transfer to more real-world settings, such as school, home, and other child-centered activities where both noise and reverberation frequently exist. Further, the benefits of bilateral listening might extend beyond increased word and sentence recognition to improved localization skills, comprehension, attention, and academic achievement. Longitudinal follow-up in these additional areas of development might help determine if the benefits observed in the laboratory influence functional skills needed to participate in all daily living activities. Related to this, is a need to better define the role of communication mode and the optimal educational environment for pediatric CI recipients with more residual hearing. Our results certainly do not answer the question of whether this population benefits most from a TC or an OC environment, but they do imply that these children are capable of capitalizing on the hearing that they do have by using primarily oral modes of communication. Finally, research comparing children who use CIs and HAs in contralateral ears to children with bilateral CIs would be of great benefit. Quantifying any performance differences between these two groups of children would have important implications regarding cost-effectiveness and risk of additional surgery in bilateral cochlear implantation. If significant performance differences are not found, the combination of CIs and nonimplanted-ear HA use arguably allows for improved spoken word recognition skills over a CI alone, while simultaneously reducing auditory deprivation in the nonimplanted ear, thereby preserving that ear for future technological advances in cochlear implantation or hearing restoration.

References

Armstrong, M., Pegg, P., James, C., & Blamey, P. (1997). Speech perception in noise with implant and hearing aid. *American Journal of Otology, 18,* S140-S141.

Blamey, P.J., Armstrong, M., & James, C. (1997). Cochlear implants, hearing aids, or both together? In G.M. Clark (Ed.), *Cochlear Implant,* (pp. 273-277). Monduzzi Editore: Bologna.

Blamey, P.J., Pyman, B.C., Gordon, M., Clark, G.M., Brown, A.M., Dowell, R.C., et al. (1992). Factors predicting postoperative sentence scores in postlinguistically deaf adult cochlear implant patients. *Annals of Otology, Rhinology & Laryngology, 101,* 342-348.

Boothroyd, A. (1993). Profound deafness. In R. S. Tyler (Ed.), *Cochlear Implants: Audiological Foundations* (pp. 1-33). San Diego, CA: Singular Publishing Group, Inc.

Boothroyd, A., & Boothroyd-Turner, D. (2002). Postimplantation audition and educational attainment in children with prelingually acquired profound deafness. *Annals of Otology, Rhinology, & Laryngology, 111* (Suppl. 189), 79-84.

Carney, A.E., Osberger, M. J., Carney, E., Robbins, A. M., Renshaw, J., & Miyamoto, R. T. (1993). A comparison of speech discrimination with cochlear implants and tactile aids. *Journal of the Acoustical Society of America, 94,* 2036-2049.

Ching, T., Incerti, P., & Hill, M. (2001). Binaural benefits for adults who use hearing aids and cochlear implants in opposite ears. *Ear and Hearing, 25,* 9-21.

Ching, T.Y.C., Psarros, C., & Hill, M. (2000). Hearing aid benefit for children who switched from the SPEAK to the ACE strategy in their contralateral Nucleus 24 Cochlear Implant System. *The Australian and New Zealand Journal of Audiology, 22,* 123-132.

Ching, T., Psarros, C., Hill, M., Dillon, H., & Incerti, P. (2001). Should children who use cochlear implants wear hearing aids in the opposite ear? *Ear and Hearing, 22,* 365-380.

Cohen, N.L., Waltzman, S.B., & Fisher, S.G. (1993). A prospective, randomized study of cochlear implants. The Department of Veterans Affairs Cochlear Implant Study Group. *New England Journal of Medicine, 328,* 233-237.

Dooley, G., Blamey, P., Seligman, P.M., Alcantara, J.I., Clark, G.M., Shallop, J.K., et al. (1993). Combined electrical and acoustical stimulation using a bimodal prosthesis. *Archives of Otolaryngology-Head and Neck Surgery, 119,* 55-60.

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

Eisenberg, L.S., & House, W.F. (1982). Initial experience with the cochlear implant in children. *Annals of Otology, Rhinology, & Laryngology, 91* (Suppl. 91), 67-73.

Eisenberg, L.S., Kirk, K.I., Martinez, A.S., Ying, E.A., & Miyamoto, R.T. (2004). Communication abilities of children with aided residual hearing: Comparison with cochlear implant users. *Archives of Otolaryngology–Head & Neck Surgery, 130,* 563-569.

Gantz, B., Tyler, R.S., Knutson, J.F., Woodworth, G., Abbas, P.J., McCabe, B.F., et al. (1988). Evaluation of five different cochlear implant designs: Audiologic assessment and predictors of performance. *Laryngoscope, 98,* 1100-1106.

Gatehouse, S. (1992). The time course and magnitude of perceptual acclimatization to frequency responses: Evidence from monaural fitting of hearing aids. *Journal of the Acoustical Society of America, 92,* 1258-1268.

Giolas, T. & Wark, D. (1967). Communication problems associated with unilateral hearing loss. *Journal of Speech and Hearing Disorders, 32,* 336-343.

Hamzavi, J., Pok, S., Gstoettner, W., & Baumgartner, W. (2004). Speech perception with a cochlear implant used in conjunction with a hearing aid in the opposite ear. *International Journal of Audiology, 43,* 61-65.

Haskins, H.A. (1949). *A phonetically balanced test of speech discrimination for children.* Unpublished master's thesis, Northwestern University.

Hattori, H. (1993). Ear dominance for nonsense-syllable recognition ability in sensorineural hearing-impaired children. Monaural vs. binaural amplification. *Journal of the American Academy of Audiology, 4,* 319-330.

Henry, B.A., & Turner, C.W. (2003). The resolution of complex spectral patterns by cochlear implant and normal-hearing listeners. *Journal of the Acoustical Society of America, 113,* 2861-2873.

Henry, B.A., & Turner, C.W. (2003). *Spectral shape perception and speech recognition in normal hearing, hearing impaired, and cochlear implant listeners.* Paper presented at the Association for Research in Otolaryngology, Palm Beach, FL

Holt, R.F., & Kirk, K.I. (2005). Speech and language development in cognitively delayed children with cochlear implants. *Ear and Hearing, 26,* 132-148.

Keys, J.W. (1947). Binaural versus monaural hearing. *Journal of the Acoustical Society of America, 19,* 629-631.

Kirk, K.I., Diefendorf, A.O., Pisoni, D.B., & Robbins, A.M. (1997). Assessing speech perception in children. In L.L. Mendel & L.J. Danhauer (Eds.), *Audiologic Evaluation and Management and Speech Perception Assessment* (pp. 101-132). San Diego, CA: Singular Publishing Group, Inc.

Kirk, K.I., Eisenberg, L.S., Martinez, A.S., & Hay-McCutcheon, M. (1999). Lexical neighborhood test: Test-retest reliability and interlist equivalency. *The Journal of the American Academy of Audiology, 10,* 113-123.

Kirk, K.I., Miyamoto, R.T., Ying, E.A., Perdew, A.E., & Zuganelis, H. (2002). Cochlear implantation in young children: Effects of age at implantation and communication mode. *The Volta Review, 102,* 127-144.

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.

Knecht, H.A., Nelson, P.B., Whitelaw, G.M., & Feth L.L. (2002). Background noise levels and reverberation times in unoccupied classrooms: Predictions and measurements. *American Journal of Audiology, 11,* 65-71.

Konkle, D., & Schwartz, D. (1981). Binaural amplification: A paradox. In: F. Bess, B. Freeman, & E. Sinclair (Eds.), *Amplification in Education.* Washington, DC: Alexander Graham Bell Association for the Deaf.

Mendel, L.L., & Danhauer, J.L. (1997). Test development and standardization. In L.L. Mendel & L.J. Danhauer (Eds.), *Audiologic Evaluation and Management and Speech Perception Assessment* (pp. 7-13). San Diego, CA: Singular Publishing Group, Inc.

Miller, A.L. (2001). Effects of chronic stimulation on auditory nerve survival in ototoxically deafened animals. *Hearing Research, 151,* 1-14.

Moeller, M.P., Osberger, M.J., & Eccarius, M. (1986). Receptive language skills. In M.J. Osberger (Ed.), *Language and Learning Skills of Hearing-Impaired Students, ASHA Monogram, 23,* 41-54.

Moog, J.A., Kozak, V.J., & Geers, A.E. (1983). *Grammatical Analysis of Elicited Language–Presentence Level.* St. Louis, MO: Central Institute for the Deaf.

Nilsson, J.J., Soli, D.D., & Gelnett, D.J. (1996). *Development of the Hearing in Noise Test for Children (HINT-C).* Los Angeles: House Ear Institute.

Osberger, M.J., & Fisher, L. (2000). Preoperative predictors of postoperative implant performance in children. *Annals of Otology, Rhinology, & Otolaryngology, 109* (Suppl.), 44-46.

Osberger, M.J., Robbins, A.M., Miyamoto, R.T., Berry, S.W., Myres, W.A., Kessler, K.S., & Pope, M.L. (1991). Speech perception abilities of children with cochlear implants, tactile aids, or hearing aids. *American Journal of Otology, 12* (Suppl.), 105-115.

Pollack, I. (1948). Monaural and binaural threshold sensitivity for tones and white noise. *Journal of the Acoustical Society of America, 20,* 52-58.

Reynell, J.K., & Huntley, M. (1985). *Reynell Developmental Language Scales (2nd ed.).* Windsor, United Kingdom: NFER-Nelson.

Robbins, A.M. (1994). *The Mr. Potato Head Task.* Indianapolis, IN: Indiana University School of Medicine.

Shallop, J.K., Arndt, P.L., & Turnacliff, K.A. (1992). Expanded indications for cochlear implantation: Perceptual results in seven adults with residual hearing. *Journal of Spoken Language Pathology and Audiology, 16,* 141-148.

Skinner, M.W., Fourakis, M.S., Holden, T.A., Holden, L.K., & Demorest, M.E. (1996). Identification of speech by cochlear implant recipients with the Multipeak (MPEAK) and Spectral Peak (SPEAK) speech coding strategies. *Ear and Hearing, 17,* 182-197.

Sommers, M.N. (1991). Speech perception abilities in children with cochlear implants or hearing aids. *American Journal of Otology, 12* (Suppl.), 174-178.

Staller, S., Arcaroli, J., Parkinson, A., & Arndt, P. (2002). Pediatric outcomes with the Nucleus 24 contour: North American clinical trial. *Annals of Otology, Rhinology, & Laryngology, 111,* 56-61.

Staller, S.J., Beiter, A. ., & Brimacombe, J.A. (1991). Children and multichannel cochlear implants. In H. Cooper (Ed.), *Practical Aspects of Audiology: Cochlear Implants: A Practical Guide.* Singular Publishing Group, Inc.: San Diego, CA.

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

Thornton, A., & Raffin, M.J.M. (1978). Speech discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research, 36,* 380-395.

Tyler, R.S., Parkinson, A.J., Wilson, B.S., Witt, S., Preece, J.P., & Noble, W. (2002). Patients utilizing a hearing aid and a cochlear implant: Speech perception and localization. *Ear and Hearing, 23,* 98-105.

Waltzman, S.B., Cohen, N.L., & Shapiro, W.H. (1992). Sensory aids in conjunction with cochlear implants. *The American Journal of Otology, 13,* 308-312.

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 and Hearing, 18,* 240-251.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

When and Why Feedback Matters in the Perceptual Learning of Visual Properties of Speech[1]

Stephen J. Winters, Susannah V. Levi and David B. Pisoni

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

[1] This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We would like to thank Christina Fonte, Jen Karpicke, Sara Phillips, and Melissa Troyer for their help in running subjects, constructing stimuli, and analyzing data.

When and Why Feedback Matters in the Perceptual Learning of Visual Displays of Speech

Abstract. This study investigated how feedback can be used to improve the perception of speech from visual-only displays. Participants saw English words spoken in two different, visual-only displays: full-face displays, in which a speaker's whole face is seen under normal lighting conditions, and point-light displays, which preserve only some of the dynamic information that is visible in speech. The participants attempted to identify words in one of these display formats, and then received feedback information about the identity of the words after each trial. Six different forms of feedback were provided to the participants. Participants who saw point-light displays improved most at the visual-only word identification task when they received feedback which re-presented the stimulus in its original, visual form; these participants also improved more rapidly when they received feedback in audio, rather than in orthographic form. However, the form in which feedback was presented to participants who saw full-face displays of speech did not have as strong an effect on their rate of perceptual improvement. In both display conditions, feedback only improved identification accuracy on stimuli which participants had seen before, without facilitating generalization to novel stimuli. These results suggest that the information contained in visual displays of speech is retained in memory in a highly-detailed, modality-specific format, and that observers draw upon this detailed, episodic information in memory in the process of perceptual learning.

Introduction

Formal linguistic theory (e.g., Chomsky & Halle, 1968) typically represents the phonological structure of language in highly abstract terms. These formal structures are assumed to represent the knowledge that an "ideal" speaker-hearer has of the sound structure of his or her language, and are thus independent of the particular physical system in which they may be manifested (Chomsky, 1965). Whenever we perceive speech in the course of everyday life, however, we perceive it in a particular set of idiosyncratic circumstances, as produced by a particular speaker, through a particular medium. The phonological structures underlying the speech that we hear therefore never come to us in an ideal form. Moreover, they are shaped in a wide variety of ways by the unique characteristics, or "indexical properties" of the speakers that we hear (Abercrombie, 1968). For this reason, it is often possible to identify certain personal characteristics of the speakers we hear—such as their age, their gender, their socio-economic or geographic background—as we interpret the linguistic content of what they are saying. Similarly, it is often possible to identify specific characteristics of the medium through which speech is transmitted to us—for example, over the telephone, over the radio, in a noisy (reverberant) room, etc. None of these "indexical properties" or medium-specific characteristics of the speech that we hear, though, are ever encoded into a formal description of phonological structure. Since these properties of speech cannot affect the meaning or content of a linguistic message, they have all traditionally been considered "extra-linguistic" properties of a speech signal (cf. Laver, 1994; Kreiman, VanLancker-Sidtis, & Gerratt, 2005).

It has often been assumed that these extra-linguistic details are discarded in the perception of speech, as listeners pare down what they hear to an essential linguistic core. For instance, Halle (1985) claimed that:

> "...when we learn a new word we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears; for example, we do not remember the voice quality, speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance." (p. 101)

The process by which extra-linguistic details are filtered or removed from the speech signal in order to yield a formal, abstract and sparsely detailed linguistic representation in memory is known as "perceptual normalization". (see Pisoni, 1997, for a review). A growing body of evidence suggests, however, that the perception of speech does not "normalize" away extra-linguistic details, but actually yields representations that include far more talker-dependent and medium-specific detail than is typically contained in a formal description of language (Pisoni & Levi, 2005). It has been shown, for instance, that listeners retain highly detailed information about the voice of the speaker in recognition memory experiments using spoken words (Palmeri, Goldinger, & Pisoni, 1993). Palmeri et al. showed that listeners can recognize words more quickly and accurately when they are re-presented to them in the same voice, rather than in a different voice. This talker repetition effect is robust across varying numbers of speakers and occurs even when listeners are not asked to attend to the voices of the speakers who are producing the words they hear. Similarly, Nygaard, Sommers, and Pisoni (1994) and Nygaard and Pisoni (1998) have shown that listeners can better identify words and sentences in noise if they are spoken by familiar talkers than if they are spoken by unfamiliar talkers. This finding indicates that listeners not only encode and store talker-specific information in memory when listening to speech, but also that they actively use this information when processing the semantic content of spoken messages. These findings are similar to earlier studies showing that readers store information in memory about the orientation and font face of typewritten words that they have read (Kolers, 1973).

On the basis of such evidence, some theoreticians (Johnson, 1997; Pierrehumbert, 2001) have proposed that specific experiences of speech and language are stored in a highly-detailed, medium-specific format in memory. This view of speech perception holds that normalization is unnecessary; instead, linguistic generalizations emerge during perception from the process of extracting meaningful information out of a wide variety of similar category "exemplars" in memory. Further evidence in support of these exemplar models of speech perception comes from Goldinger (1997), who studied the role of lexical frequency in a phenomenon he dubbed "spontaneous vocal imitation." Goldinger reported that listeners who are asked to repeat words produced by other speakers will reflexively mimic the low-level acoustic characteristics of the words that they hear. That is, listeners' repetitions of spoken words will more closely match the acoustic characteristics of the originals if the words are low in frequency. Goldinger hypothesized that this frequency effect occurs because the acoustic structure of specific word repetitions is based on a combination of what the listener hears and the aggregate average of the acoustic details of the listener's experiences of that word in memory. The repetition of low frequency words, which have fewer exemplars in memory, will thus be more heavily influenced by the acoustic structure of the input signal than high frequency words.

While studies such as Goldinger (1997), Palmeri, Goldinger, and Pisoni (1993) and Nygaard, Sommers, and Pisoni (1994) have shown that the auditory processing of speech preserves highly detailed, "extra-linguistic" information in memory, much less is known about the extent to which human observers preserve extra-linguistic or episodic information in the processing of speech in the visual domain. It has been established in a variety of studies that normal-hearing listeners can extract some meaningful linguistic information from visual recordings of speech which completely lack acoustic information (Breeuwer & Plomp, 1986; Demorest & Bernstein, 1992). Sumby and Pollack (1954) found that visual speech signals significantly augment the intelligibility of speech in adverse listening conditions, while McGurk and McDonald (1976) demonstrated that the visual cues to certain speech sounds may override

the audio cues to different speech sounds in the perception of audio-visually mismatched stimuli. The visual perception of speech thus appears to be highly robust and pervasive (Rosenblum, 2005).

Evidence has begun to emerge from recent work that talker- and token-specific details from particular productions of speech are preserved in the process of visual speech perception. Rosenblum, Yakel, Basser, Panchal, Nodarse and Niehus (2002) showed that observers can match talkers in visual-only speech stimuli across point-light and full-face display formats. Lachs and Pisoni (2004b,c) have reported that observers can match individual tokens of words across modalities and various acoustic transformations. The available evidence thus suggests that observers retain "extra-linguistic" characteristics of the visual signal in memory, just as they preserve the fine-grained acoustic-phonetic details of speech in memory during the process of perception.

Pilot Study

In an earlier pilot study (Winters & Pisoni, 2004), we investigated whether the extra-linguistic, modality-specific details of visual experiences of speech could be used to facilitate the perceptual learning of the visual properties of speech. We asked participants to identify isolated, monosyllabic English words from visual-only displays of speech and then provided them with feedback. This feedback information was presented to different groups of participants in one of three different formats. One group received audio-visual feedback, in which they saw the original, visual stimulus again, while simultaneously hearing the word that was spoken in the video. Another group of participants received audio-only feedback, in which they only heard the audio track from the original stimulus video. The third group of participants received orthographic-only feedback, in which an alphabetic display of the word that had been spoken in the original stimulus video was presented to them on the video screen. A fourth group of participants, in a control condition, received no feedback on their responses.

Prior to this study, we expected feedback to improve the participants' ability to identify the words in each silent video, since feedback information would provide the participants with a linguistic interpretation of the speech events they had seen in the silent, visual displays. Furthermore, we hypothesized that improvement would be proportional to the amount of information provided in feedback to the participants about the speech events they saw in each video. In particular, we expected audio-visual feedback to improve participants' identification accuracy more than audio-only or orthographic-only feedback because it re-presented the stimulus in a visual form that exactly matched the participants' memory of their initial experience of that stimulus. We also expected audio-only feedback to improve identification accuracy more than orthographic-only feedback because the audio-only signal would more closely match the idiosyncratic, dynamic structure of the speech events in the original, visual-only stimulus. Orthographic-only feedback, on the other hand, would only provide the participants with a static, symbolic linguistic representation of the word which had been spoken, which would not provide the participants with any detailed information about the dynamics of the speech events in the original visual signal.

We quantified the amount of perceptual learning the participants made by comparing the participants' accuracy in identifying whole words and sub-lexical units (such as phonemes) between the first and the second halves of the experiment. We found that accuracy did improve over the duration of the experiment, but that the amount of improvement in identification accuracy was almost always unrelated to the type of feedback the participants received. Statistically equivalent gains in whole word identification accuracy were made by all groups of participants—even those who received no feedback at all. The only significant effect of feedback type on perceptual improvement emerged in the identification of word-initial phonemes. However, the observed effects of feedback on identification accuracy in this

context did not match what had been predicted. Participants who received orthographic-only and audio-only feedback identified a higher percentage of word-initial phonemes correctly in the second half of the experiment than they did in the first half. The participants who received audio-visual feedback and no feedback, however, made no comparable gains in perceptual improvement.

We speculated that the specific type of feedback might not have affected whole word identification accuracy because none of the word stimuli were ever re-presented to participants after they had received feedback on them. Therefore, the participants could not apply what they had learned through feedback to the process of identifying the same words on subsequent experimental trials. The effect of feedback type on phoneme identification accuracy, on the other hand, may have emerged because certain phonemes were presented in more than one word in the experiment. Participants could therefore apply what they had learned through feedback about the visual properties of phonemes to the process of identifying those same phonemes on subsequent experimental trials.

The metric that was used to assess perceptual improvement in Winters and Pisoni (2004) may have actually obscured gains in improvement made during each half of the experiment itself. The audio-visual feedback group had a small, but not significant advantage in identification accuracy for word-initial phonemes over the other feedback groups. This advantage may have been the result of rapid perceptual learning during the first half of the experiment by the audio-visual feedback group. The audio-visual feedback group may also have been better at identifying word-initial phonemes at the beginning of the experiment than the other groups, independent of their ability to improve in word identification accuracy throughout the experiment.

Current Study

For the present study, we modified the experimental paradigm used in the pilot study in order to determine whether the expected effects of feedback on perceptual learning and the visual perception of speech would emerge under more relevant testing conditions. The visual-only word identification task remained the same in this study, but the number of experimental trials was expanded and split into three separate phases: pre-test, training, and post-test. In the pre-test, participants saw 16 video stimuli without receiving feedback. Performance in this pre-test thus provided a baseline measure for the inherent ability of each group of participants to do the visual-only word identification task. In the training phase, participants saw 64 videos and received feedback after each trial. Most of the videos they saw during training were also presented to them again, during training, after they had already received feedback on those videos. By re-presenting stimuli in this way, we enabled participants to apply what they had learned through feedback directly to the identification of the stimuli they had received feedback on. Recently, Pashler, Cepeda, Wixted, and Rohrer (2005) have shown that re-presenting test stimuli, after participants have received feedback on them, is an effective way of improving the identification of lexical items in a unfamiliar language; hence, we also expected observer identification accuracy to improve after repeated viewings of the same visual stimuli in training. Comparing identification accuracy after successive presentations of each stimulus in training also provided a more objective means by which to gauge the effects of feedback on perceptual improvement in the task than did the arbitrary first half/second half split that had been used in the pilot study. Finally, in the post-test phase, participants saw 16 new videos without receiving feedback on any of them. The structure of the post-test was thus identical that of the pre-test. Participant performance in the post-test could thus be directly compared to their performance in the pre-test to gauge how well the participants had improved in lip-reading accuracy over the course of the experiment. Comparing identification accuracy between pre-test and post-test phases also provided a direct measure of whether any gains in identification accuracy which had been made during training would generalize to novel video stimuli.

Along with the three forms of feedback which were used in the pilot study—audio-visual (AV), audio-only (A) and orthographic-only (O) feedback—participants in this investigation also received feedback in three new forms which combined a simultaneous or sequential presentation of the visual signal with either audio or orthographic information. These new forms of feedback were included in order to provide an equitable means of testing the effects of combining audio and orthographic feedback with visual information on perceptual learning. In the orthographic-visual (OV) feedback condition, orthographic and visual information were simultaneously presented to the observers by superimposing an orthographic representation of the spoken word on the silent visual stimulus. In the sequential feedback conditions, observers first received information about the identity of the spoken word through either an acoustic-only signal (A-then-V feedback) or an orthographic-only presentation of the word on the computer screen (O-then-V feedback) prior to viewing the silent video stimulus again. With these six forms of feedback, the effects of dynamic audio feedback on perceptual learning could be directly compared to the effects of static orthographic feedback along three separate dimensions, two of which involved a re-presentation of the original video stimulus. A summary of these feedback conditions is provided in Table 1.

	Audio	Orthographic
Simultaneous feedback	AV	OV
Sequential feedback	A-then-V	O-then-V
Non-visual feedback	A	O

Table 1. Summary of feedback types.

As in Winters and Pisoni (2004), we expected that feedback would not only improve observers' identification accuracy for visual stimuli on repeated presentations during training, but that certain types of feedback would improve identification accuracy more than others. When observers see a stimulus that they have seen before on a previous trial, they can use what they have learned about the linguistic properties of that stimulus through feedback to help them identify its linguistic content on the repeated presentation. The ability of observers to do this, however, will depend on how much feedback information they encode and store in memory. If observers store all the modality-specific details that they see during feedback (e.g., the visual properties of the spoken word that they have seen in audio-visual feedback, or the dynamic spectral properties of speech that they have heard in audio-only feedback), then feedback which shares more features in common with the visual-only stimuli should improve identification accuracy more than feedback which does not. However, if such modality-specific detail is discarded in the perceptual analysis of the visual or auditory properties of speech, then the type of feedback the observers receive should not affect how much perceptual improvement observers make in the visual-only word identification task. Performance should only improve if they receive some kind of feedback, regardless of the form in which it is presented to them.

By hypothesizing that observers do not discard modality-specific, "extra-linguistic" details of visible speech tokens from memory, we expected that visual feedback would improve identification accuracy more than non-visual feedback. We also expected that audio feedback would facilitate perceptual learning better than orthographic feedback, since audio feedback matches the dynamic information in the visual speech signal while orthographic feedback does not. It was unknown whether sequential feedback would facilitate perceptual learning better than simultaneous feedback. However, we had a priori reasons for expecting that OV feedback would not facilitate perceptual learning as well as AV feedback, because observers must divide their visual attention in attempting to perceive both an

orthographic and a visual representation of a word at the same time. Observers do not need to divide their attention between modalities when either audio or orthographic feedback is presented in sequence with a repetition of the visual speech signal. The sequential feedback conditions were therefore expected to provide a clearer test of the effects of presenting dynamic (audio) vs. non-dynamic (orthographic) feedback, in conjunction with the original visual signal, on observer accuracy in the visual-only word recognition task.

This study investigated the perceptual learning of the visual properties of speech by using two different kinds of visual displays: full-face displays and point-light displays. Full-face displays of speech present a speaker's face under normal, visible lighting conditions. It has been known since Sumby and Pollack (1954) that normal-hearing observers can readily extract meaningful linguistic information from full-face displays of speech. It has also been shown in a wide variety of studies that the ability of observers to perceive speech in visual-only full-face displays improves over the course of a short training experiment, especially if the participants receive feedback (Bernstein, Auer, & Tucker, 2001; Black, O'Reilly, & Peck, 1963; Gesi, Massaro, & Cohen, 1992; Massaro, Cohen, & Gesi, 1993; Massaro & Light, 2004; Walden, Erdman, Montgomery, Schwartz, & Prosek, 1981; Walden, Prosek, Montgomery, Scher, & Jones, 1977).

Point-light displays are animated sequences of illuminated dot patterns (Johansson, 1973). Figure 1 shows two example frames from a point-light display of a person executing a placekick.

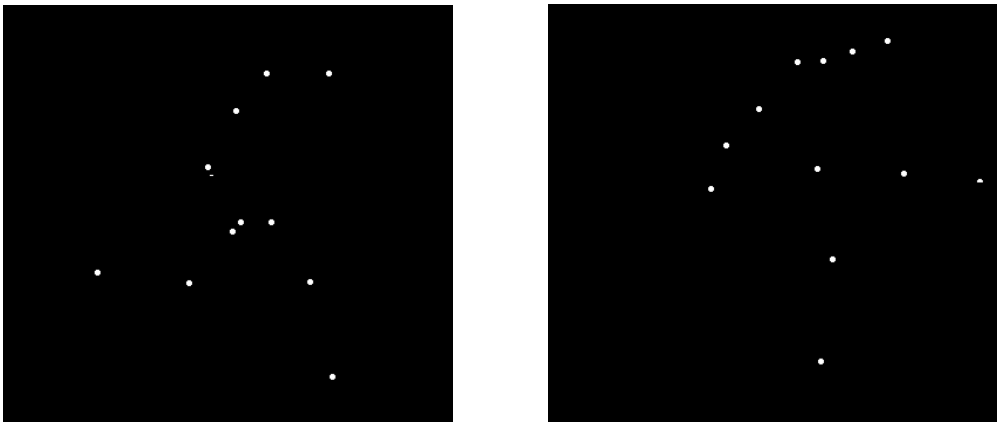


Figure 1. Example frames from a point-light display of a person executing a placekick.

Such point-light displays may be constructed by attaching luminescent points to the major joints on a person's body (e.g., shoulders, elbows, wrists, hips, knees, ankles) and then filming that person executing some motion under darkened lighting conditions. Johansson found that observers could identify human motions in point-light videos made in this way, even though the point-light videos contained much less information than fully illuminated videos of the same motions. Point-light displays of speech were first constructed by Summerfield (1979), who investigated whether they contained enough information to support word identification in adverse listening conditions. Summerfield's point-light displays consisted of only four luminescent points which had been attached to a talker's lips in video-recordings of speech made under darkened lighting conditions. Summerfield presented these point-light displays to observers in conjunction with acoustic speech signals in noise, but found that they did not significantly increase the intelligibility of words over a control condition in which listeners saw no visual information whatsoever. Rosenblum, Johnson, and Saldana (1996), however, found that point-light displays that were made with more than four point-lights in the configuration did improve the intelligibility of speech in noise.

Furthermore, Rosenblum and Saldana (1996) demonstrated that point-light displays also induce McGurk-like effects in audio-visually mismatched tokens of speech. Both Rosenblum, Johnson, and Saldana (1996) and Winters and Pisoni (2004) have also shown that the perception of speech in point-light displays improves rapidly over the course of a short experiment. The visual perception of speech in point-light displays thus exhibits the same basic properties as the visual perception of speech in full-face displays, despite the fact that point-light displays contain much less visual information than fully illuminated displays of speech. It is unknown, however, whether feedback can facilitate the perceptual learning of the visual properties of speech in point-light displays, as it does for full-face displays of speech. Experiment 1 reports the results of using the proposed experimental paradigm to test the effects of feedback on the perceptual learning of the visual properties of speech in point-light displays, while Experiment 2 reports the results of using the same paradigm with full-face displays of speech.

Experiment 1: Perceptual Learning of Point-light Displays of Speech

Methods

Participants. Participants were introductory psychology students at Indiana University in Bloomington, Indiana. A total of 147 subjects participated in the study; seven were removed from analysis (one because of computer failure, one because of self-reported hearing impairment, one who was bilingual, and four because they did not provide responses), resulting in twenty participants in each of the six feedback conditions and twenty in the control condition. All participants were between the ages of 18 and 25, native speakers of English, with normal or corrected-to-normal vision and no reported hearing or language deficits at the time of testing. None of the participants had any previous experience with the audio-visual speech stimuli used in this experiment. All participants received partial course credit for participation in the experiment.

Materials. The point-light displays of speech that were used in this experiment were selected from a digital database originally created by Lachs and Pisoni (2004a). A single talker produced all stimuli. In each video, the talker read one of 96 English words of the form consonant-vowel-consonant (CVC) (e.g., “base”). The talker was video-recorded with glow-in-the-dark dots attached to her face, under black light illumination. The dots were each approximately 3 mm in diameter and were attached to the talker’s face in the pattern shown in Figure 2. There were five dots on each cheek, one on the nose, two on the chin, four on the lower edges of the lips, four on the outer edges of the lips, two on the corners of the lips, one on the tip of the tongue, and two dots each on the lower and upper rows of teeth. Figure 3 shows an example from one of the finished point-light videos.

Figure 2. Configuration of point-lights

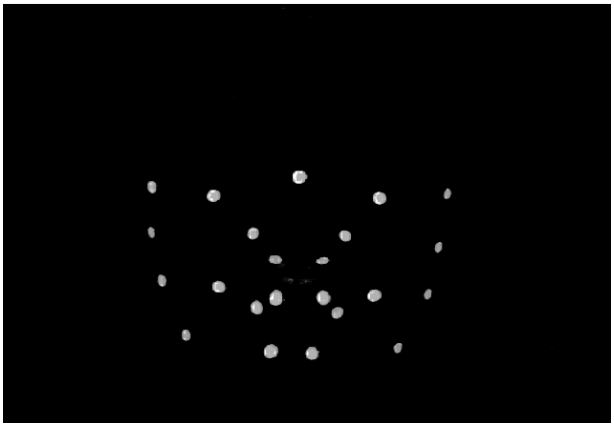


Figure 3. Example frame from point-light video

During pilot-testing, several observers complained that the motion of these point-light displays began before they could orient themselves to the pattern of lights they saw on the screen. The difficulty of interpreting the point-light displays was alleviated by extending the first frame of each video by 500 milliseconds.

For the audio-only feedback condition, the audio track from each video was saved as an .AIFF file using QuickTime software. For the simultaneous orthographic-visual feedback condition, a duplicate set of point-light videos were constructed, using FinalCut Pro Software, in which the spoken word appeared, in lucida grande font face, size 36, centered just beneath the speaker's chin, for the duration of the audio signal in the original video.

Procedure. The experiment was implemented on a customized SuperCard (version 4.1.1) stack, running on a Macintosh G3 computer in a quiet testing room. Participants sat in front of a computer while wearing Beyer Dynamic DT-100 headphones. Their primary task was to watch individual, silent point-light videos and then identify the word that was spoken in each video. After the presentation of each video, the participants answered the on-screen question, "What did the speaker say?" by typing in a response on the computer keyboard.

A brief description of how the point-light stimuli were created was given to the participants prior to the experiment. The participants were told that it might be difficult to perceive speech in the point-light videos, but that they should always provide a response to each video on every trial, even if they had to guess.

The experiment consisted of three phases: pre-test, training and post-test. In the pre-test, the participants saw 16 different point-light videos and responded to each. The participants received no feedback during this portion of the experiment, and none of these videos were shown again during the study.

During training, the participants saw 64 different point-light stimuli. 16 of these stimuli were presented eight times, 16 were presented four times, 16 were presented twice, and 16 were presented only once. The participants received feedback after they responded to each video during training. After the

presentation of feedback, participants clicked on an on-screen button to move on to the next trial. None of the videos which were presented during the training session had been presented during the pre-test.

The experiment used a between-subjects design, so that each group of participants received only one form of feedback each during training. Details of the method used to provide each form of feedback to the various groups of participants are summarized below in Table 2.

Feedback Form	Method of Presentation
AV	Participants saw the original video stimulus again and also heard the original audio track from the test stimulus.
A-then-V	Participants first heard the audio track from the original stimulus and then, after a 500 millisecond pause, they saw the silent, video test stimulus again.
A	Participants heard the audio track from the test stimulus, without seeing the visual stimulus again.
OV	Participants saw a simultaneous presentation of the original, silent visual stimulus and an orthographic representation of the spoken word.
O-then-V	Participants saw the word which had been spoken in text on the computer screen for 1000 milliseconds, and then, after a 500 millisecond pause, saw the silent, visual stimulus again.
O	Participants saw the word which had been spoken in text on the computer screen for 1000 milliseconds, without seeing the visual stimulus again.
N	Participants received no feedback.

Table 2. Method of presenting each type of feedback during training.

The form of the post-test was identical to that of the pre-test. The participants saw 16 different point-light videos and responded to each. The participants received no feedback during this portion of the experiment, and none of the videos which were presented during the post-test had been presented during either training or in the pre-test.

Videos were presented to the participants in random order, with the restriction that no video was ever shown on two consecutive trials during training. The groups of videos which were selected for the pre-test, post-test, and the four different presentation groups in training were also selected at random for each participant. Most participants completed the experiment within 45 minutes.

Analysis. The participants' responses were scored using three levels of analysis: whole word, phoneme, and viseme. A "viseme" denotes a visually equivalent class of sounds (Walden, Prosek, Montgomery, Scher, & Jones, 1977); for instance, the bilabials /b/, /p/ and /m/ belong to the same category. All stimuli and responses were converted into phonetic transcriptions by matching them with entries in the Carnegie Mellon pronouncing dictionary, which lists transcriptions for English words in ARPA notation. Every word in the original video stimuli had a consonant-vowel-consonant form, so the phonetic transcriptions in the dictionary for each match were segmented into an "onset", a "nucleus" and

a “coda.” The vowel in each stimulus word formed its “nucleus”, while the initial consonant was its “onset” and the final consonant was its “coda.”

Even though all participants were informed, prior to the experiment, that they would only see monosyllabic words in each video, many of their responses contained more than one syllable. The “nucleus” of all participant responses—no matter how many syllables they contained—was therefore taken to be the vowel with the highest stress level in the response word or phrase. All segments—including any consonants or vowels—which preceded this response “nucleus” were then taken to be the “onset” of the response, while all segments which followed it were taken to be the response’s “coda.”

For example, one participant gave the response “camera” to the point-light stimulus “thumb.” The phonetic transcription for “camera” in the CMU pronouncing dictionary is /k ae1 m ax0 r ax0/. The /ae1/ vowel has the highest stress level in the word, so it formed the “nucleus” of the response. The /k/ which preceded it then formed the response “onset,” while the final /m ax0 r ax0/ sequence formed the “coda.”

A response was scored correct at the whole word level if the phonetic transcription of its onset, nucleus and coda matched the corresponding transcriptions of the stimulus onset, nucleus and coda. Homonyms (e.g., “wear” and ‘ware”) were thus considered to be correct identifications of whole stimulus words. At the phoneme level of analysis, response onsets, nuclei, and codas were only considered to be correct identifications of their counterparts in the original stimuli if the two matched perfectly. Thus, response onsets or codas which contained more than one segment were considered to be incorrect even if one of those segments formed the original stimulus onset or coda. Thus, the /m ax0 r ax0/ coda of “camera” did not count as a correct identification of the /m/ coda in the “thumb” stimulus, even though an /m/ formed part of the response coda.

The onset and coda of all stimuli and responses were also classified by viseme. The different viseme types included bilabials (/p/, /b/, /m/), labio-dentals (/f/, /v/), interdentals (/θ/, /ð/), dorso-linguals (/t/, /d/, /n/, /k/, /g/, /ŋ/), palato-alveolars (/ʃ/, /ʒ/), and separate categories for /s/, /r/, /h/, /l/, and /w/ (Walden, Prosek, Montgomery, Scher, & Jones, 1977). (Corresponding viseme categories for vowels have not been defined.) Those response onsets and codas which contained more than one segment were classified as having a “mixed” viseme type—unless all of the segments in those onsets and codas happened to agree in viseme type. In this case, the common viseme category or place of articulation was then taken to be the appropriate classification for that portion of the response.

The viseme type of the response onsets and codas were only counted as “correct” identifications if they exactly matched the corresponding viseme features of the stimulus. One participant, for instance, gave the response “damp” to the “dame” stimulus. In “damp,” the coda /mp/ was classified as having a bilabial viseme type, since both /m/ and /p/ are bilabial consonants. This was scored as a correct identification of the stimulus coda viseme, since the coda /m/ in “dame” is also a bilabial. Another participant, however, identified the same “dame” stimulus as “table.” Since the coda of “table” includes both /b/ and /l/ segments, which have a bilabial and a lateral viseme classification, respectively, the coda was categorized as having a “mixed” viseme type. This response was therefore scored as an incorrect identification of the bilabial viseme type in the stimulus coda /m/.

Many of the participants’ responses could not be matched to any entry in the CMU pronouncing dictionary. Responses that were obvious misspellings (e.g., “cheif”) were corrected in the original data file and then matched with the corresponding dictionary entry, while responses that were not obviously

English words (e.g., "rith") were given onset-nucleus-coda transcriptions by hand and then scored accordingly.

Results

Analyses of variance (ANOVAs) were run on the percentages of whole words, phonemes and visemes correctly identified by the participants in order to determine the effects that testing session, feedback type and repetition number had on participants' response accuracy. Three separate ANOVAs were run for all three levels of analysis (words, phonemes, visemes): one comparing participant performance in pre- vs. post-test, another analyzing participant performance on the initial presentation of each group of stimuli during training, and another analyzing participant performance on the final presentation of each group of stimuli during training. Essentially the same pattern of effects on identification accuracy emerged from the separate ANOVAs at the three different levels of analysis, so only the results of the whole word ANOVAs will be reported here, since this is the linguistic level at which participants entered their responses and at which feedback was given.

Pre- vs. Post-test. A repeated measures ANOVA with testing session (pre-test vs. post-test) as a within-subjects factor and feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor revealed a significant main effect of testing session (F (1,132) = 22.634; $p < .001$) but no main effect of feedback. The percentage of words correctly identified in the post-test (3.9%) was significantly higher than the percentage of words correctly identified in the pre-test (1.4%). There was no significant interaction between feedback condition and test session.

Training: Initial Presentation. In order to establish a baseline to measure the effects of stimulus repetition during training on participant response accuracy, a two-way repeated measures ANOVA was run using the percentages of words correctly identified on the initial presentation of each point-light stimulus in training as a dependent variable, feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor and presentation group (one, two, four or eight) as a within-subjects factor. Presentation group number was included as a factor in the ANOVA in order to establish that there were no pre-existing differences in ease of identifiability between the words in each group of stimuli prior to their repetition during training. This ANOVA failed to reveal any significant main effects for feedback or presentation group on whole word identification accuracy. There was also no significant interaction between these two factors. None of the presentation groups thus contained words with inherent differences in intelligibility.

Training: Final Presentation. A repeated measures ANOVA was run using the percentages of words correctly identified on the final presentation of each stimulus during the training phase as a dependent measure in order to determine what effects feedback type and stimulus repetition had on participants' response accuracy. The independent factors in this ANOVA included presentation number (one, two, four, eight) as a within-subjects factor and feedback type (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor. This analysis revealed significant main effects of both presentation number (F (3,130) = 127.193; $p < .001$) and feedback type (F (6,132) = 10.248; $p < .001$).

Table 3 shows the percentage of words correctly identified during training on the final presentation of words in each presentation group. Paired samples t-tests on the main effect of presentation number showed that participants identified words more accurately on the eighth presentation than they did on the fourth, second and first presentations (all $p < .001$). Likewise, they identified more words correctly on the fourth presentation than they did on the second and first presentations (both $p <$

.001), and they identified a significantly higher percentage of words correctly on the second presentation than they did on the first ($p < .001$).

Presentation	% Correct
1	3.2%
2	8.2%
4	15.0%
8	25.6%

Table 3. Percentage of words correctly identified, by presentation group.

Table 4 lists the percentages of words correctly identified by participants in each feedback condition during training. Post-hoc Tukey tests on the main effect of feedback condition indicated that participants in the N feedback group identified fewer words correctly than participants in all of the other feedback groups (AV, A-then-V, O-then-V: $p < .001$; O: $p = .009$; OV: $p = .01$; A: $p = .022$). The AV and A-then-V feedback groups also correctly identified a significantly higher percentage of words than participants in the A feedback group ($p = .037$ in both cases). Comparisons between all other groups yielded no significant differences in word identification accuracy.

Video Presentation	A	O	N
Simultaneous	18.1%	11.8%	---
Sequential	18.1%	16.4%	---
None	11.1%	11.7%	3.7%

Table 4. Percentage of words correctly identified, by feedback condition.

The repeated measures ANOVA also yielded a significant interaction between presentation number and feedback condition (F (18,396) = 3.252; $p < .001$). Figure 4 shows the percentages of words correctly identified by participants in each feedback condition, on the final presentation of each word during training. Post-hoc Tukey tests on the interaction between feedback group and presentation number revealed that, on the second presentation of each stimulus, the AV and A-then-V groups were significantly more accurate than the N feedback group ($p = .004$ and $p = .012$, respectively). On the fourth presentation of each stimulus, the AV, A-then-V and O-then-V groups were significantly more accurate than the N feedback group ($p < .001$, $p < .001$ and $p = .001$, respectively). All groups receiving feedback were significantly more accurate than the N feedback group on the eighth presentation of each stimulus (AV, A-then-V, O-then-V: $p < .001$; A: $p = .002$; OV: $p = .003$; O: $p = .001$). Comparisons between all other feedback groups, for all presentation numbers, yielded no significant differences.

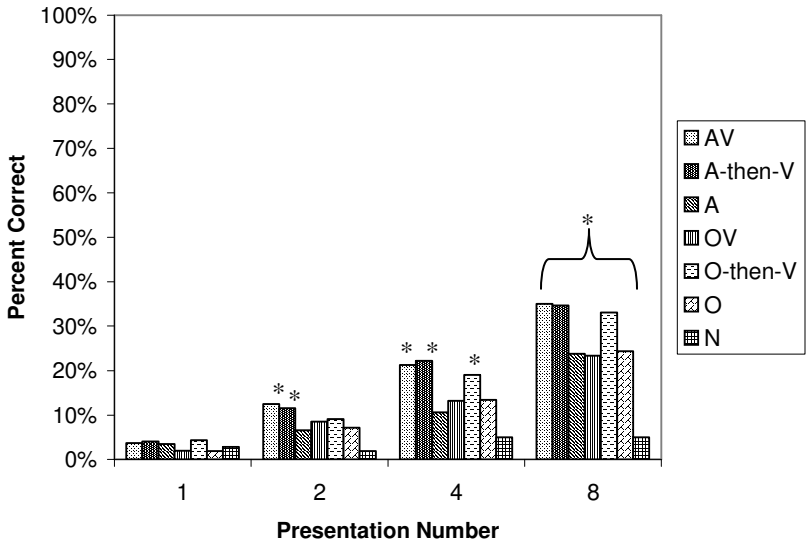


Figure 4. Percentage of whole words correctly identified, by feedback condition and presentation number, on final presentation in training. (* denotes that identification accuracy in that feedback condition was significantly higher at the p=.05 level than identification accuracy in the N feedback condition, after an equivalent number of stimulus presentations.)

Discussion

Two kinds of perceptual learning emerged from the results of this study—one which relied on the type of feedback that participants received, and one which relied on practice. The perceptual learning that relied on practice did not depend on feedback type and took place between pre-test and post-test. All groups of participants, regardless of whether or not they received feedback during training, improved in their ability to identify words from visual-only speech stimuli between pre-test and post-test. As even the group of participants which did not receive any feedback improved in identification accuracy between pre- and post-test, this form of perceptual learning seems to be just a generalized practice effect—the result of acclimating to the experimental procedure and task.

The form of perceptual learning which did rely on feedback type emerged during the training session. The results from training consistently showed two broad trends in identification accuracy: performance improved with successive re-presentations of the same stimuli, and performance was also higher for the groups which received feedback than it was for the group which received no feedback. These two trends also interacted in that several feedback groups performed significantly better than the group which received no feedback after fewer repetitions of the same stimuli.

Taken together, these results indicate that participants were able to use what they had learned through feedback to improve their ability to identify stimuli they had seen before on previous training trials. In other words, simply re-presenting stimuli to participants after they have received feedback on them enabled that feedback information to improve identification accuracy. This effect of feedback did not emerge in the earlier pilot study by Winters and Pisoni (2004) because none of the stimuli were ever presented more than once to participants. Within the confines of a short-term learning experiment, it is clear that feedback only improves participants' ability to identify the stimuli for which they have received feedback. For this reason, the differential effects of feedback on identification accuracy which

emerged in the training session did not generalize to any of the novel stimuli the participants saw in the post-test.

Even when participants attempt to identify stimuli they have seen before and have already received feedback on, they show greater improvement in identification accuracy with some types of feedback than others. In general, the participants who received feedback that re-presented the original point-light stimulus in visual form improved more rapidly in word identification accuracy than did the groups of participants who received only audio or orthographic feedback during training. For instance, the AV and A-then-V feedback groups displayed significantly better word identification accuracy than the N feedback group after the second presentation of the same word during training. The O-then-V feedback group attained the same level of performance after the fourth presentation of the same word in training. The A and O feedback groups, however, did not identify a significantly higher percentage of words than the N feedback group until the eighth presentation of repeated words in training. The group of participants who received OV feedback did not improve in word identification more rapidly than the A and O feedback groups because of the expected difficulties in dividing visual attention between the simultaneous orthographic and visual presentation of the target word.

The pattern of results observed in this study suggests that participants preserved in memory the fine-grained visual details of what they saw during feedback, and they were able to use this information to facilitate their perception of identical visual stimuli on subsequent training trials. When participants do not receive any visual feedback—as in the orthographic-only and audio-only feedback conditions—they cannot directly apply what they learn during feedback to the perception of identical stimuli on subsequent training trials. In order for participants to improve their identification accuracy in these conditions, they must learn to interpret the visual-only stimuli they see on each training trial in terms of the audio or orthographic representations of feedback they have in memory. Since this is more difficult than simply using visual representations in memory to interpret visually identical test stimuli, participants in the A or O feedback conditions did not improve as quickly in identification accuracy during training as the participants who received visual feedback.

We suggested earlier that audio feedback would improve identification accuracy more than orthographic feedback, because audio and visual representations of spoken words share dynamic features which static, orthographic representations do not preserve. The results of the present study only confirmed this hypothesis in the visual feedback conditions. The A-then-V and AV feedback groups both identified a significantly higher percentage of words after fewer stimulus repetitions in training than did the O-then-V and OV feedback groups. There was, however, no difference in the time-course of perceptual improvement between the A and O feedback groups. The fact that audio and visual representations of spoken words share dynamic properties thus only affected perceptual learning when the spoken word was presented to participants in both modalities during feedback. Observers, that is, are apparently only able to use the dynamic information in audio feedback to improve their perception of visual-only stimuli when the dynamic connection between audio and visual representations of speech is explicitly shown to them in feedback. Dynamic information may not affect perceptual learning in the non-visual feedback conditions because it is more difficult for observers to notice the shared dynamic properties of audio and visual representations of speech when there is a longer lag between the presentation of audio feedback and the re-presentation of the original visual stimulus.

The results of this study also showed that, in certain circumstances, sequential feedback facilitated perceptual learning better than simultaneous feedback. The sequential presentation of orthographic and visual (O-then-V) feedback consistently improved identification accuracy more rapidly than the simultaneous presentation of orthographic and visual (OV) feedback. This result is confounded,

however, by the aforementioned difficulties that the simultaneous presentation of two different types of visual information present in OV feedback. No differences emerged in the rate of perceptual improvement between the sequential A-then-V and simultaneous AV feedback groups in whole word identification accuracy, although the A-then-V feedback group did improve more quickly than the AV feedback group in both phoneme and viseme identification accuracy.[2] These results thus provide only limited evidence confirming the efficacy of providing feedback in sequential form on perceptual learning.

Comparing the rate of improvement across different levels of analysis—whole word identification accuracy, phoneme identification accuracy and viseme identification accuracy—revealed very few differences between the various feedback groups. For example, the AV feedback group correctly identified a significantly higher percentage of words than the N feedback group after only two presentations of words in training, but only reached the same level of performance in phoneme and viseme identification accuracy after four presentations of words in training. Other than minor differences such as these, most feedback groups progressed in identification accuracy at comparable rates, regardless of whether their responses were scored in terms of whole word, phoneme or viseme accuracy. The fact that there are no substantial differences in improvement between the whole-word and sub-lexical levels suggests that participants are processing stimuli and making use of feedback on a relatively holistic level, rather than building up their knowledge of the visual properties of speech from smaller perceptual units at the segmental or featural levels. The fact that the participants received more feedback and experience with the various phoneme and viseme categories, in a wider variety of phonological environments, than they did with whole words during training may make the absence of stronger learning effects at the sub-lexical level seem surprising. However, a pattern of learning suggesting that participants might have perceived the visual-only speech stimuli in a holistic fashion echoes earlier findings that the visual perception of speech is largely a holistic, top-down process (Heider & Heider, 1940). The top-down nature of the perceptual learning of the visually degraded point-light stimuli in this experiment provides converging evidence for Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan's (2005) finding that the perceptual learning of noise vocoded speech depends on access to top-down lexical information.

In summary, the results of Experiment 1 indicate that participants preserve more than just formal, abstract and symbolic linguistic structures in memory when they perceive speech in the visual domain. The observers of visual-only speech stimuli in this study preserved in memory modality-specific details of the information they received in feedback, and used that information to help identify the linguistic content of the same stimuli when they were re-presented on subsequent trials in training. Observers do not discard these modality-specific details in the visual perception of speech through some sort of perceptual normalization process. It is for this reason that feedback which re-presents stimuli in their original, visual form facilitates perceptual learning better than other forms of feedback. The results of Experiment 1 also showed, however, that this form of perceptual learning does not generalize to novel stimuli; the observers in this study improved their perception of novel visual-only speech stimuli through practice alone.

Experiment 2: Perceptual Learning of Full-Face Displays of Speech

Experiment 1 showed that feedback facilitates the perceptual learning of the visual properties of speech in point-light displays, which are unusual, highly degraded visual displays of speech. Experiment 2 investigated whether the same forms of feedback would affect the perceptual learning of the visual

[2] A-then-V participants identified a significantly higher percentage of phonemes and visemes than the N feedback group after only two presentations of the same stimuli in training ($p = .014$ and $p = .044$, respectively), but the AV feedback group did not reach the same level of performance until the fourth presentation of the same stimuli in training ($p = .012$ and $p = .008$, respectively).

properties of speech in full-face displays in the same way. Full-face displays of speech differ from point-light displays in two important ways: first, they present more information about speech, and second, observers have more experience perceiving full-face displays of speech than they do perceiving point-light displays of speech. For both of these reasons, we expected participants' ability to perceive speech in full-face displays to be significantly better than their ability to perceive speech in point-light displays.

Based on the results of previous research, we also expected that feedback would improve observers' perception of visual-only, full-face displays of speech. Black, O'Reilly, and Peck (1963), Massaro, Cohen, and Gesi (1993) and Bernstein, Auer, and Tucker (2001) have all shown that orthographic-only feedback improves the visual-only perception of speech. Gesi, Massaro, and Cohen (1992) have shown that audio-visual feedback also improves the visual-only perception of speech. Other studies, such as Walden, Prosek, Montgomery, Scher, and Jones (1977), Walden, Erdman, Montgomery, Schwartz, and Prosek (1981) and Massaro and Light (2004), have provided feedback by informing their participants whether or not their responses in a visual-only speech perception task were correct or incorrect—and then repeated the presentation of those same stimuli until the participants responded correctly. This form of feedback also significantly improved observers' perception of visual-only speech stimuli. While all of these forms of feedback improve the visual-only perception of speech in full-face displays, it is unknown whether some forms of feedback might improve visual-only speech perception more than others, as was shown for point-light displays of speech in Experiment 1.

Methods

Participants. Participants were drawn from the same pool of subjects as in Experiment 1 and met the same criteria for inclusion. A total of 143 subjects participated in the study. Three were not included in the analysis of the response data (two for self-reported hearing impairment, one for a bilingual language background), thus resulting in twenty participants in each of the seven feedback conditions.

Materials. Full-face visual stimuli were selected from the Hoosier Audiovisual Multi-Talker Database (Lachs & Hernandez, 1996). This database consists of digitized videos of ten different talkers (five males and five females) producing 300 CVC English words under normal lighting conditions. Only stimuli produced by one (female) talker were included in Experiment 2; the words produced in those videos were identical to the list of 96 CVC words produced in the point-light videos in Experiment 1. An example frame from one of these videos is shown in Figure 5.

Figure 5. Example frame from a full-face video.

To parallel the presentation of the point-light videos in Experiment 1, the first frame of each full-face video was extended by 500 milliseconds at the beginning of the clip. For the OV feedback condition, a duplicate set of full-face videos were constructed using FinalCut Pro Software in which the spoken word appeared in lucida grande font face, size 36, centered just beneath the speaker's chin, for the duration of the audio signal in the original full-face video. For the audio-only feedback condition, audio files were constructed by simply saving the audio track of each full-face video to an .aiff file using QuickTime software.

Procedure. The procedures for Experiment 2 were identical to those used for Experiment 1. Participants were encouraged to provide a response to each stimulus, even if they were not sure what word had been spoken in the full face video.

Analysis. The analysis of participant responses to the fully illuminated videos in Experiment 2 was identical to the analysis of responses in Experiment 1. This process thus yielded correct identification percentages for whole words, phonemes and visemes, in each of the three parts of the experiment.

Results

Analyses of variance (ANOVAs) were run on the percentages of whole words, phonemes and visemes correctly identified by the participants in order to determine what effects testing session, feedback type and repetition number had on the participants' response accuracy. Three separate ANOVAs were run for all three sets (words, phonemes, visemes) of response accuracy data: one comparing participant performance in pre- vs. post-test, another analyzing participant performance on the initial presentation of each group of stimuli during training, and another analyzing participant performance on the final presentation of each group of stimuli during training. Once again, the same pattern of results emerged from the separate ANOVAs at the different levels of analysis. Therefore, only the results of the whole word ANOVAs are reported here.

Pre- vs. Post-test. A repeated measures ANOVA with testing session (pre-test vs. post-test) as a within-subjects factor and feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) as a between-subjects factor revealed a significant main effect of testing session (F (1,133) = 4.567; p = .034) but no main effect of feedback. The percentage of words correctly identified in the post-test (24.1%) was significantly higher than the percentage of words correctly identified in the pre-test (21.3%). There was no significant interaction between feedback condition and test session.

Training: Initial Presentation. A two-way repeated measures ANOVA was run using the percentage of words correctly identified on the initial presentation of each point-light stimulus in training as a dependent variable and both feedback condition (AV, A-then-V, A, OV, O-then-V, O, N) and presentation group (one, two, four or eight) as independent factors. The results of this ANOVA did not reveal any significant main effects for feedback type or presentation group. There were also no significant interactions between these two factors. None of the presentation groups in Experiment 2 thus contained words with inherent differences in intelligibility.

Training: Final Presentation. A repeated measures ANOVA was run using the percentage of words correctly identified on the final presentation of each stimulus as a dependent measure in order to determine what effects feedback type and stimulus repetition had on participants' response accuracy. The independent factors in this ANOVA were presentation group number (one, two, four, eight), a within-

subjects factor, and feedback type (AV, A-then-V, A, OV, O-then-V, O, N), a between-subjects factor. This ANOVA revealed significant main effects of both presentation number (F (3,131) = 248.461; p < .001) and feedback type (F (6,133) = 12.496; p < .001).

Table 5 shows the percentages of words correctly identified during training on the final presentation of words in each presentation group. Paired samples t-tests on the main effect of presentation number showed that participants identified words more accurately on the eighth presentation than they did on the fourth, second and first presentations (all p < .001). Likewise, they identified more words correctly on the fourth presentation than they did on the second and first presentations (both p < .001), and they identified a significantly higher percentage of words correctly on the second presentation than they did on the first (p < .001).

Presentation	% Correct
1	21.9%
2	34.5%
4	49.3%
8	59.6%

Table 5. Percentage of whole words correctly identified, by presentation number.

Table 6 lists the percentages of words correctly identified during training by participants in the different feedback conditions. Post-hoc Tukey tests on the main effect of feedback condition indicated that participants in the N feedback group identified fewer words correctly than participants in all of the other feedback groups (p < .001 in all cases). Comparisons between all other groups yielded no significant differences in word identification accuracy.

Video Presentation	**A**	**O**	**N**
Simultaneous	45.7%	45.5%	---
Sequential	41.7%	44.2%	---
None	42.2%	46.4%	23.5%

Table 6. Percentage of whole words correctly identified, by feedback condition.

The repeated measures ANOVA also revealed a significant interaction between presentation number and feedback condition (F (18,399) = 5.194; p < .001). Figure 6 shows the percentages of words correctly identified by participants in each feedback condition, on the final presentation of each word in each presentation group, during training. Post-hoc Tukey tests on the interaction between feedback group and presentation number revealed that, on the second presentation of each stimulus, all groups receiving feedback except for the A group correctly identified a significantly higher percentage of words than the N feedback group (O: p < .001; AV: p = .001; O-then-V: p = .007; OV: p = .017; A-then-V: p = .038). On the fourth and eighth presentations of words, all feedback groups identified a significantly higher percentage of words than the N feedback group (p < .001 in all cases). No comparisons between any other feedback groups, for all presentation numbers, yielded significant differences in performance.

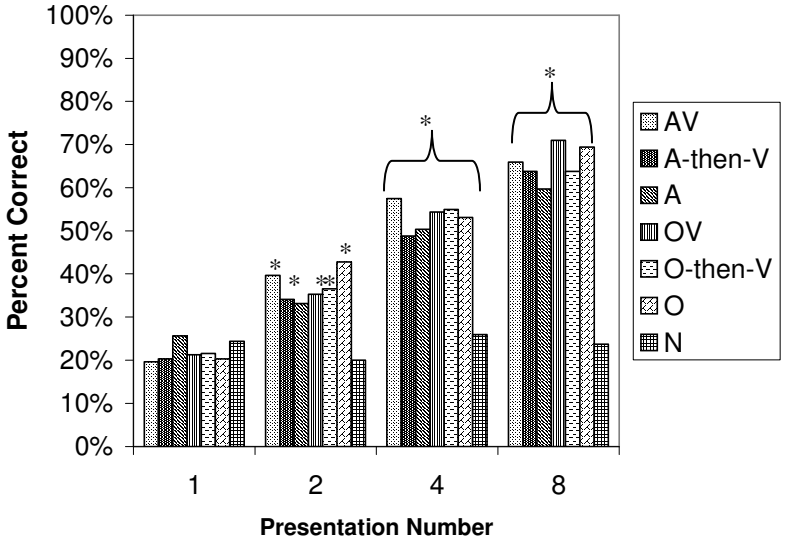


Figure 6. Percentage of whole words correctly identified, by feedback condition and presentation number, on final presentation in training. (* denotes that identification accuracy in that feedback condition was significantly higher at the p=.05 level than identification accuracy in the N feedback condition, after an equivalent number of stimulus presentations.)

Discussion

Taken together, the results obtained in Experiment 2 were quite similar to the findings obtained in Experiment 1. Participants consistently improved in their ability to identify the words, phonemes and visemes in the full-face stimuli between pre- and post-test, regardless of which type of feedback they received. Participants were also able to improve their perception of novel full-face stimuli through practice alone, even when they received no feedback during training at all. Those participants who did receive feedback during training, however, accurately identified a significantly higher percentage of words in repeated presentations of the same stimuli than did participants who received no feedback. The participants in Experiment 2 were thus able to use the information they received in feedback to improve their perception of stimuli they had seen before on previous trials. As in Experiment 1, however, feedback only improved the ability of participants to identify stimuli they had seen before and did not generalize to the novel stimuli presented during the post-test.

As expected, the level of identification accuracy was much higher for the full-face stimuli in Experiment 2 than it was for the point-light stimuli in Experiment 1. A variety of participants who received feedback in Experiment 2 also improved in identification accuracy more rapidly with respect to the participants who did not receive feedback. On only the second repetition of words in Experiment 2, for instance, all groups of participants who received feedback (except for the A feedback group) identified a significantly higher percentage of whole words than the N feedback group. In contrast, only the AV and A-then-V feedback groups reached the same level of performance in whole word identification accuracy on the second presentation of training stimuli in Experiment 1.

Although the effect of feedback on perceptual learning which emerged in training in Experiment 2 did not differ as much between feedback types as it did in Experiment 1, the A feedback group lagged

behind the other groups in its rate of perceptual improvement. The A feedback group did not identify a significantly higher percentage of words than the N feedback group until the fourth presentation of stimuli in training. In this respect, the A feedback group was actually outperformed by the O feedback group, even though the opposite was true for participants in Experiment 1. This result is surprising, given our expectation that audio feedback should facilitate perceptual learning better than orthographic feedback. Aside from this difference, though, all other feedback groups—including those who received AV, OV, O-then-V and O feedback—improved in identification accuracy at the same rate in comparison to the N feedback group.

In Experiment 1, we found that the rate of perceptual improvement which emerged in training depended on the specific type of feedback the participants received. This result provided evidence for the hypothesis that observers store specific instances of visual speech in a highly detailed, modality-specific format in memory. Since the feedback-based improvement which emerged in Experiment 2 did not depend as strongly or as consistently on the type of feedback the participants received, it is difficult to draw similar conclusions about the representation of feedback information in memory for full-face displays of speech. What appears to be the case, instead, is that it is simply easier for observers to make use of feedback—in a variety of different forms—to improve their perception of full-face, visual-only speech stimuli that they have seen before. This may be the case for the same reasons why it is easier to perceive speech in full-face displays than it is to perceive speech in point-light displays: full-face stimuli not only contain more visual information, but observers also have more experience viewing speech in full-face form in everyday life. Participants in an experiment such as this one thus have much more information and knowledge to draw from—as well as more practice perceiving speech in full-face displays—to help them interpret the visual-only stimuli.

Previous studies have shown that the perception of speech in full-face displays improves when participants are provided with either AV or O feedback during a short training experiment. The results of this study corroborate those earlier findings, but also show that, for full-face displays of speech, neither AV nor O feedback improves the visual-only perception of repeated stimuli better than the other. Interestingly, however, one form of feedback that was not used in any of the previous studies—A feedback—did not improve the perception of whole words quite as rapidly as the other forms of feedback that were used in this study. It is possible that participants in the A feedback condition lagged behind the other feedback groups in their rate of perceptual improvement because the structure of the experiment required participants to identify words spoken in visual-only stimuli by typing them into a computer. In order to do this, all participants had to access the orthographic representations of each word. Audio-only feedback is the only form of feedback which did not present the spoken word to the participants in either visual or orthographic form. Without receiving information in either of these forms, it may have been more difficult for the participants in the A feedback condition to use the feedback information they received to interpret visual-only stimuli in orthographic terms than it was for participants in the other feedback groups. This hypothesis could be tested by investigating whether A feedback would facilitate the perceptual learning of the visual properties of speech more rapidly in an experimental paradigm where participants speak their responses, rather than type them. In this paradigm, the output of the participants' spoken responses would be in the same modality as the feedback they receive in the A condition. Participants would not have to generalize across modalities when interpreting their spoken responses in terms of the feedback information they receive under these conditions. It might therefore be easier for them to use A feedback to modify their responses to more closely match what they see in the visual-only speech stimuli.

In summary, the results of Experiment 2 confirmed that the perception of visual-only full-face displays of speech also improved when stimuli were re-presented to participants and feedback was

provided after each trial in a short training experiment. As with the point-light displays of speech in Experiment 1, this perceptual learning effect was only observed in stimuli the participants had seen before and did not, therefore, generalize to novel stimuli. The form in which participants received feedback did not affect the rate of perceptual learning for full-face displays as much as it did for point-light displays, however. The perception of speech in visual-only full-face displays improved rapidly when observers received several different forms of feedback, regardless of whether or not feedback re-presented the stimulus in its original, visual form.

General Discussion

This study investigated whether modality-specific information is preserved in memory when observers are asked to identify spoken words from visual displays of speech. Evidence from the perceptual learning of visual-only point-light displays of speech indicated that highly detailed, modality-specific information was preserved in the visual perception of speech. Feedback that re-presented point-light speech stimuli in their original, visual form to participants improved the perception of point-light displays of speech better than feedback which did not. This result indicates that the fine-grained visual details of point-light stimuli are encoded and retained in memory, and are used to facilitate the perception of previously seen speech stimuli, rather than being discarded in favor of an abstract, linguistic representation resulting from perceptual normalization processes at the time of initial encoding. The visual perception of speech thus preserves "extra-linguistic" visual details in memory just as the auditory perception of speech preserves speaker-specific information (Goldinger, 1997; Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni, 1998; Palmeri, Goldinger, & Pisoni, 1993) and reading preserves information about the font face and orientation of written material (Kolers, 1973).

The effect of visual vs. non-visual feedback on perceptual learning in this study was, however, limited to the perceptual learning of point-light displays of speech. The rate of perceptual learning of full-face displays of speech was, in contrast, largely independent of the form in which feedback was provided to the participants. The perceptual learning of full-face displays of speech may have been insulated from the form in which feedback was provided to participants for at least two reasons. First, compared to point-light displays, full-face displays provide more visual information to observers about the speech events they are trying to perceive, and second, observers have far more experience perceiving full-face speech displays outside the laboratory setting. In future research, it may be possible to test how much each of these two factors interacts with feedback in the perceptual learning of the visual properties of full-face displays of speech by varying them independently in a study on the perception of point-light displays of speech. For instance, more visual information could be provided in point-light displays of speech by simply adding more points of light to the articulators, or even by completely illuminating some articulators, such as the lips, without showing the speaker's whole face. A range of point-light displays along a continuum of informativeness could be created and then presented to observers in a perceptual learning paradigm such as this one. The perceptual learning of point-light displays which are highly visually informative should, presumably, be less susceptible to differences in feedback form than those point-light displays which are less visually informative. Similarly, participants' experience with point-light displays could also be increased through a passive viewing task, in which they simply watch speech in point-light displays (with sound) without responding to what they see. Participants with varying amounts of exposure to point-light displays in such a task could then be tested in a perceptual learning experiment. The gains in perceptual accuracy made by those participants with greater amounts of exposure to point-light displays of speech should be less sensitive to the form of feedback they receive than the gains in perceptual accuracy made by those participants made with less experience to point-light displays. This line of inquiry might, however, prove impractical because enormous amounts of exposure

to point-light displays might be required before the amount of experience observers have with point-light displays would begin to approximate their level of experience with full-face displays of speech.

It is important to note that the feedback-based gains in perceptual accuracy made by the participants in this study were limited to stimuli they had seen before on previous trials. No group of participants in either experiment displayed generalization of what they had learned through feedback to improve their perception of novel visual-only stimuli. However, participants in all feedback conditions, in both experiments, were able to improve their identification accuracy between pre-test and post-test through practice alone. This effect of practice on perceptual learning suggests that there is more to the process of visual speech perception than the mere retention of episodic details in memory. Through practice, observers can apparently "tune in" to the properties of visual-only displays of speech which may be relevant for the identification of linguistic information. In other words, practice enables observers to improve in their ability to pick up information from the visual signal per se, independently of how well they can match up those visual stimuli with feedback information in memory. Since perceptual learning due to practice generalizes to novel stimuli, it likely reflects some form of higher-order knowledge of the articulation of speech sounds in a variety of different phonetic environments.

An important question for future research to answer is what—if any—role the "extra-linguistic" details stored in memory from previous experiences with speech play in the perception of novel speech stimuli. It may be possible to answer this question by modifying the training paradigm that was used in this study in some way so that feedback improves the perceptual identification of novel speech stimuli, as well as repeated stimuli. One modification which may make such generalization possible is to incorporate more variability into the training stimuli, in order to force observers to abstract away from arbitrary, idiosyncratic, instance-specific details which are specific to particular training stimuli. For instance, generalization of feedback-based knowledge may be facilitated by training participants to perceive visual-only speech tokens produced by a wide variety of talkers, rather than just one, as used in the present set of experiments. Similarly, generalization might also be facilitated by training participants on sentence-length stimuli, rather than on individual words. Variations of this "High Variability Training Paradigm" have been used with success in previous work on training Japanese listeners to identify the English /r/-/l/ distinction (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994), and normal-hearing listeners to both understand synthetic speech (Greenspan, Nusbaum, & Pisoni, 1988) and identify dialects of American English (Clopper & Pisoni, 2004). It is also possible that training observers with nonsense words or semantically anomalous sentences may facilitate generalization, because observers would be forced to extract linguistic information solely from what they perceive in the visual-only speech signals, without relying on higher-order knowledge to facilitate processing (Burkholder, 2005).

Developing a training paradigm which can improve the visual perception of speech is, of course, important for practical as well as theoretical reasons. The ability to perceive speech through the visual domain can dramatically improve the intelligibility of speech in adverse listening conditions for normal-hearing listeners, as well as improve the ability of the hearing-impaired to communicate (Bergeson & Pisoni, 2004). The results of this study, along with that of previous research, have shown that incorporating feedback into a training paradigm is an effective way of improving the visual perception of speech. The results of this study also suggest, however, that future researchers should consider the form in which they provide feedback to participants in any given training paradigm. With respect to the perceptual learning of the visual properties of speech, not all forms of feedback are created equal.

References

Abercrombie, D. (1967). Elements of General Phonetics. Edinburgh: Edinburgh University.

Bergeson, T.R. & Pisoni, D.B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implementation. In G. Calvert, C. Spence, & B.E. Stein (Eds.), *Handbook of multisensory processes* (pp. 749-772). Cambridge, MA: MIT Press.

Bernstein, L.E., Auer, E.T., & Tucker, P.E. (2001). Enhanced speechreading in deaf adults: can short-term training/practice close the gap for hearing adults? *Journal of Speech, Language and Hearing Research, 44,* 5-18.

Black, J.W., O'Reilly, P.P., & Peck, K. (1963). Self-administered training in lipreading. *Journal of Speech and Hearing Disorders, 28,* 183-186.

Breeuwer, M. & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America, 79,* 481-499.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Clopper, C.G. & Pisoni, D.B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47,* 207-239.

CMU pronouncing dictionary, version 0.6 Available at http://www.speech.cs.cmu.edu/cgi-gbin/cmudict.

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

Demorest, M.E. & Bernstein, L.E. (1992). Sources of variability in speechreading sentences: a generalizability analysis. *Journal of Speech and Hearing Research, 35,* 876-891.

Gesi, A.T., Massaro, D.W., & Cohen, M.M. (1992). Discovery and expository methods in teaching visual consonant and word identification. *Journal of Speech and Hearing Research, 35,* 1180-1188.

Goldinger, S.D. (1997). Words and voices: perception and production in an episodic lexicon. In K.A. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing*. (pp. 33-66). Academic Press: San Diego, CA.

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

Halle, M. (1985). Speculations about the representations of words in memory. In V. Fromkin (Ed.), *Phonetic linguistics*. (pp. 101-114). Academic Press: Orlando.

Heider, F. & Heider, G.M. (1940). An experimental investigation of lipreading. *Psychological Monographs, 52,* 124-153.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201-211.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K.A. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing*. (pp. 145-166). Academic Press: San Diego, CA.

Kolers, P.A. (1973). Remembering operations. *Memory and Cognition, 1,* 347-355.

Kreiman, J., VanLancker-Sidtis, D., & Gerratt, B.R. (2005). Perception of voice quality. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception*. (pp. 338-362). Blackwell Publishing: Malden, MA.

Lachs, L. & Hernandez, L.R. (1998). Update: the Hoosier Audiovisual Multitalker Database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.

Lachs, L. & Pisoni D.B. (2004a). Specification of crossmodal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America, 116,* 507-518.

Lachs, L. & Pisoni, D.B. (2004b). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 378-396.

Lachs, L. & Pisoni, D.B. (2004c). Crossmodal source identification in speech perception. *Ecological Psychology, 16 (3),* 159-187.

Laver, J. (1994). Principles of phonetics. New York: Cambridge University Press.

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96,* 2076-2087.

Massaro, D.W., Cohen, M.M., & Gesi, A.T. (1993). Long-term training, transfer and retention in learning to lipread. *Perception & Psychophysics, 53,* 549-562.

Massaro, D.W. & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language and Hearing Research, 47,* 304-320.

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

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

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.

Pashler, H., Cepeda, N., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words and facts? *Journal of Experimental Psychology: Learning, Memory and Cognition, 31,* 3-8.

Pierrehumbert, J.B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. (pp. 137-158). John Benjamins: Amsterdam.

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

Pisoni, D.B. & Levi, S.V. (2005). Some observations on representations and representational specificity in speech perception and spoken word recognition. In *Research on Spoken Language Processing Progress Report No. 27*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Rosenblum, L.D. (2005). Primacy of multimodal speech perception. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception*. (pp. 51-78). Blackwell Publishing: Malden, MA.

Rosenblum, L.D., Johnson, J.A., & Saldana, H.M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research, 32,* 921-929.

Rosenblum, L.D. & Saldana, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 318-331.

Rosenblum, L.D., Yakel, D.A., Baseer, N., Panchal, A., Nodarse, B.C., & Niehus, R.P. (2002). Visual speech information for face recognition. *Perception & Psychophysics, 64,* 220-229.

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

Summerfield, A.Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36,* 314-331.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scher, C.K. & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130-145.

Walden, B.E., Erdman, S.A., Montgomery, A.A., Schwartz, D.M., & Prosek, R.A. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech and Hearing Research, 24,* 207-216.

Winters, S.J. & Pisoni, D.B. (2004). Some effects of feedback on the perception of point-light and full-face visual displays of speech: a preliminary report. In *Research on Spoken Language Processing Progress Report No. 26* (pp. 139-164). Bloomington, IN: Speech Research Laboratory, Indiana University.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Sound Similarity Relations in the Mental Lexicon: Modeling the Lexicon as a Complex Network[1]

Vsevolod Kapatsinski

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

[1] Preparation of this chapter was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012). Many thanks to Luis Hernandez for designing the program used to create neighbor pairs from a list of words, as well as to Katy Börner, Tom Gruenenfelder, and David Pisoni for helpful feedback on this project.

Sound Similarity Relations in the Mental Lexicon: Modeling the Lexicon as a Complex Network

Abstract. The standard definition of neighborhood density defines two words to be neighbors if they differ by one and only one segment (Luce & Pisoni, 1998). This definition assumes that the length of the shared part is irrelevant to sound similarity. However, confusability of non-linguistic sound sequences (Fallon Coble & Robinson, 1992; Kidd & Watson, 1992) as well as judgments of sound similarity between spoken pseudowords (Kapatsinski, in press, b) depend on the proportion of total duration of the word or sound sequence that is mismatched, and not on the absolute duration of the mismatch. To bring the definition of neighborhood into alignment with these results, this paper defines words to be neighbors if they share at least two thirds of their total duration, measured in segments. This simple change reduces the proportion of words with no neighbors to all words in the lexicon from over 58 percent to just 7 percent, increasing the applicability of the Neighborhood Activation Model (Luce & Pisoni, 1998). Lexical decision, naming, and familiarity judgment data indicate that speaker/hearers are sensitive to the more distant neighbors brought in by the new neighborhood definition. The large-scale properties of the network are discussed and future directions are indicated.

Introduction

The perceptual similarity relations of spoken words have attracted the attention of researchers since the seminal study by Greenberg and Jenkins (1964). Examination of sound similarity is fundamental for studying the structure of the mental lexicon and the nature of the linguistic system.

For instance, in the debate on whether rules are required to explain linguistic productivity, proponents of the rule-based models have claimed that in each inflectional domain there is one default, rule-based morpheme that is more productive than its competitors and bears little regard to how similar a novel word is to existing ones (Pinker & Prince, 1988). Kapatsinski (2005a, b, in press a) has demonstrated that there are dissociations between productivity and sensitivity to similarity effects, a conclusion that crucially depends on a psychologically real measure of phonological similarity.

A psychologically real measure of phonological similarity is also necessary for creating successful models of analogical extension of linguistic patterns, as in morphological productivity or the lexical diffusion of sound change. For instance, Albright and Hayes (2003) have shown that a model that weighs mismatches in segments adjacent to the affix whose behavior is to be predicted more heavily than more distant mismatches outperforms a model that weighs mismatches in all positions equally in predicting the past tenses of novel verbs produced by native English speakers.

Phonological similarity interacts with word recognition. For instance, Marslen-Wilson's (1990) Cohort Model of word recognition predicts that initial mismatches should lead to more between-word inhibition than later mismatches, as found by Radeau et al. (1995).

Finally, phonological similarity has been argued to influence the direction of sound change and phonological alternations. Steriade (2001) analyzed regressive and progressive place assimilation in CC clusters. She hypothesized that if /anpa/ is perceived to be more similar to /ampa/ than to /anta/, it will be realized as /ampa/. Fujimura et al. (1978) found that when a pause is preceded by transitions

indicating one consonant and followed by transitions indicating another, listeners make the judgment of what the consonant is based on the CV transition, not the VC one in both English and Japanese. On the other hand, retroflection is primarily determined by VC transitions (Ladefoged & Maddieson 1986). Thus, assimilation in retroflection should affect C_2. This is precisely what has been found by Steriade (2001) in a cross-linguistic study. Steriade (2004) argued that phonological similarity can also influence loanword adaptation. For instance, in deciding to simplify a CVC_1C_2 input as CVC_2 the speaker judges C_1 to be less perceptible than C_2 and thus that C_1C_2 is more similar to C_2 than to C_1, and that C_1C_2 is more similar to C_2 than to C_1VC_2.

Recently, researchers have begun to explore the large-scale structure of the phonological mental lexicon using graph-theoretic tools. Vitevitch (2004) and Gruenenfelder and Pisoni (2005) modeled the lexicon as a network in which nodes are words and where two words are connected to each other if they differ only by the addition, deletion or substitution of one segment. This "one phoneme deletion, addition, substitution metric" has been the standard criterion used to determine whether or not two words are lexical neighbors (Luce & Pisoni, 1998).

Unfortunately, this metric has limited applicability since even among the 20000 most common English words more than half have no neighbors (Gruenenfelder & Pisoni, 2005). In addition, the metric contradicts results of confusability studies that have found that the confusability of two sounds depends on the *proportion* of the total duration that is mismatched and not on the absolute duration of the mismatched parts (Fallon Coble & Robinson, 1992; Kidd & Watson, 1992). It is also at odds with the finding that judged sound similarity of two words depends on how many segments they share as well as on how many segments they differ by (Kapatsinski, 2005b, in press b).

The aim of the present study was to examine the large-scale structure of the mental lexicon using a more psychologically plausible definition of neighbors based on the "proportion-of-total-duration rule" derived from confusability studies (Kidd & Watson, 1992). We define a word B to be a neighbor of word A if and only if it shares at least two thirds of A's segments. That is, if A is six segments long its neighbors can be derived from it by at most two phoneme changes (deletions additions, or substitutions), while if A is nine segments long its neighbors can differ from it by at most three segments. Under this metric, the proportion of hermit words (i.e., words with no neighbors) decreased from 58% of the lexicon to 7%. We show that the new metric outperforms the old metric in modeling reaction times and accuracy in the lexical decision and visual word naming tasks as well as in predicting familiarity judgments. Finally, we discuss the present limitations of the metric and ways to further improve and test it.

Another aim of this paper is to help resolve the debate on whether the mental lexicon has scale-free structure. Vitevitch (2004) has claimed that the histogram of number of neighbors per word (i.e., the degree distribution) follows a power law and thus, the lexicon is a scale-free network. Gruenenfelder and Pisoni (2005) have argued that this result is simply due to the relationship between length and the number of neighbors, which is an artifact of the one-phoneme deletion, addition, substitution metric. They have examined the set of monosyllabic words and found that the degree distribution did not follow a power law. In fact, they argued that it resembles much more a Poisson distribution. By eliminating the length bias with the new metric, we show that longer words still tend to have fewer neighbors. In addition, through fitting a number of curves to the data, we find that the best fit to the lexicon's degree distribution is provided by an exponential equation rather than a power law. While the power law accounts for 83% of the variance (85% under the old metric, Gruenenfelder & Pisoni, 2005), the exponential distribution accounts for 97% of the variance.

Finally, we will argue that neighborhood density should not be modeled as the degree of a word, i.e., the number of links connecting the word to other words, but rather as the sum of strengths of those links.

Methods

In this paper, we will be modeling the lexicon as a network in which words are nodes. A link is drawn from node A to node B if at least 2/3 of the segments that occur in the word represented by A also occur in the word represented by B.

The database analyzed was the Hoosier Mental Lexicon (Nusbaum et al., 1984). The phonologically transcribed form of each word in the lexicon was subjected to the new metric. The resulting set of nodes and links, which excluded nodes that had no links, was analyzed using Pajek (Batagelj & Mrvar, 2003). Random networks used for comparison with the actual networks were created using the Erdos-Renyi method in Pajek and had the same number of nodes and links as the actual networks. They differed from the actual networks in that nodes were connected randomly rather than based on the similarity metric.

The network created was "directed" because a long word can have a short word as a neighbor without the short word having the long word as a neighbor. For instance, 'moat' is a neighbor of 'demote' but 'demote' is not a neighbor or 'moat'. This is because 'demote' differs from 'moat' by two segments, which is 1/3 of the duration of 'demote' but more than 1/3 of the duration of 'moat.' The reason the ratio of 1/3 was used is because the traditional metric has performed well in predicting reaction times and familiarity ratings in experiments that mostly used CVC words, which consist of 3 segments (Luce & Pisoni, 1998). Ideally, the **ratio** of mismatch to total length should be derived empirically for each task by seeing which ratio provides the best fit to the data, which is a direction for future research.

For word lengths that are not divisible by three, rounding was used. Thus the maximum number of segments that neighbors could differ by was 1 for 2-segment, 3-segment, and 4-segment words, 2 for 5-, 6-, and 7-segment words, and 3 for 8-, 9-, and 10-segment words. Words that were longer than 10 segments (n=955) or shorter than 2 segments (n=5) were excluded from being heads of links. That is, they could be neighbors of other words but could not have neighbors themselves or, in graph-theoretic terms, all such words have an **output degree** (our operational definition of **neighborhood density**) of 0 and thus are only included in the network if they are pointed to by other words. We will call the word for which we are searching for neighbors the **base word** from now on**.** The reason we are using the output degree, that is, number of links pointing from the base word to other words, rather than input degree (i.e., number of links pointing to the base word) as a measure of neighborhood density is that we take neighborhood density of a word to be the number of words activated by the base word when the base word is presented. That is, we take neighborhood density to be a postlexical variable, which is in agreement with findings that neighborhood density starts to affect processing at a stage indexed by a later ERP component than sublexical variables, such as phonotactic probability (Pylkkanen et al., 2002). The total number of base words (size calculated using MonoConc Pro) was 18360.

Figure 1 shows the unfortunate side effect of rounding up the neighborhood radius for words that are not divisible by three. There is a large jump in average degree as the radius increases from 1 to 2 and then from 2 to 3 segments. There is no evidence that these bumps in the distribution are psychologically real.

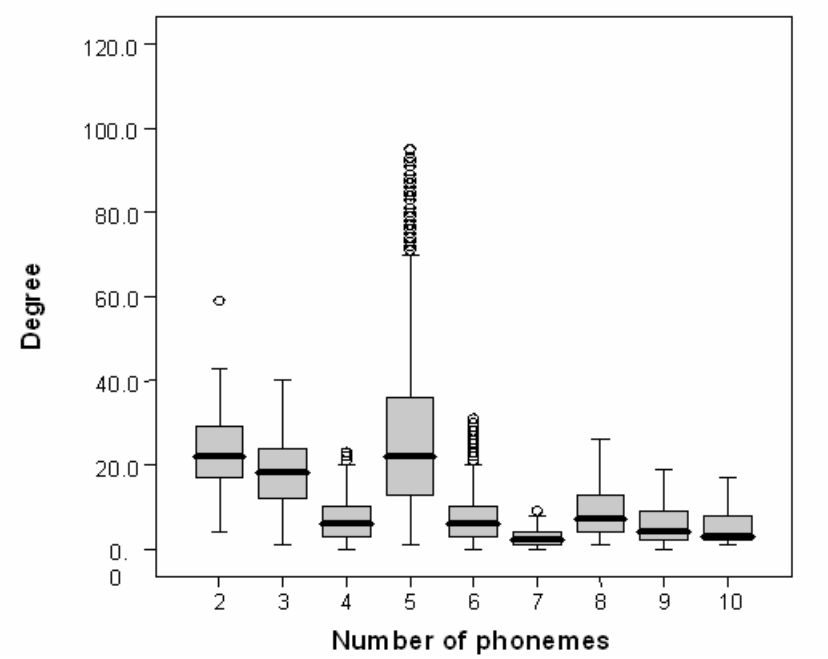


Figure 1: The relationship between base word length in segments and output degree (neighborhood density).[2]

One way to deal with this issue would be to make sure that the number of links between two words always represents the ratio of shared segments to the length of the base word. That means that the maximum number of links between two words would have to be divisible by all possible numbers of segments a base word may have, or, in our case, 2-10. This turns out to be 2520 with two-segment base words being connected to their two-segment neighbors by 1260 links pointing in each direction and ten-segment words being connected to neighbors that differ from them by 3 segments by 252 links. Clearly, this solution is extremely computationally expensive.

The alternative is to use link weights instead of numbers of links to express connection strength. With this approach, the degree of a node does not reflect how strongly it is connected to other nodes. Thus, degree stops being a theoretically justified predictor of behavioral and electrophysiological data. Rather, behavioral and electrophysiological dependent variables should be influenced by the sum of strengths of all links the node has. Future work should examine the characteristics of the link-strength-sum distribution in addition to the degree distribution for the lexicon.

We have used the English Lexicon Project (Balota et al., 2002), a repository of reaction times from the lexical decision task and the naming task for over 40,000 words collected from 1200 subjects, each of whom responded to all the words. The overlap between the 18,360 base words used for network creation and the English Lexicon Project consisted of 13,458 words. These are the words for which we have neighborhood density estimates as well as lexical decision and naming reaction times and familiarity ratings.

[2] The middle line on the rectangle indicates mean output degree, rectangle sides index the 25th and 75th percentile. Vertical lines indicate cases within 3 lengths of box length from the upper or lower edge of the box. Points indicate cases with values more than three box lengths removed from the nearest edge of the box.

Perhaps the biggest drawback of the English Lexicon Project for testing the influence of neighborhood density is that the words were presented to subjects visually. Facilitatory density effects are usually found for visually presented words (Andrews, 1997) while inhibitory effects are found for auditorily presented words (Luce et al., 2000). It is likely that the facilitatory effects found with orthographic presentation are sublexical in nature since high-density words contain high-frequency grapheme chunks whose high frequency can facilitate the orthography-to-phonology mapping (cf. Plaut et al., 1996; Pylkkanen et al., 2002). Therefore, while the new metric is shown to be better at predicting reaction times from the English Lexicon Project, this result should be taken with caution since the neighborhood density effect is facilitatory in this database. A definitive test would come from ERP studies where sublexical and lexical effects can be disambiguated and studies using auditorily presented stimuli, which eliminate the extra processing stage involved in converting orthographic representations into phonological ones.

Results

Small-world Characteristics

A network is considered to have small-world characteristics if it has short average path length and diameter and a clustering coefficient that is orders of magnitude higher than that of a random network with the same number of links and nodes (Watts & Strogatz, 1998). The clustering coefficient CC1 is defined as the proportion of a node's neighbors that are also neighbors of each other. CC2 is the proportion that links between neighbors of a word form out of all links the word's neighbors have. Table 1 shows that the lexicon is characterized by very high clustering but also by relatively long average path length and diameter. That is, like a small-world network, the lexicon contains neighborhoods in which all words are densely interconnected and between which the connectivity is lower. However, the between-neighborhood links that would allow access from a node in one neighborhood to a node in another neighborhood are harder to find than in a small-world network. Importantly, high clustering does not depend on the inclusion of morphologically complex words, although the exclusion of morphologically complex words does decrease the lexicon's clustering coefficient relative to a random network. Thus, morphology increases clustering but is not exclusively responsible for it. Interestingly, mean path length for the entire lexicon does not decrease noticeably from its value with the old metric (6.08 in Vitevitch 2004, 6.06 here), despite a large increase in the network's size.

Table 1. Small-world properties (actual data compared to random nets with the same number of nodes and links/node created using the Erdos-Renyi method in Pajek)

	Entire lexicon		**2-4 phoneme words**		**5-7 phoneme words**		**8-10 phoneme words**		**Monomorphemic words**	
	Real	**Rand.**	**Real**	**Rand.**	**Real**	**Rand.**	**Real**	**Rand.**	**Real**	**Rand.**
Average distance	6.06	4.30	4.86	4.08	5.46	4.74	7.79	6.24	5.27	3.88
Diameter	20	7	15	7	16	8	26	13	21	7
CC1	.235	.0007	.262	.002	.208	.0006	.173	.0006	.253	.0016
CC2	.040	.0005	.039	.001	.030	.0005	.047	.0005	.040	.0009

Figure 2 shows the entire lexicon. As we can see, the lexicon is a disconnected graph, which has a giant component comprising the vast majority of words in the lexicon. The visualization is derived using the Fruchterman-Reingold algorithm, which separates the giant component from the rest of the network and arranges the nodes in such a way that distance in terms of number of links corresponds to physical distance on the graph.

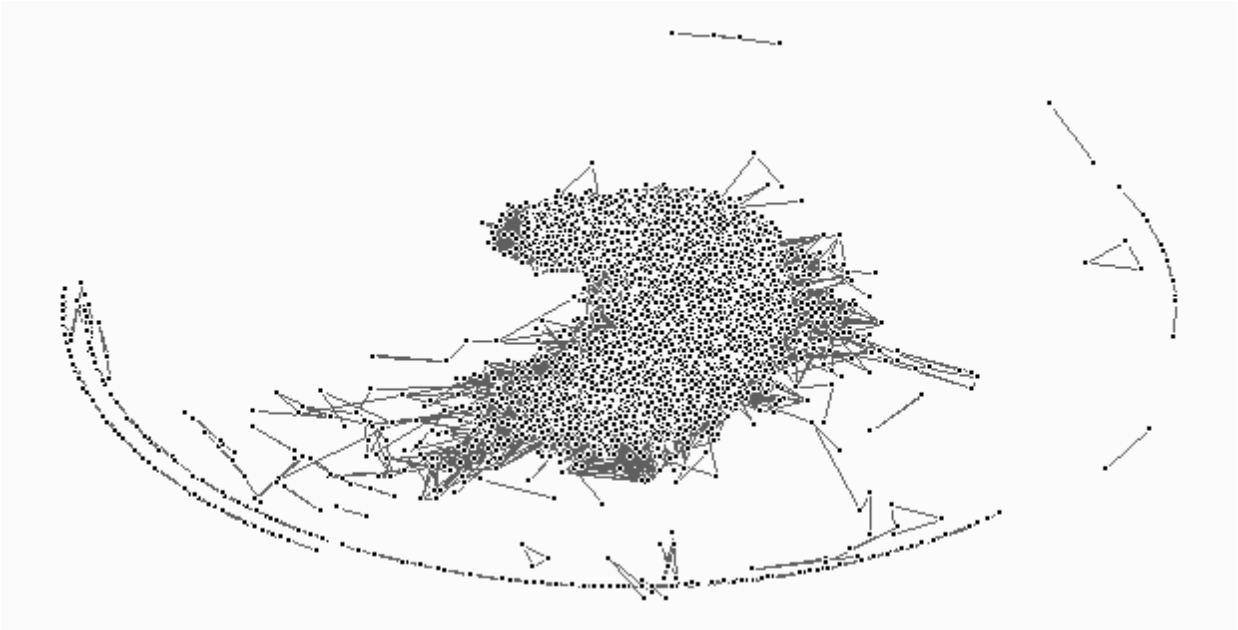


Figure 2. The complete lexicon. Dots are words, lines are connections. Two dots are connected if they share at least 2/3 of their segments.

The giant component is elongated, showing that some nodes can only reach each other through a long chain of intermediaries, thus leading to a relatively long mean path length. Interestingly, this elongation is obtained only if all words are included. It is not obtained with only monomorphemic words, or with words that are 2-4, 5-7, or 8-10 phonemes long. Thus, the elongation seems to result from the fact that the longest words in the complete lexicon cannot be connected to the shortest words: words of particular lengths form distinct connected regions of the giant component, such that the length of the words increases as one proceeds from the top of the giant component to its bottom. This hypothesis is confirmed by the finding that mean path length is larger relative to the corresponding random net for the entire lexicon than for any of its word-length-limited subsets (1.41 vs. 1.2, 1.15, 1.25).

In fact, the lexicon tends to fall apart into separate clusters when nodes with low degrees are eliminated. When nodes with degrees below 15 are eliminated, there is a noticeable bottleneck separating long words (more than 7 phonemes long) from shorter words. This bottleneck, which connects long words to short words, is formed entirely of /ʃən/-final words. Of the 2228 nodes with degrees above 25, the elimination of 4 nodes - ‘coalition’ (degree=40) or ‘colon’ (degree=43), ‘passion’ (degree=41), ‘nation’ (degree=39), and ‘fixation’ (degree=40) or ‘fission’ (degree=45) - would render the network disconnected. Of the six, only 2 (‘passion’ and ‘nation’) have more than one link to each mega-neighborhood. Thus, in this network, it appears that if one wants to render the network

disconnected, it is not the biggest hubs (like 'pastor' with degree=112) that need to be taken out. Figure 3 shows a visualization of the lexicon with all nodes with output degrees below 40, i.e., all words with fewer than 40 neighbors, eliminated. The figure shows that the lexicon 'falls apart', showing that high-degree nodes fall into two sparsely-connected giant neighborhoods. The lower neighborhood comprises 8-10 segment words while the upper component contains shorter words with 24 segment words confined to the left side of the upper neighborhood.

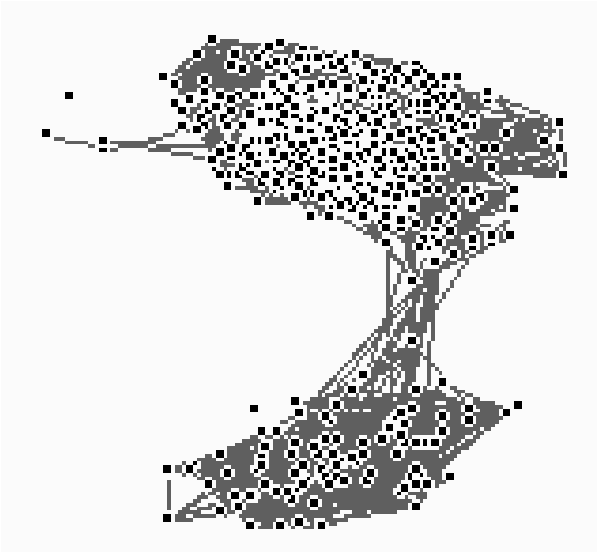


Figure 3. The network comprising all words with more than 40 neighbors.

Table 2 shows how the various reductions of the lexicon compare to the corresponding random networks. The table shows that as nodes with lower output degrees are introduced to the network, the network does not become less connected. That is, the low degree nodes do not just attach themselves to the outskirts of the network but also form shortcuts between high-degree nodes. Interestingly as nodes with lower degree are introduced, the average degree grows (until minimum degree reaches around 20), which indicates that the lower-degree nodes connect to many of the high-degree nodes, rather than forming long chains of low-degree nodes that must be traversed to reach a high-degree node from a randomly chosen low-degree node. In some cases, this even increases the connectivity of the network, e.g., when nodes with degrees between 35 and 39 are introduced, mean path length remains constant and diameter shrinks despite an increase in the size of the network. The same occurs when nodes with degrees between 15 and 19 are introduced.

In a small-world network, the growth of diameter and mean path length with the introduction of low-degree nodes would likely be steeper than in a random net because in such a network high-degree nodes are the ones that are more likely to have connections linking different neighborhoods. Traveling from a randomly chosen node down a randomly chosen link one is more likely to end up in a high-degree node than in a low-degree node. If the high-degree nodes are also more likely to provide a link to another neighborhood, the average path length between the neighborhoods is shortened. It makes functional sense for the lexicon to consist of poorly connected neighborhoods, since one would not want spreading activation to easily activate or inhibit neighbors of the stimulus's neighbors that are not neighbors of the stimulus and are therefore not at all similar to the stimulus.

Table 2. Elimination of low-degree nodes does not lead to decrease in average path length.

Minimum output degree	Mean path length relative to random net	Diameter relative to random net	Mean degree
40[3]	3.86/2.68=**1.44**	13/4=**3.3**	**16.21**
35	3.86/2.77=**1.39**	10/4=**2.5**	**16.68**
30	3.91/2.89=**1.35**	11/5=**2.2**	**16.98**
25	4.28/2.95=**1.45**	14/4=**3.5**	**17.73**
20	4.80/3.05=**1.57**	16/5=**3.2**	**18.04**
15	4.80/3.21=**1.50**	14/5=**2.8**	**18.02**
10	5.02/3.43=**1.46**	20/5=**4.0**	**17.38**
5	5.33/3.72=**1.43**	17/6=**2.8**	**15.22**
2	5.78/4.04=**1.43**	19/7=**2.7**	**12.77**
1	5.98/4.23=**1.41**	20/7=**2.9**	**11.73**
0	6.06/4.30=**1.41**	20/7=**2.9**	**11.33**

Low mean path lengths occur in networks which need to be traversed quickly. One purpose of such traversal is search. In the lexicon, on the other hand, search is unlikely to occur by traversing links between distantly located nodes. Rather, search in the lexicon involves activation of structured neighborhoods of words that all share a single sublexical chunk, the chunk consistent with the acoustic evidence at that point in word recognition (Marslen-Wilson, 1990). Below we will see that structured neighborhoods are a major feature of the lexicon.

Degree Distribution

Figures 4-8 show the degree distributions for the entire lexicon and its subsets modeled as a network in which two words are connected by only one link regardless of the strength of the connection. The trendlines, fitted in Microsoft Excel, show that the exponential distribution provides a much better fit to the data than the power-law-based one.

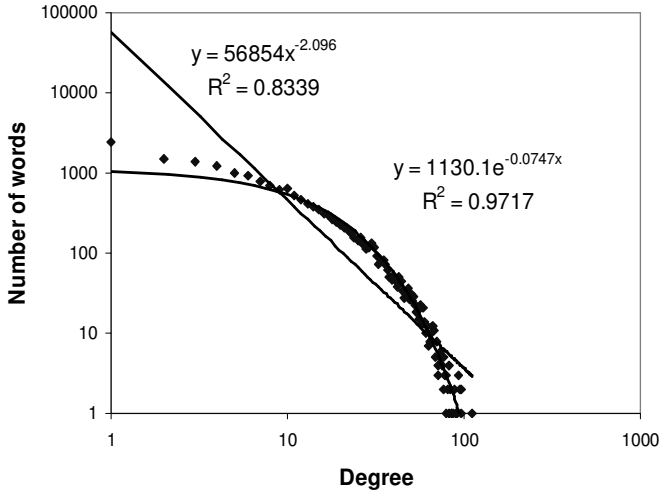


Figure 4. Degree distribution for the entire lexicon modeled with a power law and an exponential distribution.

[3] This was the highest degree for which mean path length and diameter measures were collected because the network no longer has a giant component at higher degrees.

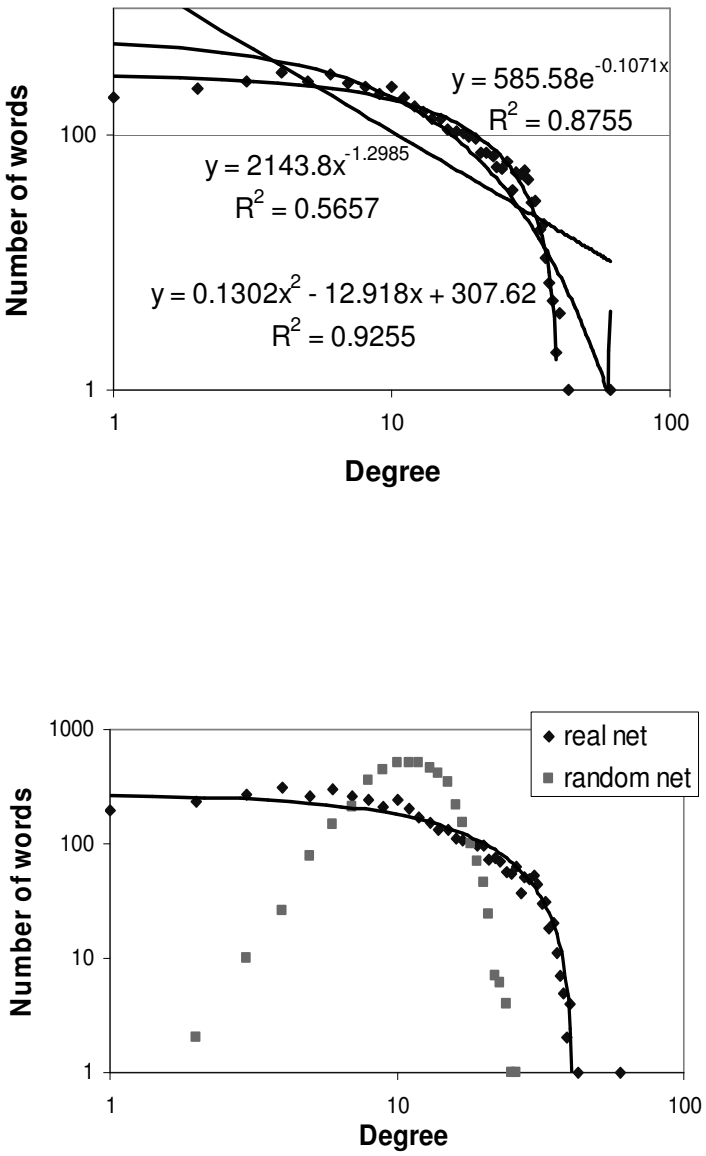


Figure 5. Degree distribution for words 2, 3, and 4 segments long with power law, exponential and parabolic models and compared to the Poisson degree distribution of a random network.

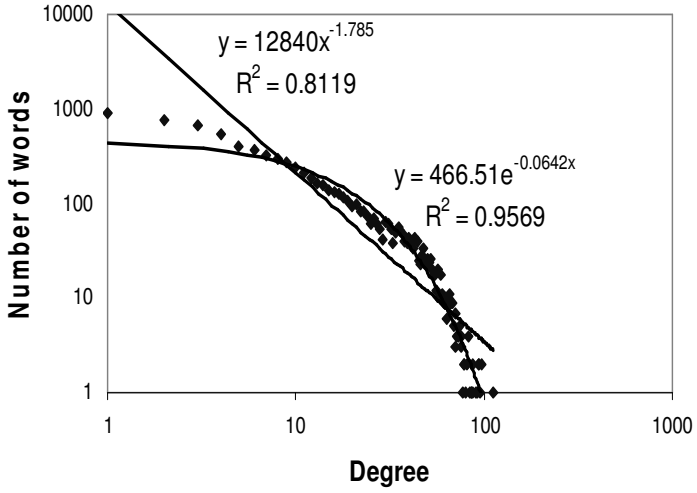


Figure 6. Degree distribution for words 5, 6, and 7 segments long with power law and exponential models.[4]

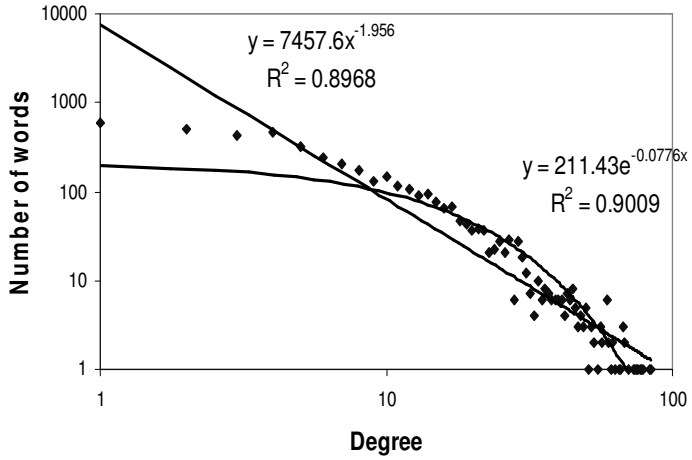


Figure 7. Degree distribution for words 8, 9, and 10 segments long with power law and exponential models.

[4] The reason the graphs do not show a random net distribution is that there is no lowering in the left-hand tail of the observed distribution, making a Poisson fit highly inappropriate.

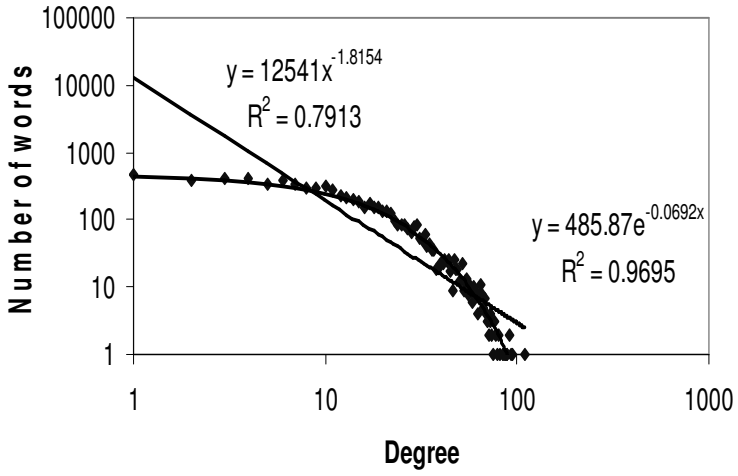


Figure 8. Degree distribution for the monomorphemic lexicon modeled with a power law and an exponential distribution.

To summarize, the degree distributions for the entire lexicon, the monomorphemic lexicon, and the set of words that are 5, 6, or 7 segments long are best approximated by an exponential curve, rather than a power law. The degree distribution of words that are 2, 3, or 4 segments long is best approximated by a parabola and not a Poisson distribution. The degree distribution of 8-, 9-, and 10-segment words is approximated equally well by an exponential curve and a power law. Where the curves diverge most is when predicting the number of words with very low degrees: the power law systematically overpredicts while the exponential distribution underpredicts. If the data for the entire lexicon are fit to a power law, the exponent is 2.1.

Perhaps, the most important feature of the new metric, however, is that the marked reduction achieved in the number of words with no neighbors (hermits), shown in Table 3.

Table 3. Number of hermits under old and new metrics.

Length of word	Percent hermits under old metric	Percent hermits under new metric
2-4 segments	2.4%	2.4%
5-7 segments	67.6%	8.6%
8-10 segments	92.7%	10.4%
Whole lexicon	58.1%	7.3%
Monomorphemic lexicon	33.2%	7.0%

Relations between Old Density, New Density, and Other Independent Variables[5]

Table 4 shows that new density correlates with other independent variables more highly than does old density. Thus to show that new density is a better predictor of behavior than old density it will not be sufficient to show that new density shows better correlation with behavioral dependent variables. Rather, we will need to show that it does better even when the other partially correlated variables are competing against density in a regression. Old density correlates with new density at r=.623. We are going to concentrate on modeling reactions to monomorphemic words to avoid confounding phonological and morphological links.

Table 4. Correlations between old density, new density, and other lexical variables for all monomorphemic words that are longer than four phonemes (n=4146).

		Number of syllables	Number of letters	Number of phonemes	Mean phoneme frequency	Mean bigraph frequency	Log word frequency
Old density	Pearson r	**-.373**	**-.339**	**-.387**	**-.029**	**.069**	**.103**
	Sig. (2-tailed)	.000	.000	.000	.061	.000	.000
New density	Pearson r	**-.490**	**-.429**	**-.499**	**.118**	**.115**	**.150**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000

Modeling Behavioral Data

Table 5 shows that new density demonstrates stronger correlations with behavioral variables than does old density.

Table 5. Old density vs. new density as correlates of behavioral variables for all monomorphemic words that are longer than four phonemes. The reason shorter words were excluded is because the new metric and the old metric make identical predictions for those words.

		Naming accuracy	LDT accuracy	Naming RT	LDT RT	Familiarity
Old density	Pearson r	**.142**	**.114**	**-.245**	**-.204**	**.136**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
New density	Pearson r	**.181**	**.151**	**-.310**	**-.273**	**.182**
	Sig. (2-tailed)	.000	.000	.000	.000	.000

Table 6 shows that when old density and the difference between old and new density are entered into a regression, the difference between old and new density makes a significant contribution

[5] All results reported in this section and the next one were derived with the network in which all relations are represented by a single link.

to predicting behavior. Thus subjects are sensitive to the word's distant neighbors brought in by the new metric but not by the old metric. Another interesting aspect of these data is that the clustering coefficient CC2 and hubness, which are measures that take into account characteristics of the neighborhoods of the stimulus's neighbors, are significant only when multimorphemic words are included. That suggests that words that are neighbors of neighbors of the stimulus play a role in the stimulus's processing only to the extent that they share morphemes with the stimulus or are neighbors of the stimulus's root.

Table 6. Old density vs. new density as predictors of naming and lexical decision accuracy and reaction time and familiarity judgments. The top row in each cell shows results for all word that are more than four phonemes long (n=10732). The bottom row shows results for monomorphemic words that are more than four phonemes long (n==4146). Significant effects are in bold.

Predictor	**Naming accuracy**		**Lexical decision accuracy**		**Naming time**		**Lexical decision time**		**Familiarity rating**	
	Coeff.	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**
Old density	.018	.207	.020	.123	-.021	.081	-.005	.689	.017	.145
	.034	.063	**.033**	**.050**	**-.031**	**.048**	-.026	.103	.023	.121
New density minus old	**.030**	**.021**	**.051**	**.000**	**-.038**	**.001**	**-.050**	**.000**	**.046**	**.000**
	.082	**.000**	**.072**	**.000**	**-.093**	**.000**	**-.076**	**.000**	**.073**	**.000**
Log word frequency	**.424**	**.000**	**.585**	**.000**	**-.472**	**.000**	**-.541**	**.000**	**.595**	**.000**
	.476	**.000**	**.625**	**.000**	**-.499**	**.000**	**-.567**	**.000**	**.645**	**.000**
Mean phoneme frequency	**-.051**	**.000**	**-.033**	**.001**	**.114**	**.000**	**.042**	**.000**	**-.051**	**.000**
	-.061	**.000**	**-.034**	**.023**	**.127**	**.000**	**.043**	**.003**	**-.057**	**.000**
Mean bigraph frequency	-.004	.693	-.002	.834	-.006	.465	**.018**	**.034**	.012	.163
	-.019	.244	.009	.551	-.023	.091	.015	.287	**.032**	**.012**
Number of phonemes	.023	.276	.022	.248	**.073**	**.000**	**.057**	**.001**	-.012	.497
	.090	**.001**	**.053**	**.036**	.008	.731	.040	.098	.003	.891
Number of letters	**.095**	**.000**	**.175**	**.000**	**.113**	**.000**	**.150**	**.000**	**.154**	**.000**
	.029	.236	**.083**	**.000**	**.180**	**.000**	**.093**	**.000**	**.127**	**.000**
Number of syllables	**-.319**	**.000**	**-.154**	**.000**	**.257**	**.000**	**.169**	**.000**	**-.153**	**.000**
	-.162	**.000**	**-.089**	**.000**	**.133**	**.000**	**.123**	**.000**	**-.114**	**.000**
Number of morphemes	**.137**	**.000**	**.139**	**.000**	**-.102**	**.000**	**-.050**	**.000**	**.169**	**.000**
CC2[6] (new)	**.045**	**.000**	**.027**	**.006**	-.015	.093	-.017	.055	.026	.004
	.027	.113	.025	.107	-.015	.296	-.003	.817	**.005**	**.721**
Hubness (new)	-.004	.717	**-.029**	**.003**	**.025**	**.006**	**.022**	**.010**	**-.040**	**.000**
	.011	.447	.000	.997	-.007	.587	.011	.374	.003	.824
r^2	.227		.354		.438		.477		.383	
	.272		.405		.457		.465		.444	

Table 7 shows that new density can make a significant contribution to successfully predicting subjects' reactions to the words that the old metric considers to be hermits. Table 8 shows that the correlations between new density and subjects' behavior in response to 'hermits' are significant. These results suggest that, contrary to the old metric, these words do not all have the same density. It is important to point out that the old metric cannot predict differences in behavior in response to different hermits even if neighbors are defined as words that are 1 or 2 links away from the stimulus (Gruenenfelder & Pisoni, 2005).

[6] The other clustering coefficient, CC1, is not a significant predictor for any dependent variable.

Table 7. Predicting reactions to former hermits. The top row in each cell shows results for all hermits that are more than four phonemes long (n=7795). The bottom row shows results for monomorphemic words that are more than four phonemes long (n==2054). Significant effects are in bold.

Predictor	Naming accuracy		Lexical decision accuracy		Naming time		Lexical decision time		Familiarity rating	
	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.	Coeff.	Sig.
New density	.009	.520	**.036**	**.005**	**-.030**	**.015**	**-.028**	**.017**	**.033**	**.004**
	.058	**.011**	**.070**	**.001**	**-.076**	**.000**	**-.052**	**.009**	**.080**	**.000**
Log word frequency	**.430**	**.000**	**.582**	**.000**	**-.495**	**.000**	**-.548**	**.000**	**.586**	**.000**
	.507	**.000**	**.650**	**.000**	**-.540**	**.000**	**-.570**	**.000**	**.654**	**.000**
Mean phoneme frequency	**-.059**	**.000**	**-.044**	**.000**	**.116**	**.000**	**.050**	**.000**	**-.065**	**.000**
	-.060	**.005**	**-.042**	**.027**	**.105**	**.000**	**.048**	**.011**	**-.065**	**.000**
Mean bigraph frequency	.008	.509	.014	.202	-.007	.492	.015	.150	**.031**	**.002**
	-.006	.757	.032	.080	-.019	.303	.011	.559	**.058**	**.000**
Number of phonemes	.026	.265	.013	.528	**.060**	**.003**	**.040**	**.038**	-.010	.592
	.099	**.007**	**.065**	**.046**	-.022	.487	.009	.782	.022	.465
Number of letters	**.085**	**.000**	**.162**	**.000**	**.094**	**.000**	**.170**	**.000**	**.138**	**.000**
	.025	.435	**.066**	**.019**	**.177**	**.000**	**.130**	**.000**	**.111**	**.000**
Number of syllables	**-.277**	**.000**	**-.136**	**.000**	**.261**	**.000**	**.160**	**.000**	**-.145**	**.000**
	-.147	**.000**	**-.060**	**.013**	**.147**	**.000**	**.118**	**.000**	**-.101**	**.000**
Number of morphemes	**.139**	**.000**	**.148**	**.000**	**-.106**	**.000**	**-.059**	**.000**	**.179**	**.000**
CC2 (new)	**.055**	**.000**	**.031**	**.007**	**-.024**	**.028**	**-.029**	**.006**	**.030**	**.005**
	.019	.381	.029	.123	-.032	.087	-.011	.541	-.003	.880
Hubness (new)	-.001	.911	**-.037**	**.002**	**.033**	**.005**	**.026**	**.019**	**-.040**	**.000**
	.009	.656	-.005	.768	-.004	.826	.010	.547	-.010	.532
r^2	.271		.404		.456		.464		.443	
	.277		.426		.453		.443		.444	

As seen from Table 8, new density is equally good at predicting behavior to former hermits as to former non-hermits in terms of lexical decision accuracy and judgments of familiarity but is slightly less successful on the hermits with the other dependent variables. Notably, the new density measure shows a better correlation with behavioral measures even when words considered hermits by the old metric are eliminated from comparison.

Table 8. Correlations between new density and behavioral variables for morphologically simple words considered hermits by the old density metric (top row, n=2054) and those words considered non-hermits by the old metric (bottom row, n=1628).

		Naming accuracy	LDT accuracy	Naming RT	LDT RT	familiarity
New density	Pearson Correlation	.107	.107	-.225	-.187	.136
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	Pearson Correlation	.162	.105	-.238	-.209	.135
	Sig. (2-tailed)	.000	.000	.000	.000	.000
Old density	Pearson Correlation	.090	.049	-.151	-.096	.075
	Sig. (2-tailed)	.001	.067	.000	.000	.002

Discussion

In this paper, we have proposed a definition of neighborhood that incorporates findings from confusability (Kidd & Watson, 1992) and sound similarity judgment data (Kapatsinski, 2005b, in press b) in that two words are defined to be neighbors if they share a certain proportion of their length, and not some absolute number of segments. We have seen that even under this metric longer words are more likely to have fewer neighbors but that the proportion of words that have no neighbors decreases dramatically. We have also seen that the lexicon has an exponential degree distribution but for very long words a scale-free distribution is indistinguishable from an exponential one in terms of goodness of fit to the data while the distribution for very short words is roughly parabolic due mainly to differences in numbers of words that have very few neighbors. Thus, lexicon structure is not consistent with growth via preferential attachment (Barabasi & Albert, 1999) under the existing neighborhood density metrics.

We have also seen that the lexicon is highly clustered and has a higher mean path length than a random network. Thus, the lexicon is not a small-world network. One factor contributing to relatively high average path length is that while in a random network a node can link to any other node, in the lexicon a word can only be linked to another word with which it shares at least 2/3 of its segments. Therefore, some nodes are guaranteed not to be directly connected.

We have also seen evidence that the more distant neighbors have an effect on how fast the subject can recognize a word. In what follows, I will touch upon the reasons for the high clustering of the lexicon and discuss some limitations of the present metric.

High clustering results naturally if words tend to be similar to many neighbors in the same way. That is, if a word shares some part with many of its neighbors, then the neighbors automatically share the part and are likely to be neighbors with each other. Thus the high clustering of the lexicon results from certain suprasegmental parts being reused more often than others. That is, the lexicon consists of a large number of mega-neighborhoods or "gangs" (Bybee & Moder, 1983), each of which is characterized by what Albright and Hayes (2003) called "structured similarity." One such gang is the gang of words ending in –ter, which includes the highest-degree words 'pastor,' 'canter,' 'caster,' 'master,' etc. Another large gang consists of words starting with str-. The vast majority of long, 8-10-segment words belong to the gang of words ending in -tion, which is subdivided into words ending in -ation and those ending in -ition. If the neighborhood is highly structured, the clustering coefficient CC1 is high because neighbors of the stimulus are very likely to be neighbors of each other: they all share the same parts.

One could imagine a lexicon in which there would not be gangs, as all parts of a word would be equally likely to occur in another word. However, the process of chunking ensures the emergence of these larger units: segments that are used together fuse together through Hebbian learning (Bybee, 2002). Product-oriented generalizations of the type "words (that mean/are X) have Y" are formed (Burzio, 2002; Bybee, 1995) making new words that conform to the generalizations more learnable and leading to analogical change in old words that do not conform to the generalizations. Thus, a rich-get-richer phenomenon occurs at the sublexical level: frequently used units are likely to be used even more frequently in the future.

If no phonotactic constraints on the shape of possible words existed, then under the single-phoneme deletion/addition substitution metric, a word consisting of n segments in a language that has k

phonemes could be neighbor to a maximum of n words that are shorter than it by 1 segment, n*k words that are of the same length, and (n+1)*k words that are longer than it by 1 segment. Therefore the old metric would make the prediction that short words should have more neighbors than long words since a short word can link to more long words than a long word can link to short words and the space of possible words is more sparsely populated at the longer word lengths.

As shown in Table 9, the number of neighbors two phonemes away depends on the length of the word more than does the number of neighbors one phoneme away. Table 9 shows that when neighborhood radius, in terms of number of phonemes, is kept constant, number of neighbors under the new metric is even more sensitive to word length than number of neighbors under the old metric because of including more remote neighbors.

Table 9. Correlations between neighborhood density and length for old and new metrics when neighborhood radius is kept constant.

	Old metric	**New metric**
Words 5-7 segments long (n=3545)	-.38	-.56
Words 8-10 segments long (n=598)	-.10	-.25

The fact that the lexicon falls apart along length boundaries is a testament to the strength of the phonotactic constraints of English, which, for an average word, rule out many more additions than substitutions. It is not predicted by the above formulas since the maximum number of additions is slightly greater than the maximum number of substitutions.

There are certain limitations of our metric of lexical/phonological similarity. A fundamental limitation is that the raw number of links is used to predict reaction times and familiarity judgments, rather than the sum of the strengths of the links.

Several ways in which links could be weighted are apparent. One is that the weight of a link should reflect the proportion of the word's segments that are mismatched.[7] Vitz and Winkler (1973) have found that in their data there is a correlation of .81-.95 (depending on the experiment) between similarity judgments and the proportion of mismatched segments. In addition, we need to take into account how similar substituted segments are to each other and how salient the inserted or deleted segments are, factors shown to affect similarity and confusability (cf. Kapatsinski, 2005b for review). Furthermore, a mismatch of x phonemes has the same effect on similarity under the present metric regardless of whether the mismatched phonemes are adjacent to each other, while sound similarity judgment data show that discontinuous mismatches are more salient (Kapatsinski, 2005b, in press b), a finding that may have to do with the fact that a discontinuous mismatch is likely to involve several suprasegmental units. The number of mismatched syllabic constituents has been shown to affect sound similarity judgments in addition to the number of mismatched segments (Kapatsinski, 2005b, in press b).

[7] This does not necessarily mean that everything is connected to everything, since no link would be postulated if the ratio of segments shared by two words to the number of segments in the base word is smaller than the allowed minimum ('neighborhood radius').

Finally, changes in all positions in the word are weighted equally at present. There are two reasons why this is problematic. One is that final and initial positions are more salient in sound similarity judgments than medial positions (Kapatsinski, 2005b, in press b) and the end of a stimulus is especially important in determining confusability (Fallon Coble & Robinson, 1992; Kidd & Watson, 1992). The second reason, which is even more serious, is that in phonological priming words that share beginnings inhibit each other while words sharing ends often show facilitation (Radeau et al., 1995). Furthermore, words sharing beginnings appear to inhibit each other in picture naming (Vitevitch et al., 2004) which makes the use of position-insensitive neighborhood definitions an inadequate predictor of reaction times.

In addition, morphologically complex words were included. The results indicate that the inclusion of morphologically complex words leads to finding effects of the clustering coefficient CC2 and hubness. Given the findings that words are often (Hay, 2003) or exclusively (Stockall, 2004) accessed through their roots, the inclusion of morphologically complex words is problematic. Furthermore, it is not clear whether neighbors of all forms of the word can influence the processing of any given wordform or whether neighbors of the root can. That is, whether neighborhood relations are relations between inflected forms, derived bases, or roots.

The data against which the metric has been tested at present are not ideal. Words were presented to subjects visually in the lexical decision task, allowing facilitatory sublexical effects during orthography-to-phonology conversion to be overlaid on the inhibitory phonological neighborhood density effects. Familiarity judgments are a metalinguistic task, which involves postlexical processing.

A promising avenue for modeling behavior using neighborhood density as a predictor has been provided by Pylkkanen et al. (2002) who found two ERP (event-related potential) components on the MEG (magnetoencephalogram), only one of which was sensitive to neighborhood density. An early component occurring at about 170 ms after word presentation was found to be sensitive only to phonotactic probability but not to neighborhood density while a component peaking at 350 ms was sensitive to neighborhood density. Generally, the M350 is thought to index lexical access. Examining neighborhood density effects using MEG and EEG may allow us to investigate the effects of neighborhood density at the first stage of processing that is sensitive to lexical-level variables (M350 is also the first ERP component sensitive to word frequency, Embick et al. 2000). Such an early component is less likely to show strategic effects. Comparing M170 effects to M350 effects also provides a way to deconfound neighborhood density and phonotactic probability/sublexical unit frequency.

Conceptualizing the lexicon as a complex network provides us with a number of variables that may influence the speaker/hearer's processing of the words. Graph theory has provided us with various measures of clustering. One other theoretically promising variable for modeling priming and word recognition is the average degree of the word's neighbors, its 'neighbor density'. A promising network-based variable for modeling confusability is the number of neighbors that two potentially confusable words have in common. Finally, some node pairs provide connections between neighborhoods, while others lie within giant, almost fully interconnected clusters. For instance, two CVC words sharing a rare VC unit and containing different frequent CV units will form one of the few links connecting together two large CV-based neighborhoods and may be especially important for analogical extension of linguistic patterns.

While graph theory provides us with powerful tools for describing networks, it is not, on its own, a theory of the mental lexicon. To provide such a theory, we need a psychologically real similarity

measure. We also need to understand which of the many characteristics of a node humans are sensitive to in a particular task. That relative sensitivity will surely be constrained by as well as provide constraints for our theories of how the lexical network is used in a wide range of behavioral tasks.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/ experimental study. *Cognition*, *90*, 119-61.

Andrews, S. (1997). The role of orthographic similarity in lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, *4*, 439-61

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., & Treiman, R. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. http://elexicon.wustl.edu/, Washington University.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509.

Batagelj, V., & Mrvar, A. (2003). Pajek - Analysis and visualization of large networks. In M. Jünger & P. Mutzel (eds.), *Graph Drawing Software* (pp. 77-103). Berlin: Springer.

Burzio, L. (2002). Missing players: Phonology and the past tense debate. *Lingua*, *112*, 157-199.

Bybee, J.L. (2002). Sequentiality as the basis of constituent structure. In T. Givón and B. Malle (eds.), *The evolution of language out of pre-language* (pp. 107-132). Amsterdam: John Benjamins.

Bybee, J.L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425-455.

Bybee, J.L., & C.L. Moder. (1983). Morphological classes as natural categories. *Language*, *59*, 251-270.

Embick, D., Hackl, M., Schaeffer, J., Kelepir, M., & Marantz, A. (2001). A magnetoencephalographic component whose latency reflects lexical frequency. *Cognitive Brain Research, 10*, 345–348.

Fallon Coble, S., & Robinson, D.E. (1992). Discriminability of bursts of reproducible noise. *Journal of the Acoustical Society of America*, *92*, 2630-2635.

Fujimura, O., Macchi, M., & Streeter, L.A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, *21*, 337-346.

Greenberg, J.H., & Jenkins, J.J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*, 157-177.

Gruenenfelder, T., & Pisoni, D.B. (2005). Modeling the mental lexicon as a complex graph. Ms. Indiana University.

Hay, J. (2003). *Causes and consequences of word structure*. London: Routledge.

Kapatsinski, V.M. (2005a). Characteristics of a rule-based default are dissociable: Evidence against the Dual Mechanism Model. In S. Franks, F. Y. Gladney, & M. Tasseva-Kurktchieva (eds.), *Formal approaches to Slavic linguistics 13: The South Carolina meeting*, 136-146. Ann Arbor, MI: Michigan Slavic Publications.

Kapatsinski, V.M. (2005b). *Productivity of Russian stem extensions: Evidence for and a formalization of network theory*. M.A. Thesis: U of New Mexico.

Kapatsinski, V.M. (In press a). To scheme or to rule: Defining attributes of the dual mechanism default are dissociable. *BLS 31*.

Kapatsinski, V.M. (In press b). Phonological similarity relations: Network organization of the mental lexicon. *Proceedings of VIII Encuentro Internacional de Linguistica en el Noroeste*.

Kidd, G.R., & Watson, C.S. (1992). The "proportion-of-the-total-duration rule" for the discrimination of auditory patterns. *Journal of the Acoustical Society of America*, *92*, 3109-3118.

Ladefoged, P., & Maddieson, I. (1986). Some of the sounds of the world's languages*, UCLA Working Papers in Phonetics, 64*.

Luce, P.A., Goldinger, S.D., Auer, E.T., & Vitevitch, M.S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics*, *62*, 615–625.

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

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In: G.T.M. Altmann (ed.), *Cognitive models of speech processing: psycholinguistic and computational perspectives* (pp.148-173). Cambridge, MA: MIT.

Nusbaum, H.G., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10*. (pp. 357-376). Bloomington, IN: Speech Research Laboratory, Indiana University.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 195-247.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56-115.

Pylkkänen, L., Stringfellow, A., & Marantz, A. (2002). Neuromagnetic evidence for the timing of lexical activation: An MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain and Language*, *81*, 666–678.

Radeau, M., Morais, J., & Segui, J. (1995). Phonological priming between monosyllabic spoken words. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1297-1311.

Steriade, D. (2004). Knowledge of similarity and narrow lexical override. *BLS 29*, 583-598.

Steriade, D. (2001). Directional asymmetries in place assimilation. In E. Hume & K. Johnson (eds.), *The role of speech perception in phonology*. New York: Academic Press.

Stockall, L. (2004). *Magnetoencephalographic investigations of morphological irregularity and identity*. PhD Thesis: MIT.

Vitevitch, M.S. (2004). Phonological neighbors in a small world. Ms. University of Kansas.

Vitevitch, M.S., Armbruster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 514-529.

Vitz, P.C., & Winkler, B.S. (1973). Predicting the judged "similarity of sound" of English words. *Journal of Verbal Learning and Verbal Behavior*, *12*, 373-388.

Watts, D.J., & Strogatz, S. (1998). Collective dynamics of `small-world' networks. *Nature*, *393*, 440-442.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Audiovisual Asynchrony Detection and Speech Perception in Normal-Hearing Listeners and Hearing-Impaired Listeners with Cochlear Implants [1]

Marcia J. Hay McCutcheon,[2] David B. Pisoni[2] and Kristopher K. Hunt[2]

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

[1] This work was supported by NIH-NIDCD Training Grant T32DC00012 to Indiana University and the Psi Iota Xi Sorority. This is a draft of a paper to be submitted for publication.
[2] DeVault Otologic Research Laboratory, Department of Otolaryngology—Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Audiovisual Asynchrony Detection and Speech Perception in Normal-Hearing Listeners and Hearing-Impaired Listeners with Cochlear Implants

Abstract. This study examined the effects of hearing loss and aging on the detection of AV asynchrony in both normal hearing adults and hearing impaired listeners with cochlear implants. Additionally, the relationship between AV asynchrony detection skills and speech perception was assessed. Twenty-five cochlear implant users and 22 normal-hearing adults participated in this study. Both elderly and middle-aged sub-groups of individuals were included in the normal hearing and cochlear implant groups. Individuals were asked to make judgments about the synchrony of AV speech and to complete three speech perception tests, the CUNY, HINT and CNC. No significant differences were observed in the detection of AV asynchronous speech between the normal-hearing listeners and the cochlear implant users. The data revealed, however, that older adults had wider windows over which they identified AV asynchronous speech as being synchronous than younger adults. Additionally, for normal hearing listeners the temporal AV asynchrony processing window was found to be correlated with speech perception measures. Specifically, wider temporal windows were associated with poorer speech perception skills. This trend was not observed in the hearing impaired population. These findings suggest that there may be fundamental differences in how speech is perceived in individuals with cochlear implants.

Introduction

The assessment of speech perception skills in cochlear implant users commonly involves the use of cues provided through one modality, namely, the information acquired through listening-alone. However, in order to fully understand speech perception processes it also is appropriate to evaluate the impact that visual cues have on individual word and sentence recognition abilities. In the normal hearing population, a number of studies have shown that both visual and auditory information play an important role in speech perception. For example, Sumby and Pollack (1954) demonstrated the importance of visual information for speech understanding in the presence of background noise. This study demonstrated that for extremely poor listening environments (i.e., a -30 dB signal-to-noise ratio) a 40% to 80% increase in word recognition can be achieved when visual cues are added to the auditory stimuli. The benefits of visual cues for the identification of sentences in noise also was demonstrated by Middelweerd and Plomp (1987). Additionally, McGurk and McDonald (1976) demonstrated that when a visual production of one consonant is paired with an auditory production of another consonant, a third consonant, neither visually or auditorily presented, is perceived. Specifically, when a visual /g/ is paired with an auditory /b/, a perceived /d/ will often occur. These studies demonstrate that the use of auditory and visual cues is crucial for the understanding of speech.

The temporal alterations of audiovisual signals also has been examined to determine how speech perception is affected when listening to degraded asynchronous multimodal signals (Grant & Greenberg, 2001; McGrath & Summerfield, 1985; Pandey, Kunov, & Abel, 1986). In one study, Grant and Greenberg demonstrated that normal hearing individuals could successfully recognize Harvard/IEEE sentences when they were filtered and exposed to audio-visual delays. The sentences were filtered into 1/3-octave bands, and two of the bands, one low and one high frequency band, were used to construct the final stimuli. They found that when the auditory signal led the visual signal by up to approximately 40 ms or the visual signal led the auditory signal by up to 160 to 200 milliseconds, the stimuli could be

successfully recognized. McGrath and Summerfield also demonstrated that when using an F0-modulated pulse train audio feed as part of an audiovisual signal, AV asynchronous speech was not affected when the audio and visual delays were less than 160 milliseconds. Similar findings were reported by Pandey, Kunov, and Abel who demonstrated that in the presence of background noise, AV asynchronous sentences can be successfully perceived for asynchrony levels up to approximately 120 milliseconds. More recently, Conrey and Pisoni (2006) demonstrated that young adults were able to identify isolated AV asynchronous words as being synchronous over a temporal window of approximately 150 milliseconds. The ability to recognize speech even though the audio and visual cues are not synchronous is clearly advantageous in noisy or reverberant environments where the audio and visual cues might not be aligned.

However, very little exploration of how individuals with hearing impairment integrate auditory and visual signals has been conducted. A hearing loss would imply that not all of the auditory frequency bandwidths from an audiovisual speech signal are adequately perceived, thereby potentially preventing the complete integration of auditory and visual stimuli. Grant and Seitz (1998) studied a group of individuals with mild sloping to severe hearing losses to determine the importance of synchronous AV speech stimuli for speech understanding. Their findings showed that speech recognition for audiovisual sentences was not affected until the audio delay exceeded 200 ms.

The effects of aging on AV asynchrony detection have also received little attention in the past. It has been reported that elderly listeners have more difficulty than younger adults with temporal processing of speech (Fitzgibbons & Gordon-Salant, 1996; Gordon-Salant & Fitzgibbons, 2001, 2004; Schneider, Pichora-Fuller, Kowalchuk, & Lamb, 1994; Snell, 1997; Sommers, Tye-Murray, & Spehar, 2005; Spehar, Tye-Murray, & Sommers, 2004). For example, the detection of brief temporal gaps between pairs of tones or noisebursts is about twice the magnitude for elderly individuals as it is for younger listeners (Fitzgibbons & Gordon-Salant, 1996; Schneider, Pichora-Fuller, Kowalchuk, & Lamb, 1994; Snell, 1997). Also, processing rapid speech has been shown to be more challenging for elderly individuals compared to younger adults (Gordon-Salant & Fitzgibbons, 2001, 2004). Finally, research has suggested that although older individuals often show a decline of speechreading abilities, their ability to comprehend time-altered visual speech is not compromised (Sommers, Tye-Murray, & Spehar, 2005; Spehar, Tye-Murray, & Sommers, 2004). It could be hypothesized, therefore, that due to the changes in auditory and visual processing, elderly individuals might integrate auditory and visual signals in a fundamentally different manner than younger individuals. Differences in auditory and visual integration could affect the perception of AV asynchronous speech.

Although AV asynchrony perception has not been examined in the elderly population, the integration of auditory and visual information has been assessed using several behavioral measures. For example, Cienkowski and Carney (2002) examined the McGurk effect in younger and older adults to assess the effects of aging on the ability to integrate auditory and visual information. They found that the percentage of fused responses to an auditory /bi/ and a visual /gi/ and an auditory /pi/ and visual /ki/ were not significantly different in younger and older age groups, suggesting that older adults are just as successful at integrating auditory and visual signals as younger adults. However, because individual measures of auditory and visual performance were not measured, it is not clear whether the younger and older study participants were integrating the auditory and visual cues in similar manners. That is, older adults could have relied more heavily either on auditory or visual cues to fuse stimuli whereas the younger adults could have used auditory and visual cues equally for fusion.

To address these concerns, Sommers, Tye-Murray, and Spehar (2005) recently examined the individual contributions that auditory and visual information provide for the integration of AV stimuli.

To minimize differences in unimodal performance, all study participants had normal hearing, as defined as hearing thresholds better than 25 dB HL from 250 Hz to 4000 Hz, and normal or normal-corrected vision. The study participants were asked to repeat consonants, isolated words and sentences under several different signal-to-noise ratios that would produce similar auditory performance between the two age groups. The speech materials also were presented in audiovisual and visual-only modalities in order to assess the effects of auditory or visual cues on speech understanding. The older normal hearing participants demonstrated significantly poorer speechreading skills than the younger adults suggesting that aging may be associated with declines in mechanisms that are responsible for the successful encoding of visual information, independently of hearing status. The findings also demonstrated that auditory and visual enhancement scores were comparable for the younger and older age groups. Therefore, the age differences that were obtained in audiovisual performance appear to reflect age related differences in speechreading abilities rather than the ability to integrate or combine auditory and visual stimuli.

In order to further examine the effects of hearing loss and aging on the integration of auditory and visual information in speech perception, we examined AV asynchrony perception in middle-aged and elderly normal hearing adults and cochlear implant users. Specifically, the present study measured the AV asynchrony detection skills in normal hearing adults and cochlear implant users to determine whether individuals who use cochlear implants detect AV asynchronous stimuli differently than normal hearing persons. This study also examined the effects that aging may have on the perception of AV asynchronous stimuli. Finally, because speech understanding deteriorates as the AV signal becomes increasingly asynchronous (Grant & Greenberg, 2001; Grant & Seitz, 1998), another goal of this study was to assess the association between AV asynchrony detection and speech perception abilities. Determining how normal hearing adults and cochlear implant users perceive AV asynchrony might provide some new insights into the sources of variability that underlie the wide range of speech perception skills that are typically observed within the cochlear implant population.

Method

Study Participants

Both normal hearing listeners and cochlear implant users participated in this study. Two different groups of English speaking cochlear implant users were recruited, 13 elderly adults ranging in age from 66 to 80 (mean 73 years old), and 12 middle-aged adults ranging in age from 41 to 54 years old (mean 47 years old). These individuals received either a Cochlear Corporation, an Advanced Bionics, or a Med El cochlear implant between the years of 1995 and 2004 at the Indiana University School of Medicine, Department of Otolaryngology—Head and Neck Surgery. One elderly participant had bilateral implants; the first device was implanted in 1996 while the second device was implanted in 2004. These adults were all native English speakers and none of them reported a history of stroke or head injury. The normal hearing participants consisted of 12 middle-aged adults ranging in age from 41 to 55 years old (mean age 48 years old), and 10 elderly adults ranging in age from 65 to 79 years old with a mean age of 70 years old. All of the normal hearing participants were recruited locally through posted advertisements and word of mouth. All of the normal hearing study participants reported that English was their first language, that they did not have prior speechreading training, and that they had no history of stroke or head injury.

Screening Tests

Pure-tone air-conductions thresholds were obtained for all normal hearing listeners from octaves 250 Hz to 4000 Hz using a Grason-Stadler GSI 61 Clinical Audiometer and EAR insert earphones. For

purposes of this study, normal hearing was defined as behavioral thresholds of 25 dB HL or better at all test frequencies. Additionally, all individuals included in this study had symmetrical audiometric hearing configurations (i.e., less than 20 dB HL difference between ears at one test frequency). Additionally, the sound field audiometric behavioral thresholds also were obtained for the cochlear implant users.

Screening for vision was completed prior to testing to ensure that all participants were capable of perceiving and encoding visual speech information. Normal or corrected-to-normal visual acuity of 20/25 or better was indicated for all study participants. Additionally, the Mini Mental Status Exam was administered to all individuals to assess cognitive function (Folstein, Folstein, & McHugh, 1975). All individuals who participated in this study received a score of 27 or better out of a possible 30 points. The mean score for cognitively intact individuals in the Folstein, Folstein, and McHugh study was 27.6 with a range of 24 to 30.

Procedures and Stimuli

Three speech perception measures were administered to all study participants. The Consonant-Nucleus-Consonant (CNC) word recognition test (Peterson & Lehiste, 1962), the Hearing in Noise Test (HINT) sentence recognition test (Nilsson, Soli, & Sullivan, 1994), and the City University of New York (CUNY) sentence test (Boothroyd, Hnath-Chisolm, Hanin, & Kishon-Rabin, 1988) were presented to study participants in an IAC booth. The auditory stimuli were presented at 70 dB SPL for the cochlear implant users and at 63 dB SPL for the normal hearing study participants. Background speech noise also was used for the normal hearing participants and presented at 70 dB SPL, thereby leading to a -7 dB signal-to-noise ratio. The CNC word test was administered first, followed by the HINT and the CUNY sentence tests. Additionally, the CUNY sentence test was presented in three modalities in the following order: auditory-only (A), visual-only (V) and audio-visually (AV). All study participants were instructed to repeat the stimuli they heard or saw for these tasks. Guessing was encouraged. For all tests, a percentage correct score was obtained as the dependent measure.

The speech AV asynchrony task conducted in this experiment was the same one employed by Conrey and Pisoni (2006). A list of ten familiar English words was presented to the listeners using a single talker. The words were chosen from the Hoosier Audiovisual Multitalker Database which contains digitized movies of isolated monosyllabic words spoken by single talkers (Lachs & Hernandez, 1998). The most intelligible talker of this database, as determined by Lachs (1999), was chosen for stimulus presentation. To prepare synchronous and asynchronous AV stimuli, Final Cut Pro 3 (copyright 2003, Apple Computer, Inc.) was used to manipulate the audio and visual signals. The stimuli were prepared such that the only cues that could be used to make judgments about the synchrony of the signals were temporally based between the audio and visual leads. Specifically, the audio track did not play while the screen was blank and all of the speech sounds and active articulatory movements remained within the movie.

Previous research on AV synchrony perception has revealed that normal-hearing young adults have a fairly wide range over which they will judge AV signals as being synchronous or asynchronous. That is, AV stimuli are typically judged as being asynchronous with 100% accuracy when the audio signal leads the visual signal by 300 ms (i.e., A300V) or more and when the visual signal leads the audio signal by 500 ms (i.e., V500A) or more (Conrey & Pisoni, 2006). For this study, 25 asynchrony levels that covered a range of 800 ms from A300V to V500A were used. Each successive level of asynchrony, either audio-leading or visual-leading, differed by 33.33 ms increments. Nine stimuli had auditory leads, one was synchronous, and 15 had visual leads for each of the ten stimulus words that were used. As a result, a total of 250 trials were presented to the participants in a randomized order. The visual and audio

stimuli were presented using an Apple G4 computer and Advent sound field speakers, respectively. The speakers were placed at ± 45° azimuth from the listeners who were seated approximately 19 inches from both the speakers and a Dell flat screen computer monitor.

Before the session began, the participants were given both written and oral instructions for performing the task and were presented with examples of asynchronous and synchronous AV stimuli. For each trial, the participants were asked to judge whether the AV stimulus was synchronous or asynchronous ("in sync" or "not in sync"). They were instructed to press one button on a button box if they thought the audio and visual stimuli were synchronous and a different button if they thought the stimuli were asynchronous. In order to alert the participants for an upcoming AV token, a fixation mark ("+") flashed on the computer screen for 200 ms which was then followed by a blank screen for 300 ms.

Results

The behavioral audiometric threshold data for cochlear implant users and the normal-hearing individuals are displayed in Tables 1 and 2, respectively. The cochlear implant user data presented in Table 1 reveal similar mean behavioral threshold responses for younger and older cochlear implant users at 250 Hz and 500 Hz. A one way ANOVA revealed no significant differences in thresholds between the two groups for these two warble tone behavioral thresholds. Significant differences between the younger and older cochlear implant users were obtained for the 1000 Hz (F (1,25) = 6.16, p = 0.02), 2000 Hz (F (1,25) = 7.14, p = 0.01) and 4000 Hz (F (1,25) = 4.56, p = 0.04) behavioral thresholds.

Young Subjects	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**
abf1	28	24	24	22	26
abi1	14	22	20	22	22
abk1	22	34	22	22	28
abn1	24	26	28	22	26
abs1	24	28	26	22	38
abv1	40	40	40	35	35
abw1	18	26	24	22	24
aby1	18	28	24	32	32
acb1	22	26	28	20	28
acr1	44	36	30	28	32
adh1	36	35	32	28	28
adi1	36	12	12	18	26
Mean	27.2	28.1	25.8	24.4	28.8
Elderly Subjects	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**
abg1	24	32	36	36	32
abh1	28	36	38	28	36
abj1	24	24	28	28	30
abm1	32	32	32	31	36
abo1	26	24	30	26	28
abp1	24	26	32	30	38
abr1	30	35	25	20	20
abq1	20	22	28	28	28
abz1	32	38	36	32	36
acc1	22	22	24	28	34
acd1	26	32	28	26	30
acf1	20	30	28	28	34
adc1-R	42	42	40	34	44
adc1-L	34	36	40	36	40
Mean	27.4	30.8	31.8	29.6	33.3

Table 1. Sound field behavioral audiometric thresholds for cochlear implant users. Thresholds are listed in dB HL for each test frequency.

One way ANOVAs performed using the normal hearing behavioral threshold data presented in Table 2 revealed significant differences in thresholds between younger and older adults for the right ear at 1000 Hz (F (1,21) = 8.42, p = 0.009) and 4000 Hz (F (1,21) = 8.79, p = 0.008). Left ear significant differences between the two aged groups also were noted at 1000 Hz (F (1,21) = 4.48, p = 0.04) and 4000 Hz (F (1,21) = 4.85, p = 0.04). A significant difference in the left ear pure tone average (PTA) was revealed (F (1,21) = 5.50, p = 0.03) but no significant difference between the two aged groups for the right PTA was indicated. Previous research has documented that individuals over the age of 60 experience significant hearing loss at frequencies above 4000 Hz (Lee, Matthews, Dubno, & Mills, 2005; Pearson et al., 1995). We cannot, therefore, rule out the possibility that the older adults who participated in this study did not have significant hearing loss at 8000 Hz. A hearing loss at 8000 Hz could have implications for the outcome measures (i.e., speech perception and asynchrony detection tasks) that were obtained during the course of the project.

Young Subjects	**Right Ear**						**Left Ear**					
	250 Hz	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**PTA-R**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**PTA-L**
NH1	20	15	5	5	15	8.3	15	5	5	0	5	3.3
NH2	10	5	0	-5	10	0	15	10	0	5	5	5
NH5	5	20	10	15	15	15	10	10	5	10	10	8.3
NH9	15	15	20	10	15	15	10	10	10	10	15	10
NH10	10	10	10	10	0	10	10	10	5	10	5	8.3
NH11	15	20	10	10	15	13.3	10	10	5	10	15	8.3
NH14	5	5	0	5	5	3.3	5	5	5	5	15	5
NH15	10	10	10	15	5	11.6	10	5	10	15	15	10
NH17	0	0	10	5	5	5	10	5	5	5	5	5
NH18	10	10	10	5	5	8.3	10	15	10	15	10	13.3
NH19	15	15	10	10	5	11.7	15	10	15	5	15	10
NH20	10	5	10	10	20	8.3	15	5	5	15	10	8.3
Mean	10.4	10.8	8.8	7.9	9.6	9.2	11.3	8.3	6.7	8.8	10.4	7.9
Elderly Subjects												
NH25	20	25	20	10	25	18.3	20	25	25	10	20	20
NH28	5	10	15	10	25	11.7	10	5	5	20	25	10
NH32	10	5	20	0	15	8.3	10	10	5	5	15	6.7
NH33	15	15	15	15	15	15	10	10	10	10	10	10
NH35	20	5	10	5	10	6.7	15	10	10	15	5	11.7
NH36	10	5	15	20	20	13.3	0	15	5	10	15	10
NH42	15	10	20	10	15	13.3	20	20	20	5	15	15
NH47	10	10	15	10	15	11.7	5	10	10	15	15	11.7
NH49	5	0	5	5	10	3.3	0	0	10	5	15	5
NH48	20	10	15	25	20	16.7	10	10	15	20	15	15
Mean	13	9.5	15	11	17	11.8	10	11.5	11.5	11.5	15	11.5

Table 2. Behavioral audiometric thresholds for normal hearing listeners. Thresholds were obtained using insert earphones and are listed in dB HL.

An individual cochlear implant user example of an AV asynchrony function is displayed in Figure 1 Panel A. In this figure, the mean proportion of synchronous responses is presented as a function of the asynchrony level in milliseconds. On the abscissa, the negative asynchrony levels indicate that the auditory signal led the visual signal by a specified time (e.g., A300V), the zero point indicates that both the audio and visual signals were synchronous in time (i.e., 0), and the positive asynchrony levels indicate that the visual signal led the audio signal (e.g., V400A). The ordinate axis represents the proportion of synchronous responses that were reported at a specific asynchrony level. Recall that each AV asynchrony level was presented using 10 different words and the listener's task was to judge whether or not the stimulus was out of sync. For this particular example, the trials of A300V, A267V, V367A, V400A, V433A, V467A, and V500A were judged to be asynchronous with 100% accuracy. This

individual reported that the audio and visual signals were completely synchronous for the asynchrony levels of A67V, A33V, 0, V33A, V67A, V100A, and V133A. For all other asynchrony levels, the study participant inconsistently reported that the AV stimuli were synchronous.

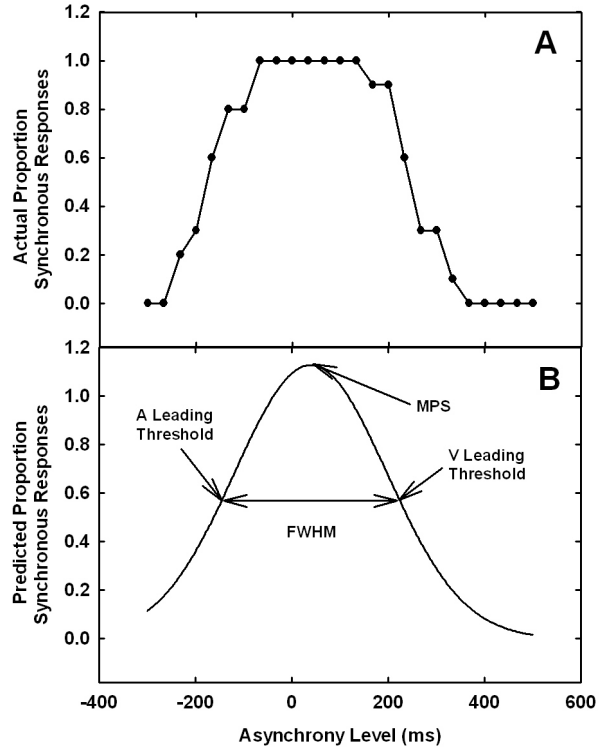


Figure 1. An individual AV asynchrony function. Panel A displays the observed function and Panel B shows the Gaussian curve fitted to the observed function. The proportion of synchronous responses is shown as a function of the asynchrony level. See text for details.

In order to quantify this AV asynchrony function, symmetrical Gaussian curves were fitted to individual asynchrony curves through the use of Sigma Plot 9.01 software and the following equation:

$$y = ae^{\left[-0.5\left(x - xo/b\right)^2\right]} \quad (1)$$

In this equation, *y* is the observed proportion of synchronous responses for each individual at each asynchrony level, *x*. The x-intercept, *xo*, represents the mean point of synchrony (MPS). Both *a* and *b* are generated parameters from the Sigma Plot software that aid with curve fitting. The Gaussian curve fitted to the individual asynchrony function shown in Panel A is displayed in Panel B of Figure 1. The four features that describe this AV asynchrony function are the MPS, the auditory (A) leading threshold, the visual (V) leading threshold and the full-width half maximum (FWHM). The A-leading threshold is the asynchrony level for the y value at 50% of the distance from the minimum to the maximum of the auditory leading portion of the curve (i.e., the left portion of the curve). Similarly, the V-leading

threshold is the asynchrony level for the corresponding y value at 50% of the distance from the maximum to the minimum of the visual leading portion (i.e., the right portion) of the Gaussian function. The FWHM is the value of the asynchrony width at the half-maxima of the function. For this individual, the MPS was 39.15 ms, the A leading threshold was -145.23 ms, the V leading threshold was 226.65, and the FWHM was 371.88 ms.

The mean AV asynchrony data for all of the cochlear implant users and the normal hearing adults are shown in Figure 2. For both panels, the mean proportion of synchronous responses is displayed as a function of the asynchrony level. The top panel of the figure shows the data for the cochlear implant users and the bottom panel displays the data for the normal hearing adults. The overall results for the younger adults (averaged over cochlear implant users and normal hearing adults) are shown with the dotted line. The data for the older adults are shown using the solid line. The MPS, the A-leading threshold, the V-leading threshold and the FWHM values for the normal hearing and cochlear implant users are presented in Table 3. Two-way ANOVA analyses were performed using age and hearing status as independent variables and MPS, A-leading threshold, V-leading threshold, and FWHM as the dependent variables. A significant age-effect finding was revealed for the A-leading threshold (F (1,46) = 4.989, p = 0.03) and for the FWHM (F (1,46) = 4.921, p = 0.03). For these two variables, the younger adults (i.e., cochlear implant users and normal hearing adults) had A-leading thresholds that were closer to the point of AV synchrony (i.e., 0 on the abscissa in Figure 2) and had narrower FWHMs than the older adults. No other main effects or interactions were obtained for any of the other analyses.

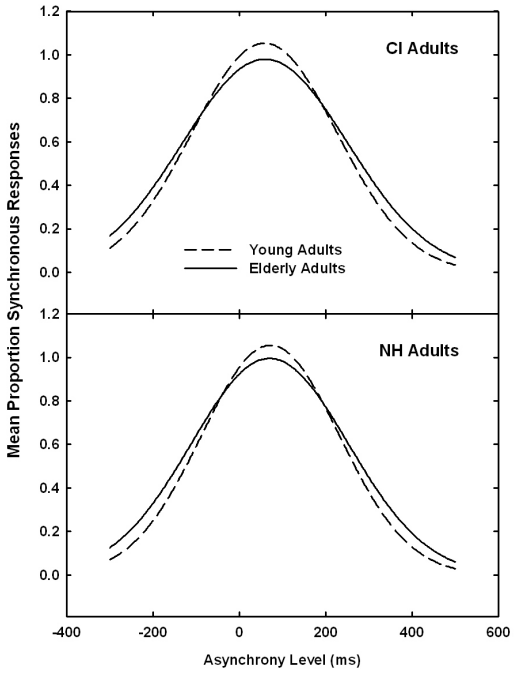


Figure 2. The mean AV asynchrony data for the cochlear implant users (top panel) and normal hearing adults (bottom panel). The mean proportion of the synchronous responses is displayed as a function of the asynchrony level.

	MPS	A leading Threshold	V leading Threshold	FWHM
Young - CI	58.4210	-135.9933	260.8817	396.8750
Elderly - CI	59.8846	-154.1708	295.8292	450.0000
Young - NH	72.3434	-112.3147	262.6853	375.0000
Elderly - NH	70.7044	-134.0327	294.0923	428.1250

Table 3. Mean Asynchrony Data. Values are in milliseconds; MPS: mean point of synchrony; FWHM: Full Width Half Maximum.

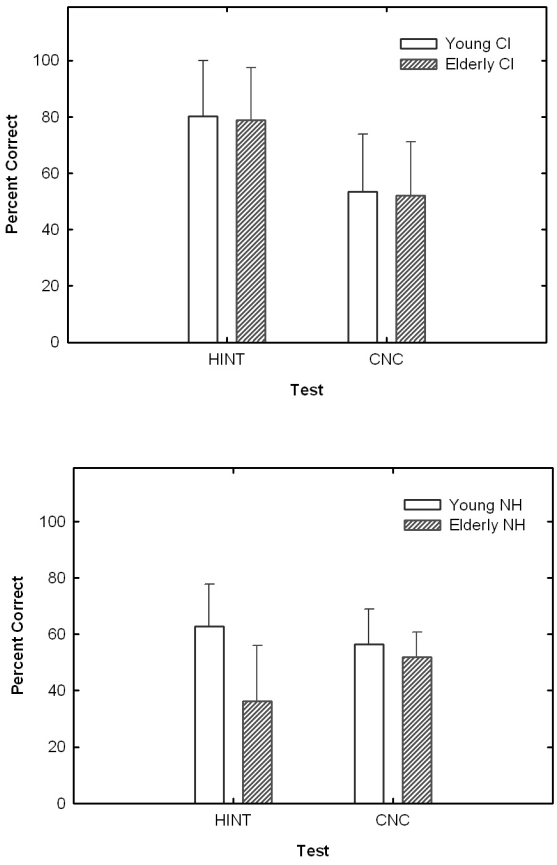


Figure 3. Mean percent correct scores and standard deviations for the HINT and CNC speech perception tests for the cochlear implant users (top panel) and normal hearing adults (bottom panel).

The results from the HINT, CNC, and CUNY speech perception tests are presented in Figures 3 and 4. Figure 3 displays the mean and standard deviations for the HINT and CNC sentence and word tests. The scores from the cochlear implant users are presented in the top panel; the scores from the normal hearing study participants are shown in the bottom panel. For both panels, the white bars show the data for the young individuals and the hatched bars show the data for the elderly participants. A two way ANOVA analysis of the HINT scores revealed a main effect for hearing status (F (1,43) = 31.34, $p <$ 0.001), and age (F (1,43) = 6.77, p = 0.013) and an interaction (F (1,43) = 5.37, p = 0.025). No significant differences were observed for the CNC data. Additionally, no significant correlations were observed between the FWHM data and the HINT and CNC data. There was, however, a trend for poorer HINT scores to be association with wider FWHMs for the young and elderly normal hearing adults, but this pattern did not reach significance (r = -0.371, p = 0.08).

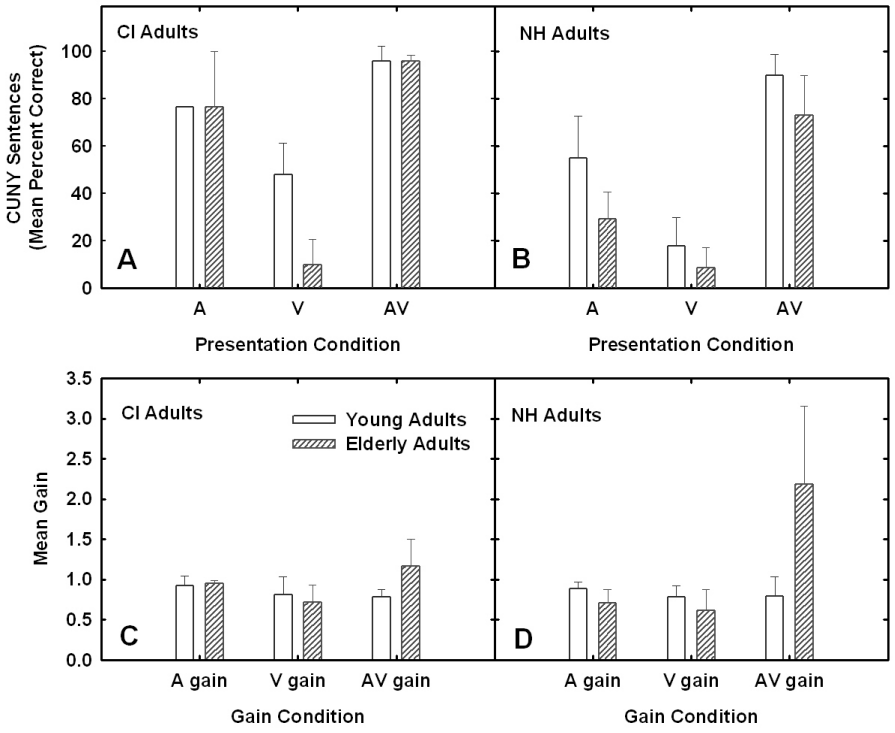


Figure 4. Mean percent correct (panels A and B) and gain scores (panels C and D) for the CUNY sentence test for the cochlear implant and normal hearing participants.

Figure 4 shows the results from the CUNY sentence test, displayed for the presentation modalities, auditory-alone (A), visual-alone (V) and audiovisually (AV). Panels A and C show the data for the cochlear implant users, and Panels B and D show the data for the normal hearing participants. The mean percent correct CUNY scores for each presentation condition (i.e., A, V and AV) are shown in Panels A and B for the cochlear implant and normal hearing participants, respectively. Additionally, Panels C and D show the A, V, and AV-gain or enhancement scores obtained from the CUNY sentence test. These gain measures assess how both auditory and visual cues contribute to overall speech

understanding and are described in further detail below. The white bars in all panels show the data for the elderly adults and the hatched bars show the data for the younger adults. The error bars represent one standard deviation around the mean.

In order to assess the effects of aging and hearing status on the perception of the CUNY sentences, a series of two-way ANOVA analyses were conducted using the scores from the A, V and AV presentations. For the CUNY A presentation, significant main effects were found for age (F (1,43) = 5.69, p = 0.02) and hearing status (F (1,43) = 41.27, $p < 0.001$), along with an interaction (F (1,43) = 5.65, p = 0.02). As shown in Figure 4 Panels A and B, the mean CUNY A results for the normal hearing participants were lower than the scores for the cochlear implant users. Recall that the cochlear implant users listened to the sentences in quiet, and the normal hearing participants listened to the CUNY sentences in the presence of background noise in order to simulate the effects of a hearing loss and reduce performance from ceiling effects.

The main effect of hearing status was most likely the result of the added background noise that was used for the normal hearing adults and not for the cochlear implant users. In addition, the interaction between age and hearing status may be due to differences in the listening conditions between the two groups. Specifically, the added background noise probably reduced the performance of the older normal hearing adults to a greater extent than it affected the performance of the younger normal hearing adults. Degraded speech understanding in noise for elderly normal hearing individuals has been previously documented (Stuart & Phillips, 1996). These differences in performance for the older and younger groups were not observed in the cochlear implant population most likely due to the absence of background noise in the listening environment.

Main effects for the V-only presentation were observed for both age (F (1,43) = 50.34, $p < 0.001$) and hearing status (F (1,43) = 22.55, $p < 0.001$). An interaction also was observed (F (1,43) = 18.95, $p < 0.001$). The data for the CUNY V-only scores in Figure 4 Panels A and B reveal that the younger cochlear implant users and normal hearing adults were better speechreaders than the older adults, a finding that has been observed previously (Hay-McCutcheon, Pisoni, & Kirk, 2005; Sommers, Tye-Murray, & Spehar, 2005). However, the younger cochlear implant users identified V speech better than the younger normal hearing individuals. In addition, the data indicate that both the older normal hearing participants and the older cochlear implant users performed similarly in this speechreading task.

Finally, a two-way ANOVA of the CUNY AV scores revealed main effects for age (F (1,43) = 8.90, p = 0.005), hearing status (F (1,43) = 26.92, $p < 0.001$), and an interaction. The data shown in Figure 3 revealed three findings: first, both the older and younger cochlear implant users performed similarly on this task; second, the cochlear implant users performed better than the normal hearing participants, and third, the younger normal hearing adults achieved higher scores than the older normal hearing adults. The results shown here, however, should be interpreted with some caution because ceiling effects were observed for the AV results obtained for the cochlear implant users.

In order to assess the separate contributions that auditory and visual cues provide for speech understanding, auditory and visual gain scores were calculated. The A, V and AV scores obtained from the CUNY speech perception test were used to calculate the benefit that audition-alone (A-gain) and the vision-alone (V-gain) cues provide for speech perception (Lachs, Pisoni, & Kirk, 2001; Sommers, Tye-Murray, & Spehar, 2005). Specifically, the "A-gain" score represents the improvement in speech perception due to the addition of visual information to the auditory signal (i.e., AV-V/100-V), and conversely, the "V-gain" score represents the improvement in speech perception due to the addition of auditory cues to the visual signal (i.e., AV-A/100-A). For the A-gain score, the contributions that the

visual cues add to the audiovisual results are subtracted and this value is subsequently divided by the difference between the possible visual-alone score and the obtained visual-alone score. Similarly, for the V-gain score, the contributions that the auditory cues add to the audiovisual results are determined and then divided by the difference between the possible auditory-alone score and the obtained auditory-alone score.

We also assessed the overall integration of auditory and visual information, and determined the superadditive nature of the use of combined modalities for speech perception (i.e., AV-gain = AV/A+V). For individual cases, if the combined AV performance is the same as the addition of the A-alone and V-alone scores, then AV-gain would be equal to one, and little AV enhancement would be indicated. Alternatively, if the combined AV scores are greater than the simple sum of the A and V scores alone, then the integration of auditory and visual information is beneficial (i.e., superadditive) for speech understanding.

The gain scores are presented in Panels C and D of Figure 4 for the cochlear implant users and the normal hearing adults, respectively. A series of two-way ANOVA analyses were conducted using the A-gain, V-gain and AV-gain scores as the dependent measures and hearing status and age as the independent variables. Because the AV results for the cochlear implant users were at ceiling, and these data were used to determine the gain measures, the following results need to be cautiously considered.

For the A-gain scores, significant main effects were obtained for age (F (1,43) = 4.823, p = 0.034) and hearing status (F (1,43) = 19.64, $p < 0.001$). In addition, an interaction was also observed (F (1,43) = 10.88, p = 0.002). For the V-gain scores, only a main effect for age was observed (F (1,43) = 4.46, p = 0.04). Main effects for age (F (1,43) = 37.67, $p < 0.001$), hearing status (F (1,43) = 12.81, p = 0.001), and an interaction effect (F (1,43) = 12.02, p = 0.001) were found for the AV-gain scores. The AV-gain scores displayed in Figure 4 reveal that the elderly normal hearing adults and cochlear implant users had scores above one, suggesting that the combined use of the auditory and visual cues provided greater benefit to the older adults than the younger adults.

To assess the relations between the CUNY speech perception scores and the AV asynchrony detection, Pearson correlations were conducted. The results are summarized in Figure 5. In this figure, the results for the normal hearing listeners are presented in the three left graphs; the results for the cochlear implant users are presented in the three right graphs. The white circles and triangles represent individual data from the young normal hearing adults and cochlear implant users, respectively. The gray diamonds and squares represent data from the elderly normal hearing adults and cochlear implant users.

The Pearson correlations, presented in each panel, reveal that for normal hearing individuals the wider the width of the FWHM the poorer the performance on CUNY A, V and AV tasks. This trend, however, was not observed in the scores obtained for the cochlear implant users. For the V results, there was a tendency for the cochlear implant users to have wider FWHMs with poorer perception, but this pattern did not reach significance (r = -0.368, p = 0.071). A significant correlation (r = -0.438, p = 0.029) was obtained for the AV data suggesting that wider FWHMs resulted in poorer speech perception scores. Because of ceiling effects, this finding needs to be viewed with some caution.

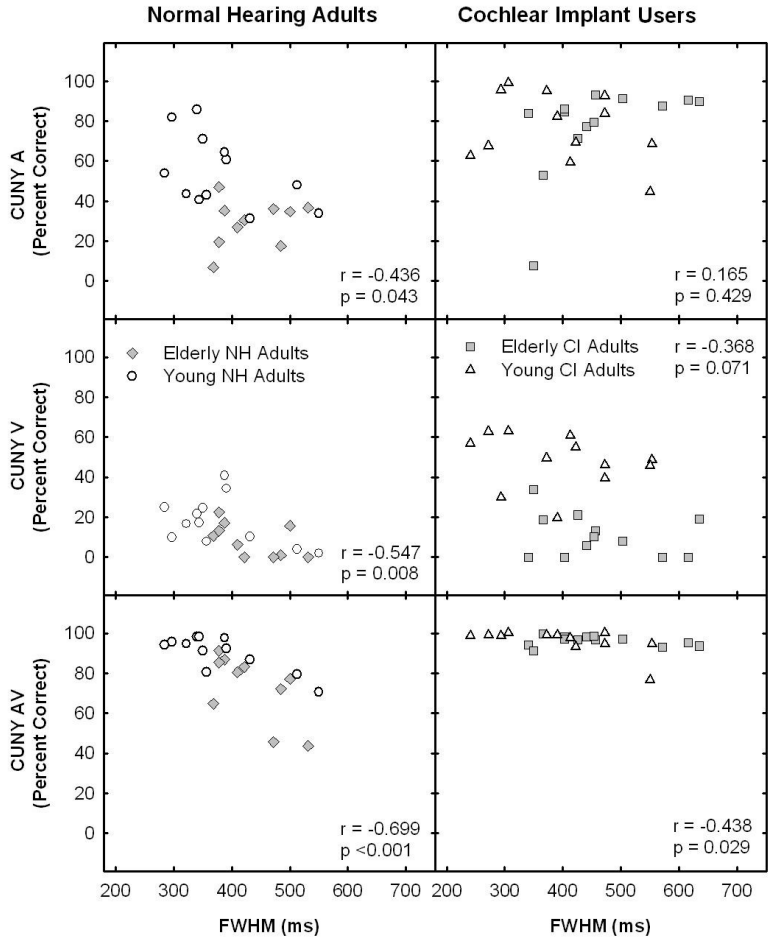


Figure 5. Correlation results for the CUNY and FWHM data for the normal hearing adults (left three panels) and cochlear implant users (right three panels).

The Pearson correlation coefficients for the gain scores and the FWHMs are presented in Figure 6. The normal hearing data are presented in the left three graphs and the data from the cochlear implant users are presented in the right three graphs. The white circles and the gray diamonds represent the data for the normal hearing young and older adults, respectively. The white triangles and the gray squares represent the data for the young and older cochlear implant users, respectively.

For the normal hearing individuals, a significant trend was observed for the A-gain (r = -0.689, $p < 0.001$) and V-gain (r = -0.659, $p < 0.001$) scores to decrease with increasing FWHM width. This trend was not observed for the AV-gain results (r = 0.114, p = 0.613). Although there was a tendency for the A-gain (r = -0.346, p = 0.090) scores to decrease with increasing FWHM width, this observation was not significant for the cochlear implant users. A significant correlation was observed for the V-gain (r = -0.610, p = 0.001) scores for the cochlear implant users. Specifically, lower V-gain scores were correlated with wider FWHMs. For both the normal hearing and the cochlear implant users, the AV-gain scores were not significantly correlated with the FWHM width.

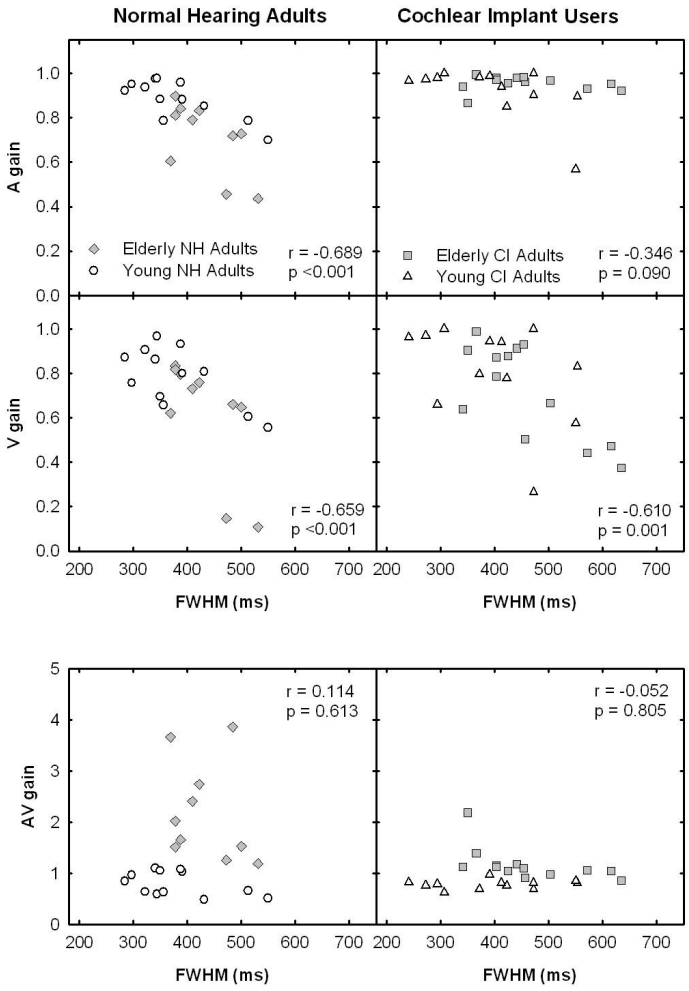


Figure 6. Correlation results for the gain scores and the FWHM (ms) data for the normal hearing adults (left three panels) and the cochlear implant users (right three panels).

Discussion

The goal of this study was to examine how normal hearing listeners and cochlear implant users perceive AV asynchrony in speech and assess the association between AV asynchrony detection and speech understanding abilities. The results of this study revealed no significant differences in performance between the normal hearing adults and the cochlear implant users in detecting and perceiving AV asynchronous single words. There was, however, a significant difference between the number of spoken words that were identified as being asynchronous for the elderly and middle-aged study participants. Specifically, the elderly normal hearing and cochlear implant individuals identified asynchronous words as being synchronous over a wider time window than did the younger adults. That is, the average FWHMs for the elderly normal hearing and cochlear implant population was approximately 440 ms compared to an average of 386 ms for the younger normal hearing individuals and cochlear implant users. Moreover, we found that the width of the asynchrony function was significantly correlated with the CUNY A, V and AV results for the normal hearing study participants. Correlations indicated that wider FWHMs were associated with poorer speech perception skills, a finding that

replicated the earlier results of Conrey and Pisoni (2006). This pattern also was observed with the HINT scores in the current study, but the findings were not significant. Conversely, for individuals with cochlear implants, the A and V CUNY scores were not significantly correlated with the FWHM data. Overall, the results suggest that older individuals have more difficulty identifying AV asynchronous stimuli and that AV skills are correlated with speech perception abilities.

AV Asynchrony Detection

The AV asynchrony findings reported here are similar to those previously reported. The findings of Conrey & Pisoni (2006) suggested that the FWHM was on average 372 ms for young adults aged 18 to 22 years old, which was very similar to the FWHMs reported in the present study for the younger adults (i.e., 386 ms). Grant and Greenberg (2001), McGrath and Summerfield (1985), and Pandy, Kunov, and Abel (1986) also reported findings on speech understanding using AV asynchronous material and all three papers noted that sentences can be successfully identified when the auditory and visual components are approximately 200-250 milliseconds out of sync. Although the findings from the current study cannot be directly compared to the results of these earlier studies because of procedural differences, it is clear that AV asynchronous speech is perceived as synchronous over a window of approximately several hundred milliseconds.

AV Asynchrony Detection and Aging

The findings from this study also suggest that age rather than hearing impairment is more closely tied with the detection of AV asynchronous speech. The data displayed in Figure 2 suggest that compared to younger adults, older individuals have a significantly wider temporal integration window over which they identify AV asynchronous speech as being synchronous. The present findings suggest that individuals with a severe-to-profound hearing loss who use a cochlear implant do not have more difficulty detecting AV asynchronous speech than individuals with normal hearing. To more fully understand the effects of cochlear implantation on the perception of AV asynchronous speech, further work should focus on the identification, rather than just the detection of AV asynchronous speech in individuals who use a cochlear implant. Additionally, future work should address the effects that the degree of hearing loss has upon both the detection and understanding of AV asynchronous speech. Through these types of studies it may be possible to more clearly describe the effect of hearing loss on the perception of AV asynchronous speech.

Age-related effects also have been reported in several previous studies that evaluated auditory perception abilities in younger and older listeners (Fitzgibbons & Gordon-Salant, 1996; Gordon-Salant & Fitzgibbons, 2004; Pichora-Fuller & Souza, 2003; Stuart & Phillips, 1996). Specifically, several studies have found that older normal hearing listeners have more difficulty than young normal hearing adults with temporal processing tasks such as gap detection, sound duration discrimination, and identifying time compressed speech (Fitzgibbons & Gordon-Salant, 1996; Gordon-Salant & Fitzgibbons, 2004). Other studies have reported that younger adults correctly answer more comprehension questions after listening to a passage presented in a -15 dB signal-to-noise ratio condition than do older adults (Schneider, Daneman, Murphy, & Kwong See, 2000). Stuart and Phillips (1996) also demonstrated that older normal hearing listeners identified fewer monosyllabic words when presented in background noise than did younger adult listeners. Evidence suggests that the declines in speech perception performance can be partially attributed to changes that have occurred within the peripheral auditory system (Humes, 1996; Schneider, Daneman, Murphy, & Kwong See, 2000; Souza & Turner, 1994). However, as noted above, the observed differences between younger and older adults with more complex tasks such as gap detection and sound duration discrimination cannot exclusively be attributed to peripheral sensory

deterioration (Fitzgibbons & Gordon-Salant, 1996; Gordon-Salant & Fitzgibbons, 2004; Pichora-Fuller & Souza, 2003). Most likely, neurophysiological changes that occur within the central auditory system also contribute to the declines in performance observed in complex listening tasks with elderly individuals.

In terms of the auditory periphery, both younger and older adults included in this study had hearing within normal limits. However, within the range of normal hearing, the behavioral audiometric threshold data suggested that the older normal hearing individuals had significantly higher thresholds than the younger normal hearing individuals at 1000 Hz and 4000 Hz. Additionally, the sound field thresholds for the cochlear implant users revealed that the older study participants had significantly lower thresholds at 1000 Hz, 2000 Hz and 4000 Hz. It is possible, therefore, that the differences in AV asynchrony detection could have been a direct consequence of the physiological differences in the peripheral auditory system between younger and older individuals.

Differences in the central auditory systems between younger and older adults also should be considered when explaining the observed differences in AV asynchrony detection. Specifically, several studies have shown that older adults experience difficulty with tasks that require them to divide their attention. Madden, Pierce, and Allen (1996) demonstrated that the reaction time to identify specific tokens from a group of distracting tokens was significantly longer in an elderly group of individuals aged 63 to 70 than in a young group of individuals aged 18 to 29 years old. Mayr (2001) reported that in a task requiring study participants to switch between different types of decisions between trials, older adults (mean age 71 years old, SD=3.3 years) had a significant longer reaction time for this task than did younger adults (mean age 33 years old, SD=1.4 years). It is possible, therefore, that the attentional demands that were required in the current study (i.e., attending to both the auditory and visual streams and making a conscious and explicit decision about their synchrony) placed greater processing demands on the older participants than the younger participants, and this could have contributed to the observed differences in performance between the two aged groups.

AV Asynchrony Detection and Speech Perception in Cochlear Implant Users

Contrary to the findings of the data for the normal hearing individuals, the relation between the AV asynchrony detection task and the speech perception scores were not highly correlated as shown in Figure 4. A correlation was observed between the AV CUNY results and the FWHM data, but this finding needs to be interpreted with some caution due to the ceiling effects noted with the AV CUNY data. Because the mechanism responsible for the correlation of the CUNY speech perception data and the FWHM in normal hearing individuals is not well understood, it is difficult to determine the reason for the lack of correlation found between the speech perception findings and the FWHM data in the cochlear implant users.

It is possible that for normal hearing individuals both the auditory and visual domains were effectively utilized to detect the presence of AV asynchrony, and that the processing of the AV input signal along the peripheral and central nervous system pathways was successfully and effectively completed. Additionally, it is possible that the processing of AV asynchronous stimuli and speech occurs using similar mechanisms in normal hearing individuals. Conversely, the cochlear implant users have had inadequate or altered peripheral auditory pathway processing prior to and following implantation. This change in processing strategies has been documented in neural plasticity data obtained from animal models (Shepherd, Baxi, & Hardie, 1999; Shepherd & Hardie, 2001). Thus altered peripheral processing could have an impact on the central processing of the signal, which would ultimately affect the perception of the AV asynchronous stimuli and the speech tokens associated with the speech perception tasks. The integration of auditory and visual information may occur in a fundamentally different manner

for cochlear implant users compared to the normal hearing individuals and this processing might have an impact on both the perception of AV synchronous and asynchronous speech stimuli.

Conclusions

In summary, the findings from this experiment suggest that aging has a greater effect on the detection of AV asynchronous speech than a severe-to-profound hearing loss that has been partially corrected through the use of a cochlear implant. For normal hearing adults, the width of the temporal processing window over which AV asynchronous speech was identified as being synchronous was correlated with speech perception skills. We found that the perception of wider temporal windows was associated with poorer speech understanding. Conversely, for cochlear implant users, the temporal width of the AV asynchrony function was not correlated with speech perception skills. The findings suggest that the perception of speech may occur in a fundamentally different manner for hearing-impaired individuals who use cochlear implants users than it does for normal hearing adults.

References

Boothroyd, A., Hnath-Chisolm, T., Hanin, L., & Kishon-Rabin, L. (1988). Voice fundamental frequency as an auditory supplement to the speechreading of sentences. *Ear & Hearing, 9*, 306-312.

Cienkowski, K.M., & Carney, A.E. (2002). Auditory-visual speech perception and aging. *Ear & Hearing, 23*, 439-449.

Conrey, B., & Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America, 119*, 4065-4073.

Fitzgibbons, P.J., & Gordon-Salant, S. (1996). Auditory temporal processing in elderly listeners. *Journal of the American Academy of Audiology, 7*, 183-189.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). Mini Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Gordon-Salant, S., & Fitzgibbons, P.J. (2001). Sources of age-related recognition difficulty for time-compressed speech. *Journal of Speech, Language, and Hearing Research, 44*, 709-719.

Gordon-Salant, S., & Fitzgibbons, P.J. (2004). Effects of stimulus and noise rate variability on speech perception by younger and older adults. *Journal of the Acoustical Society of America, 115*(4), 1808-1817.

Grant, K.W., & Greenberg, S. (2001). *Speech intelligibility derived from asynchronous processing of auditory-visual information.* Paper presented at the International Conference on Auditory Visual Speech Processing, Scheelsminde, Denmark.

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*, 2438-2450.

Hay-McCutcheon, M.J., Pisoni, D.B., & Kirk, K.I. (2005). *Speech recognition skills in the elderly cochlear implant population: A preliminary examination.* Paper presented at the MidWinter Meeting of the Association for Research in Otolaryngology, New Orleans, LA, Feb 19-24, 2005.

Humes, L.E. (1996). Speech understanding in the elderly. *Journal of the American Academy of Audiology, 7*, 161-167.

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

Lachs, L., & Hernandez, L.R. (1998). Update: The Hoosier audiovisual multitalker 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., & Kirk, K.I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report. *Ear & Hearing, 22*, 236-251.

Lee, F.S., Matthews, L.J., Dubno, J.R., & Mills, J.H. (2005). Longitudinal study of pure-tone thresholds in older persons. *Ear & Hearing, 26*, 1-11.

Madden, D.J., Pierce, T.W., & Allen, P.A. (1996). Adult age differences in the use of distractor homogeneity during visual search. *Psychology and Aging, 11*, 454-474.

Mayr, U. (2001). Age differences in the selection of mental sets: The role of inhibition, stimulus ambiguity, and response-set overlap. *Psychology and Aging, 16*, 96-109.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, 678-685.

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

Middelweerd, M.J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82*, 2145-2147.

Nilsson, M., Soli, S.D., & Sullivan, J.A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America, 95*, 1085-1099.

Pandey, P.C., Kunov, H., & Abel, S.M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research, 26*, 27-41.

Pearson, J.D., Morrell, D.H., Gordon-Salant, S., Brant, L.J., Metter, E.J., Klein, L.L., et al. (1995). Gender differences in a longitudinal study of age-associated hearing loss. *Journal of the Acoustical Society of America, 97*, 1196-1205.

Peterson, G.E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders, 27*, 62-65.

Pichora-Fuller, M.K., & Souza, P.E. (2003). Effects of aging on auditory processing of speech. *International Journal of Audiology, 42*, 12S11-12S16.

Schneider, B.A., Daneman, M., Murphy, D.R., & Kwong See, S. (2000). Listening to discourse in distracting settings: The effects of aging. *Psychology and Aging, 15*, 110-125.

Schneider, B.A., Pichora-Fuller, M.K., Kowalchuk, D., & Lamb, M. (1994). Gap detection and the precedence effect in young and old adults. *Journal of the Acoustical Society of America, 95*, 980-991.

Shepherd, R.K., Baxi, J.H., & Hardie, N.A. (1999). Response of inferior colliculus neurons to electrical stimulation of the auditory nerve in neonatally deafened cats. *Journal of Neurophysiology, 82*, 1363-1380.

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

Snell, K.B. (1997). Age-related changes in temporal gap detection. *Journal of the Acoustical Society of America, 101*, 2214-2220.

Sommers, M., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear & Hearing, 26*, 263-275.

Souza, P.E., & Turner, C.W. (1994). Masking of speech in young and elderly listeners with hearing loss. *Journal of Speech and Hearing Research, 37*, 655-661.

Spehar, B., Tye-Murray, N., & Sommers, M. (2004). Time-compressed visual speech and age: A first report. *Ear & Hearing, 25*, 565-572.

Stuart, A., & Phillips, D.P. (1996). Word recognition in continuous and interrupted broadband noise by young normal-hearing, older normal-hearing, and presbyacusic listeners. *Ear & Hearing, 17*, 478-489.

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Nonword Repetition with Spectrally Reduced Speech: Some Developmental and Clinical Findings[1]

Rose A. Burkholder,[2] Susannah V. Levi, Caitlin M. Dillon[3] and David B. Pisoni

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

[1] This research was supported by NIH-NIDCD Research Grants R01 DC00111 and R01 DC03937, Training Grant DC00012 and the American Hearing Research Foundation.
[2] Now at the University of Michigan, Ann Arbor, MI.
[3] Now at Haskins Laboratories, New Haven, CT.

Nonword Repetition with Spectrally Reduced Speech: Some Developmental and Clinical Findings

Abstract. Nonword repetition skills were examined in 24 pediatric cochlear implant users and 18 normal-hearing adult listeners. The normal-hearing adult listeners heard spectrally degraded nonwords that were processed through an acoustic simulation of a cochlear implant designed to mimic the auditory input received by cochlear implant users. Two separate groups of normal-hearing adult listeners assigned perceptual accuracy ratings to the nonword responses of the pediatric cochlear implant users and the normal-hearing adult speakers. Overall, the nonword repetitions of children using cochlear implants were rated as more accurate than the nonword repetitions of the adults. The nonword repetition accuracy ratings from both groups of subjects were correlated with their open- and closed-set word recognition scores and with their forward digit spans. However, only the accuracy scores from pediatric cochlear implant users were correlated with measures of speech production accuracy. This finding may reflect the lack of variance in the accuracy ratings and the linguistic analysis of the adults' nonword repetitions as well as differences in overall fluency of the productions. In terms of overall accuracy, the children performed better on speech perception tasks, while the adults were better on working memory tasks. These results suggest that although the pediatric cochlear implant users had more experience and success in perceiving speech under degraded auditory conditions with their cochlear implant, developmental differences in their memory skills prevent them from performing as well on working memory tasks as mature listeners who were exposed to a spectrally degraded speech for only a short period of time in the laboratory.

Introduction

For over a decade, nonword repetition has been a popular task used to assess phonological working memory in a wide range of developmental and clinical populations (Gathercole, Willis, Baddeley, & Emslie, 1994; Bishop, North, & Donlan, 1996; Edwards & Lahey, 1998; Laws, 1998; Sahlen, Reuterskiold-Wagner, Nettelbladt, & Radeborg, 1999; Briscoe, Bishop, & Norbury, 2001). Nonword repetition is assumed to be a more accurate measure of phonological memory than other simple auditory memory tasks such as forward digit span or direct assessments of immediate serial recall because it involves a more complex series of information processing operations. To complete a nonword repetition task, listeners must first accurately perceive and encode a novel linguistic pattern in the absence of any semantic or pragmatic context or lip-reading cues. After encoding, the nonword must then be retained in short-term memory using the subvocal verbal rehearsal component of the phonological loop. Finally, nonword repetition also requires that listeners reassemble the novel auditory pattern into a fluent spoken response and execute a series of motor commands to the speech articulators. Because of its complexity and the specific information processing steps it involves, the nonword repetition task has recently emerged as a diagnostic tool used by researchers and clinicians interested in the speech, language, and memory skills of deaf children who use cochlear implants (CIs).

The nonword repetition skills of pediatric CI users have been explored extensively in our laboratory in order to account more fully for the wide individual differences in speech, language, and other cognitive outcomes in this clinical population (Carter, Dillon, & Pisoni, 2002; Cleary, Dillon, & Pisoni, 2002; Dillon, Burkholder, Cleary, & Pisoni, 2004; Dillon & Pisoni, 2004; Dillon & Pisoni, under revision). Two primary methods have been used to assess pediatric CI users' nonword repetition

performance. As an alternative to scoring nonword repetitions as simply correct or incorrect, perceptual accuracy ratings and more detailed linguistic analyses—specifically segmental and suprasegmental accuracy—have been carried out on the nonword repetitions of pediatric CI users.

When scored dichotomously as either correct or incorrect, pediatric CI users' nonword repetition skills appear to be at floor and lack any informative variability (Carter et al., 2002). However, by using segmental and suprasegmental linguistic analyses along with perceptual accuracy ratings, the qualitative characteristics of pediatric CI users' nonword repetition skills have been more fully and accurately documented. In addition, by using these more descriptive and sensitive measures of nonword repetition performance, numerous correlates and predictors of pediatric CI users' nonword repetition skills have been identified. The relationships identified between CI users' nonword repetition performance and cognitive processing variables such as subvocal verbal rehearsal, memory, and reading have provided some valuable insights into the large individual differences in speech, language, and other cognitive outcomes that are frequently observed in this clinical population (Carter et al., 2002; Cleary et al., 2002; Dillon, Burkholder, et al., 2004; Dillon & Pisoni, under revision).

In a small group of pediatric CI users who were able to give a spoken response to each of the nonwords used in the study, Cleary et al. (2002) found substantial variability in the overall perceptual accuracy ratings that naïve, normal-hearing (NH) adult listeners assigned to the children's nonword repetitions. They also found that the perceptual accuracy ratings were related to a number of speech perception and production measures after demographic variables were partialled out of the analysis. Both open- and closed-set speech perception scores were positively correlated to the children's nonword repetition ratings. This result confirms that reliable initial auditory encoding of the novel nonword patterns is essential for pediatric CI users to complete the nonword repetition task successfully.

Several speech production measures were also found to be related to pediatric CI users' mean perceptual accuracy ratings (Cleary et al., 2002). Speech intelligibility scores obtained from short sentences spoken by the children were positively correlated with the overall nonword repetition rating that they received. In addition, the durations of these sentences were related to the children's nonword repetition accuracy ratings. Children who articulated the sentences more slowly received lower nonword repetition accuracy ratings. This result suggests a relationship between speaking rate and the ability to repeat novel nonword stimuli from representations in immediate memory.

In a larger study of 76 pediatric CI users who varied in their ability to provide a spoken response for each nonword token, Dillon, Burkholder, et al. (2004) confirmed some of the earlier findings from Cleary et al., (2002). They found strong relationships between several speech perception, speech production, and memory measures and nonword repetition accuracy ratings. Using linear regression, Dillon et al. found that the duration of sentences spoken by the children, which can be taken as an index of subvocal verbal rehearsal speed, was the strongest predictor of nonword repetition ratings. Two other significant predictors of nonword repetition accuracy ratings assigned to the children were scores obtained on the closed-set Word Intelligibility by Picture Identification (WIPI; Ross & Lerman, 1979) test and speech intelligibility ratings obtained from a separate group of adult listeners. Taken together, the results of these two studies indicate a close relationship between speech perception, speech production, and speaking rate measures and nonword repetition accuracy ratings.

Speaking rate may be related to the ability to reproduce a novel nonword pattern not only because it indexes pediatric CI users' abilities to produce speech in a fluent and fluid manner, but because it is also an index of subvocal rehearsal speed, that is, the speed at which verbal information can be refreshed within the phonological loop of working memory (Kail & Park, 1994; Cowan, Wood, Wood,

Keller, Nugent et al., 1998; Burkholder & Pisoni, 2003; Pisoni & Cleary, 2003). Because nonword repetition is a phonological working memory task, subvocal verbal rehearsal is a very important and integral process involved in its completion. Similarly, subvocal verbal rehearsal is also an important process that contributes to pediatric CI users' auditory and visual memory spans (Cleary, Pisoni, & Geers, 2001; Burkholder & Pisoni, 2003; Pisoni & Cleary, 2003).

It is not surprising then that auditory memory spans have also been found to be related to the pediatric CI users' nonword repetition accuracy ratings. Cleary et al. (2002) found a strong positive correlation between the pediatric CI users' forward digit spans and their average nonword repetition rating. Children with longer forward digit spans received higher nonword repetition accuracy ratings. The relationship between auditory memory span and nonword repetition has been documented previously in numerous populations of NH children (e.g. Brady, Mann, & Schmidt, 1987; Gathercole & Baddeley, 1989; 1990). In addition, both auditory memory span and nonword repetition abilities have been found to be positively correlated with NH children's vocabulary development, vocabulary size, usage of syntactically complex sentences, and word learning abilities in both native and nonnative languages (Gathercole & Baddeley, 1990; Service, 1992; Edwards, Beckman, & Munson, 2004).

Because nonword repetition is predictive of and related to such a critical set of speech and language abilities in NH children, it has been valuable to examine this ability in pediatric CI users as well. The nonword repetition skills of pediatric CI users may help explain some of the large individual differences in speech, language, and other cognitive outcomes that are frequently observed in this population and may provide insight into the processes that these children use while developing language and language-related skills. Several language and language-related skills have been found to be associated with pediatric CI users' nonword repetition accuracy. Two previous studies have shown a positive correlation between pediatric CI users' comprehension of spoken language and nonword repetition skills (Cleary et al., 2002; Dillon, Burkholder, et al., 2004). More recently, Dillon and Pisoni (under revision) found that measures of reading and lexical diversity in spontaneous speech samples were strongly correlated with the perceptual accuracy ratings of the nonword repetitions of deaf children using CIs. Taken together, research using perceptual accuracy ratings has indicated that pediatric CI users' nonword repetition skills are strongly linked to a number of speech perception, speech production, memory, and reading skills.

Several of the same measures of speech perception and production and subvocal verbal rehearsal that are related to pediatric CI users' nonword repetition accuracy ratings have also been found to be positively correlated with linguistic analyses conducted on their nonword repetition responses. A suprasegmental analysis of pediatric CI users' nonword responses indicated that both the ability to correctly reproduce primary stress and the appropriate number of syllables was related to speech perception and production scores and subvocal verbal rehearsal (Carter et al., 2002). Children's ability to produce consonants in the nonwords accurately was also related to these three variables (Cleary et al., 2002; Dillon, Cleary, Pisoni, & Carter, 2004). In a more detailed analysis of the segmental accuracy of the children's nonword repetitions, Dillon, Cleary, et al. (2004) also found that several measures of speech perception, production, and memory were strongly correlated with the number of segments reproduced correctly.

In addition to reconfirming which speech and cognitive processes are most predictive of CI implant users' nonword repetition skills, segmental and suprasegmental linguistic analyses have also been very useful in qualitatively describing the nature of these listeners' nonword repetition errors. For example, Carter et al. (2002) found that children using CIs were able to produce the correct number of syllables and the correct stress patterns in nonwords with over 60% accuracy. However, children using

CIs have been found to be less accurate in producing individual segments in novel nonword patterns. Several segmental analyses by Dillon, Cleary, et al. (2004) revealed that less than 40% of target consonants were repeated correctly. When target consonants were incorrectly reproduced, it was most often due to a substitution of another consonant. Deletions of target consonants accounted for only 25% of the segmental errors. Despite the inability to produce most segments correctly, the children with CIs reproduced manner, place, and voicing of target consonants correctly over 50% of the time. Reproduction of the correct voicing of consonants was easiest for the CI children, while reproducing the correct manner was the most difficult.

Accuracy of nonword imitations was consistent across the various voicing and manner features. However, variability was observed in the reproduction of place features. Coronals were reproduced correctly in nonword responses nearly 70% of the time. However, labials were correct only about half the time and dorsals were only correctly produced 40% of the time. A detailed analysis of substitution errors indicated that labials and dorsals were frequently replaced with coronals (see Dillon, Cleary, et al., 2004).

Dillon, Cleary, et al.'s (2004) assessment of pediatric CI users' segmental accuracy in a nonword repetition task was one of the first to find this pattern of place of articulation errors. Previous research has suggested that children with CIs reproduce labial targets more accurately than other places of articulation (Tobey, Geers, & Brenner, 1994; Dawson, Blamey, Dettman, Rowland, Barker et al., 1995). The different pattern of place of articulation errors found in the children's nonword repetitions may have occurred because the nonword repetition task was conducted in auditory-only mode with no visual cues to place of articulation (Dillon, Cleary, et al., 2004). In an audio-visual speech perception task, cues to place of articulation are readily available and likely assist pediatric CI users when they are completing open-set word recognition tasks with familiar words, especially for labial segments (Lachs, Pisoni, & Kirk, 2001; Bergeson, Pisoni, & Davis, 2003; Bergeson & Pisoni, 2004). However, when only auditory information is available and when the test stimuli are unfamiliar nonwords like the ones used in these studies, children with cochlear implants must rely exclusively on their ability to encode the speech signal in its acoustic or auditory form prior to subvocally rehearsing and repeating it.

Although nonword repetition may rely more extensively on the initial auditory encoding of a speech signal than some other speech perception tasks that are closed-set or administered in live-voice with real words, performance on the nonword repetition task has also been found to be related to pediatric cochlear implant users' speech production and working memory skills (Carter et al., 2002; Cleary et al., 2002; Dillon, Burkholder et al., 2004; Dillon, Cleary, et al., 2004). Thus, one potential problem with linguistic analyses or perceptual accuracy ratings of the nonword repetitions of deaf children using CIs is determining whether the observed performance and errors are primarily related to auditory perception and encoding, to working memory, or to speech production problems. That is, it is uncertain whether the observed nonword repetition errors committed by pediatric CI users are due primarily to perceiving the nonword incorrectly, simply articulating it improperly, or inefficiently rehearsing and maintaining the novel nonword pattern in immediate memory.

Previous studies have documented that pediatric CI users have inefficient subvocal verbal rehearsal processes in auditory, auditory-visual, and visual-spatial working memory tasks (Burkholder & Pisoni, 2003; Cleary et al., 2001). Thus, it is no surprise that inefficiencies in subvocal verbal rehearsal may also carry over to the nonword repetition task. Using linear regression, Dillon, Burkholder and colleagues (2004) found that speaking rate which can be used as an index of subvocal verbal rehearsal speed, was the strongest predictor of CI children's nonword repetition ratings. However, closed-set speech perception and speech intelligibility were also found to be significant predictors of nonword

repetition ratings. In addition, the strength of these two predictors was nearly equal. Thus, despite identifying subvocal verbal rehearsal speed as the primary predictor of nonword repetition accuracy, the relative contributions of speech perception and speech production still remains unclear.

One way to attempt to investigate the impact of speech perception and production problems on pediatric CI users' nonword repetitions is to study nonword repetition performance in listeners with normal hearing and normal speech production who are exposed to auditory conditions similar to those experienced by pediatric CI users. Using an acoustic simulation that models the input of CIs provides a way to compare nonword repetition performance in pediatric CI users and listeners with normal hearing and speech production.

The present experiment was designed to identify the locus of the problems that pediatric CI users have with nonword repetition. In the present study, the locus of "disruption" on the nonword repetition task for the NH adults listening to speech processed through a CI simulator is already known. The adults' initial perception and encoding of the nonwords is disrupted due to the degraded nature of the stimuli. However, their working memory and speech production are intact and are not disrupted in the nonword repetition task. Alternatively, for the CI children it is not clear whether they primarily have impaired or disrupted speech perception, working memory, speech production, or some combination of these processes. Comparing these two groups may provide further insight into whether speech production is a significant contributor to pediatric cochlear implant users' poor nonword repetition skills or whether the differences are primarily perceptual or memory related.

In the present study, the relationship between nonword repetition accuracy ratings and measures of speech perception, working memory, and linguistic accuracy (segmental and suprasegmental) of nonword imitations were compared for a group of pediatric CI users and a group of NH adults. In addition, overall accuracy was compared across the two groups on measures of speech perception, working memory, and linguistic accuracy of nonword imitations. It is assumed that comparisons made between NH adult speakers and pediatric CI users will help determine whether speech perception, working memory, or speech production difficulties underlie CI children's performance on the nonword repetition task. NH adults have intact speech production and working memory skills, but in this particular task, they have disrupted/altered perception since they are asked to perceive severely degraded speech processed through a CI simulator. The pediatric CI users, on the other hand, potentially have disrupted speech perception, working memory, and speech production.

In order to tease apart these causes of the variation in nonword repetition skills of CI users, and more generally their atypical language learning abilities, several comparisons were made between the NH adults and the pediatric CI users. First, patterns of correlations between perceptual accuracy ratings of nonwords and measures of speech perception, working memory, and linguistic measures of the actual nonword productions were compared for the two groups. If the same relationships between these language processing skills and nonword repetition accuracy ratings are uncovered in these two groups of listeners it may indicate that developmental and clinical differences do not influence the relationship between the component processes (i.e. encoding, memory, and production processes) used to complete a nonword repetition task under spectrally degraded listening conditions. Second, the actual performance on these language tasks was compared for the two groups. It is presumed that if NH adults and pediatric CI users demonstrate similar patterns of segmental and suprasegmental accuracy in their nonword repetitions that speech perception and working memory rather than speech production play a more critical role in pediatric CI users' nonword repetitions skills.

A second goal of the current study was to confirm the validity of this method of CI simulation. Similar patterns of nonword repetition errors in NH adults listening to spectrally degraded speech and in pediatric CI users would suggest that acoustic simulations of CIs do sufficiently model the acoustic input heard by CI users and are therefore useful when studying the effects of degraded auditory stimuli on speech perception and other cognitive skills.

Methods

Participants

Twenty-four pediatric CI users and 18 NH adults participated in this study. The children were selected from a larger group of participants who took part in the Central Institute for the Deaf (CID) 'Cochlear Implants and Education of the Deaf' project in 1999 or 2000 (see Geers & Brenner, 2003). The children were between 8 and 9 years old. All but five were deaf at birth. The average duration of deafness prior to receiving a CI for the children was 3 years. The children had between 4 and 6 years of experience with their CI. The 24 pediatric CI users included in the current study were the children who provided a response to all of the 20 nonwords (see Dillon, Burkholder, et al., 2004).

The adult participants were undergraduate students enrolled in an introductory psychology course at Indiana University and were given partial course credit for their participation. The subjects indicated through self-report that they had no history of speech, hearing, language, or attentional disorders. A short hearing screening was also conducted to confirm that the adult subjects had normal hearing at the time of testing.

Stimulus Materials

Three sets of stimulus materials were used in this study. For the nonword repetition task, the stimuli included 20 nonwords taken from the Children's Test of Nonword Repetition recorded by a female speaker of American English (Gathercole et al., 1994). Table 1 lists the nonwords and their target transcriptions. This set of nonwords is balanced for number of syllables and is the same as was used in previous studies conducted in our laboratory (see Cleary et al., 2002; Dillon et al., 2004).

In addition, two tests of speech perception were included: the Lexical Neighborhood Test (LNT: Kirk, Eisenberg, Martinez, & Hay-McCutcheon, 1999) and the Word Intelligibility by Picture Identification (WIPI: Ross & Lerman, 1979). The LNT is an open-set spoken word recognition task. The LNT test contains words which vary in *lexical difficulty*. Lexically "easy" words are high frequency words in sparse lexical neighborhoods (having few phonologically similar words) (LNTe); lexically "hard" words are low frequency words in dense lexical neighborhoods (LNTh). The WIPI, on the other hand, is a closed-set spoken word recognition task that requires participants to point to a picture that matches the auditory stimulus, thereby placing no demands on the participant's speech production system.

Finally, both forward and backward auditory digit spans (Wechsler, 1991; 1997) were obtained to assess the participants' short-term and working memory skills. Forward digit spans test verbal rehearsal and short-term immediate memory. Backward digit span, on the other hand, is assumed to measure working memory and executive functions.

Number of Syllables	Target Nonword Orthography	Target Nonword Transcription
	ballop	ˈbæ.ləp
	prindle	ˈprɪn.dəl
2	rubid	ˈru.bɪd
	sladding	ˈslæ.ɾɪŋ
	tafflist	ˈtæ.fləst
	bannifer	ˈbæ.nə.fɚ
	berrizen	ˈbɛ.rə.zən
3	doppolate	ˈda.pə.ˌlet
	glistering	ˈglɪ.stɚ.rɪŋ
	skiticult	ˈskɪ.ɾə.ˌkʌlt
	comisitate	kə.ˈmi.sə.ˌtet
	contramponist	kən.ˈtræm.pə.ˌnɪst
4	emplifervent	ɛm.ˈplɪ.fɚ.ˌvɛnt
	fennerizer	ˌfɛ.nɚ.ˈrɑ͡ɪ.zɚ
	penneriful	pə.ˈnɛ.rə.fəl
	altupatory	æl.ˈtu.pə.ˌtɔ.ri
	detratapillic	di.ˈtræ.ɾə.ˌpɪ.lək
5	pristeractional	ˌprɪ.stɚ.ˈræk.ʃə.nəl
	versatrationist	ˈvɚ.sə.ˌtre.ʃə.nɪst
	voltularity	ˈval.t͡ʃə.ˌlɛ.rə.ti

Table 1. Nonwords used in the current study (adapted from Gathercole et al., 1994).

Prior to presentation to the adult listeners, the nonwords, LNT and WIPI words, and digit span lists were processed offline using a personal computer equipped with DirectX 8.0 and a Sound Blaster Audigy Platinum sound card. The signal processing procedure used for the cochlear implant simulation was adapted from real-time signal processing methods developed by Kaiser and Svirsky (2000). The signal was lowpass filtered with a cutoff frequency of 12,000 Hz. A bank of eight filters was then used to simulate the speech processing of an 8-channel cochlear implant. The output of each filter modulated noise bands of a higher frequency range than the corresponding filter. This mismatch was designed to represent a frequency misalignment that commonly occurs between the analysis filters of a cochlear implant's speech processor and the characteristic frequency of the neurons stimulated by the corresponding electrodes. The amount of frequency mismatch used in this model was equivalent to a 6.5 mm shift within the cochlea. For a more detailed discussion of the frequency shift used in the present study see Harnsberger, Svirsky, Kaiser, Pisoni, Wright, and Meyer (2001).

Procedure

Nonword Repetition. All listeners were given instructions that they would hear a funny made-up nonword and that they should try to repeat it as accurately as possible. The adult participants were also told that the nonwords would be acoustically degraded. Before hearing and repeating any nonwords in their degraded form, the adult listeners completed nonword repetition with five unprocessed practice nonwords. The degraded nonword stimuli were played in random order to the listeners over a tabletop speaker (Cyber Acoustics MMS-1) at approximately 70 dB(A) SPL.

The nonword repetitions obtained from each of the two groups were played to separate groups of naïve, NH adult listeners to obtain "perceptual accuracy ratings" (Burkholder, Pisoni, & Svirsky, 2004; Dillon, Burkholder, et al., 2004). Listeners heard the original target nonwords and then the response of either an adult or child speaker. Listeners were asked to rate how accurate they thought the participants' nonword responses were compared to the original target nonword. Ratings were made based on a 7-point Likert scale in which 1 corresponded to a repetition that "completely failed to resemble the target" and 7 corresponded to "completely perfect rendition of the target". All listeners received partial course credit for their participation.

Speech Perception and Memory Tests. The NH adults completed both the LNT and WIPI speech perception tests prior to nonword repetition and completed different lists of forward and backward digit spans in both degraded and clear auditory conditions after the nonword repetition task (Burkholder, Pisoni, & Svirsky, 2005). The CI children also completed the LNT and WIPI tests, as well as both forward and backward digit spans. The nonword repetition responses obtained from the children and adults were recorded onto a digital audiotape (DAT) via a head-mounted microphone (Audio-Technica ATM75).

Linguistic Transcription and Accuracy Scoring

All of the adult nonword repetitions were transcribed by two phonetically trained listeners (second and third authors). Any nonword responses that were composed of real words were not transcribed and were discarded. Consensus was reached on 284/291 (97.6%) of the responses. The remaining 7/291 (2.4%) were transcribed by a third phonetically trained listener in order to resolve any disagreements.

All nonword responses were aligned with the target transcription segment by segment to ensure the maximum continuity between the target and the response. Each segment in the response that corresponded to a segment in the target was coded for accuracy along several segmental dimensions. For consonants, the segments were coded for correct global place of articulation (labial, coronal, dorsal), sonorancy ([±sonorant]), manner (stop, affricate/fricative, nasal, liquid/glide), and obstruent voicing. For vowels, the segments were coded for correct height (high, mid, low), backness (front, central, back), and roundness (round and unround). In addition to these featural codings, segments were also coded for whether the response segment and the target segments matched along these dimensions simultaneously ("whole segment correct"). It is important to note that for "whole segment correct", the segments may not actually match exactly. For example, [θ], [s], and [ʃ] were coded as exactly correct since they match for global place (coronal), manner, and voicing.

In addition to segmental coding, nonword responses were coded along several suprasegmental dimensions: correct number of segments, correct number of consonants, correct number of syllables/vowels, and correct stress placement. For the NH adult speakers, correct stress was calculated

as follows. If either a primary or secondary stress in the response matched the primary stress in the target word, it was scored as correct. This method of scoring stress was utilized because several instances of the adult responses were observed in which the degree of stress (primary vs. secondary) was difficult to determine.

The pediatric CI users' nonword responses were transcribed using similar criteria (Carter et al., 2002). Their nonword responses were not retranscribed for the current analysis, but were recoded using the same segmental and suprasegmentals dimensions as the adult NH data to allow better comparison between the adult and child data. The only dimension that was not recoded was stress placement since the original transcription of stress for the child data differed from that of the adult transcriptions. Thus, no comparisons between the two groups of speakers were carried out for stress.

Results

Figure 1 displays the adults' and children's performance on the open- and closed-set speech perception tasks. Several ANOVAs were carried out to assess differences between listener groups and lexical difficulty. A repeated-measures ANOVA on the open-set LNT with lexical difficulty (easy vs. hard) as a within-subjects factor and listener group (CI children vs. NH adults) as a between subjects factor revealed a main effect of lexical difficulty ($F(1, 40) = 18.02$, $p = 0.000$). Lexically easy words were identified better than lexically hard words. The main effect of listener group was also significant ($F(1, 40) = 83.99$, $p = 0.000$). The CI children had much better LNT word recognition scores than the adults. The interaction was not significant ($F(1, 40) = 2.95$, $p = 0.094$). A one-way ANOVA of the CI children's and NH adults' closed-set WIPI scores revealed no significant differences between the two groups ($F(1, 40) = 1.92$, $p = 0.174$). Taken together, these two tests revealed that in closed-set word recognition tasks the two groups of listeners exhibited no significant differences, whereas they did in the open-set task. Interestingly, the children with CIs performed better than the NH adults on the LNT test.

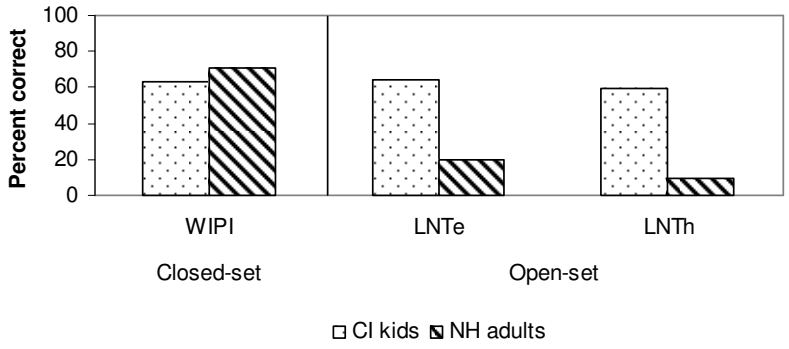


Figure 1. Performance on closed- and open-set speech perception tests by pediatric CI users and NH adults listening to an acoustic simulation of a cochlear implant.

Figure 2 shows the participants' performance on a forward and backward digit span task. The adults' digit span data reflect their performance on the digit span task when it was administered with the auditory tokens that were processed through the acoustic simulation of the cochlear implant. A repeated-measures ANOVA was conducted to determine the effects of digit span recall condition and listener

group. A main effect of recall condition was found (F (1, 40) = 59.29, p = 0.000). As expected, forward digit spans were higher than backward digit spans. The main effect of listener group was also significant (F (1, 40) = 36.41, p = 0.000). Digit span scores obtained from adults listening to the acoustic simulation of the cochlear implant were higher than the children's digit span scores. The interaction was not significant.

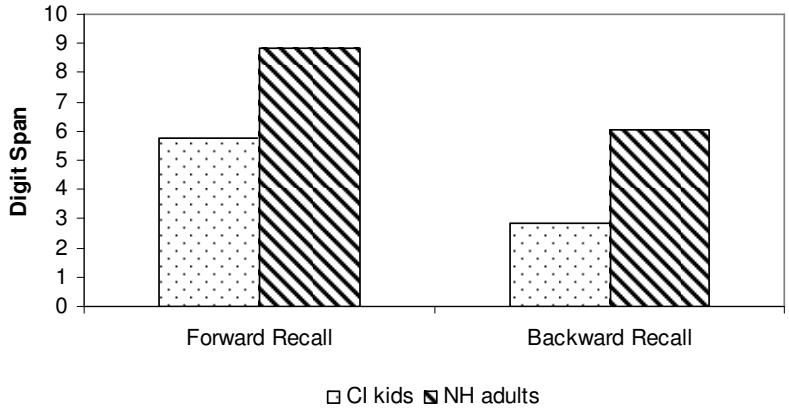


Figure 2. Forward and backward digit span scores of pediatric CI users and NH adults listening to an acoustic simulation of a cochlear implant.

Figure 3 displays the mean perceptual accuracy ratings assigned to each listener group based on the number of syllables in the nonwords. A repeated-measures ANOVA was conducted to determine the effect of syllable number and listener group. A main effect of syllable number was found (F (3, 40) = 9.53, p = 0.000). The effect of listener group was also significant (F (1, 40) = 10.65, p = 0.002). The children's nonword repetition accuracy ratings were higher than the adults. The interaction of syllable number and listener group also reached significance (F (3, 40) = 15.09, p = 0.000). This interaction indicates that the children's nonword repetition accuracy ratings decreased as the number of syllables in the nonwords increased. However, the adults' ratings remained constant across different word lengths.

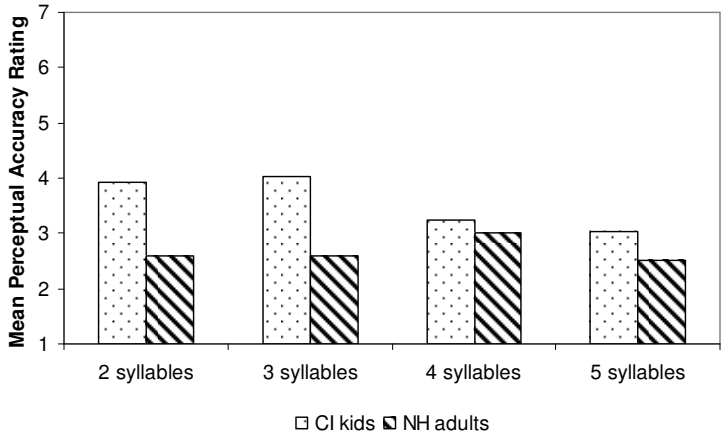


Figure 3. Mean perceptual accuracy ratings assigned to pediatric CI users and NH adults when repeating nonwords with 2, 3, 4, and 5 syllables.

Table 2 lists the results of bivariate correlations conducted using the listeners' nonword repetition accuracy scores, speech perception measures, and digit spans. The children's nonword repetition accuracy ratings were highly correlated with both the closed- and open-set speech perception tests. Children with higher scores on the LNTe, LNTh, and WIPI received higher nonword repetition accuracy ratings. The children's nonword repetition ratings were also strongly correlated with forward digit span scores, but not with backward digit span scores. This same pattern of results was obtained in the adults, although the magnitudes of the correlations were smaller. Positive correlations were found between the nonword repetition ratings and the three measures of speech perception, as well as between the ratings and the forward digit spans. As with the children, the correlations with backward digit span were weaker and did not reach significance.

	Perceptual Accuracy Ratings	
	CI Kids	**NH Adults**
Lexical Neighborhood Test (easy words)	**.71** (<.001)	**.62** (.006)
Lexical Neighborhood Test (hard words)	**.67** (<.001)	*.54* (.022)
Word Intelligibility by Picture Identification	**.69** (<.001)	*.57* (.013)
Forward Digit Span	**.77** (<.001)	*.56* (.016)
Backward Digit Span	.39 (.063)	.05 (.834)

Table 2. Bivariate correlations between the adults' and children's nonword repetition perceptual accuracy ratings and several speech perception tests and digit spans. *P*-values are provided in parentheses. Correlations with a *p*-value of .01 are considered significant (Bonferroni correction: .05/5). Significant correlations are indicated in bold. Correlations that approach significance are indicated in italics.

Several one-way ANOVAs were carried out on the linguistic analyses of the repetition responses in order to assess the differences in the accuracy in reproductions of the two groups. The results of these ANOVAs are presented in Table 3. The only variables that yielded significant differences between the two groups were obstruent voicing, vowel height, and vowel rounding. The CI children were more accurate in reproducing vowel height, whereas the NH adults were more accurate in reproducing obstruent voicing and vowel rounding. No other linguistic dimensions exhibited significant differences.

		Means		*p*-value
		CI Kids	NH Adults	
Supra-segmentals	Number of syllables/vowels	64 % (17)	74 % (8)	.082
	Number of consonants	32 % (18)	30 % (12)	.736
	Number of segments	29 % (19)	25 % (13)	.441
Segmentals	Place of Articulation	77 % (12)	73 % (4)	.302
	Sonorancy	83 % (8)	86 % (4)	.271
	Manner	75 % (11)	74 % (5)	.690
	Obstruent Voicing	**77 % (11)**	**83 % (4)**	.048
	Vowel Height	**71 % (9)**	**63 % (7)**	.004
	Vowel Backness	67 % (10)	67 % (9)	.903
	Vowel Rounding	**87 % (5)**	**92 % (4)**	.004

Table 3. Results of one-way ANOVAs comparing the average percent correct performance of CI kids and NH adults along several linguistic measures. Standard deviations are provided in parentheses. Significant differences are indicated in bold.

Table 4 lists the results of bivariate correlations conducted using the listeners' nonword repetition accuracy scores and the measures obtained from linguistic analysis of the nonword repetitions. Several differences between the two groups emerged when examining the correlations between linguistic measures and perceptual accuracy ratings. Although the CI children's nonword rating scores were found to be highly correlated with 10/11 of the linguistic measures, the adults' nonword accuracy ratings were only correlated with one of the eleven linguistic measures.

		Perceptual Accuracy Ratings	
		CI Kids	**NH Adults**
Supra-segmentals	Number of syllables/vowels	**.75** (<.001)	.28 (.262)
	Number of consonants	**.66** (<.001)	-.08 (.741)
	Number of segments	**.70** (<.001)	.07 (.783)
	Correct stress placement	**.60** (.002)	.23 (.360)
Segmentals	Place of Articulation	**.82** (<.001)	-.32 (.189)
	Sonorancy	**.80** (<.001)	.04 (.863)
	Manner	**.81** (<.001)	.22 (.382)
	Obstruent Voicing	**.74** (<.001)	-.02 (.944)
	Vowel Height	**.82** (<.001)	**.75** (<.001)
	Vowel Backness	**.60** (.002)	.20 (.426)
	Vowel Rounding	.30 (.152)	-.07 (.773)

Table 4. Bivariate correlations between the adults' and children's nonword repetition accuracy ratings and several linguistic measures. *P*-values are provided in parentheses. . Correlations with a *p*-value of .0045 are considered significant (Bonferroni correction: .05/11). Significant correlations are indicated in bold. Correlations that approach significance are indicated in italics.

Discussion

Several interesting and novel results emerged from the present analyses conducted on the nonword repetitions of pediatric CI users and NH adults listening to an acoustic simulation of a CI. Overall, the CI children had better spoken word recognition scores than the NH adults. This result would initially suggest that the deaf children with CIs performed better on the speech perception tests because they had more experience listening to spectrally degraded speech and/or because the spectrally mismatched speech that the adults were listening to was more degraded than the input from the children's devices. However, the children only performed better than the adults on the open-set LNT; no significant difference between the groups was found for the closed-set WIPI. This pattern suggests that the adults and children may have used different strategies when choosing responses in the closed-set speech perception task.

The closed-set WIPI is a 6-alternative, forced-choice test in which the five response alternatives were all minimal pairs or close neighbors of the target words which were all appropriate for use with children. Carrying out this task requires that listeners make discriminations between words based on the perception of fine acoustic-phonetic detail. The performance on the WIPI suggests that adults and children were able to make fine acoustic-phonetic discriminations. The LNT with hard words requires similar abilities in an open-set format since hard words have many acoustically similar neighbors. However, pediatric cochlear implant users performed much better than adults on this task. This pattern of results suggests that in the forced-choice task adults may have used a global pattern recognition strategy and linguistic knowledge to choose the correct response alternative. The adults' decision strategies and their more extensive linguistic knowledge and experience may have compensated for their overall poor speech feature discrimination abilities when they were completing the closed-set task. It is likely that the higher performance of CI children on the open-set test occurred because the children have more experience listening to a degraded speech input through their CI.

Developmental differences in working memory processes may also underlie the differences observed in the digit spans between the pediatric CI users and NH adults listening to the acoustic simulation of a CI. Given that the pediatric CI users had better speech perception scores than the adults it is unlikely that errors in speech perception were the primary cause of the children's poorer performance on the digit span task. In addition, earlier studies have found that pediatric CI users have shorter visual-spatial memory spans than NH children in a task in which no spoken response is required (Cleary, Pisoni, & Geers, 2001). These two findings suggest that speech perception and speech production problems are not the primary cause of pediatric CI users' shorter memory spans. Rather, slower memory processing strategies such as subvocal verbal rehearsal and scanning for these items in working memory may be the major factors contributing to the relatively short digit spans of pediatric CI users (Burkholder & Pisoni, 2003).

However, in NH adults listening to an acoustic simulation of a cochlear implant, it has been suggested that perceptual encoding errors, rather than memory processing errors, are responsible for shorter digit spans in spectrally degraded conditions (Burkholder et al., 2005). Taken together, the previous findings and the current results suggest again that adults' extensive linguistic experience and their more mature processing strategies can be used to compensate for perceptual and encoding difficulties that are the result of listening to spectrally degraded speech in speech perception or immediate serial recall tasks. Similarly, the present results suggest that the delayed memory processing strategies of pediatric CI users are not sufficient to compensate for auditory encoding problems.

The comparisons between the adults' and children's nonword repetition perceptual accuracy ratings also provide support for this proposal. Pediatric CI users' nonword repetition accuracy ratings were significantly higher than those assigned to the NH adults listening to the acoustic simulation of the cochlear implant. Moreover, the effect of syllable length was observed only in the children. The children's mean nonword repetition accuracy ratings decreased as the number of syllables in the nonwords increased. This suggests that limitations in the children's ability to rehearse and retain longer nonword sequences in phonological working memory is responsible for the syllable-length effect (Carter et al., 2002; Dillon, Burkholder, et al., 2004). The children's repetitions of the shorter nonwords may have been rated as more accurate than the adults' simply because the children had much more experience listening to degraded input. That the children's ratings decrease to that of the adults' for longer words further suggests that limitations to working memory are responsible for the syllable-length effects.

Similar interpretations of the nonword repetition syllable-number effect in NH children have been proposed by Gathercole et al. (1994). In addition, when NH adults complete the nonword repetition

task in clear listening conditions using unprocessed speech signals, they also demonstrate a syllable-number effect as a result of increased memory load (Gupta, 2003). The lack of the syllable-number effect with the NH adults listening to an acoustic simulation of a cochlear implant suggests that difficulty in encoding the degraded speech stimuli may have blocked or inhibited the used of normal phonological memory processes that contribute to the syllable-number effect. In addition, the current group of NH adults may have shown no evidence of the syllable-length effect because their nonword repetition performance and ratings were simply near or at the floor. This floor effect may have resulted because the adults in this study had very little experience listening to spectrally degraded speech compared to the pediatric CI users who have used their implant for several years before the present study. In addition, the adults may have performed poorly because of the large spectral mismatch used in the acoustic model of the CI which made the task perceptually harder.

Despite having what appears to be a near-floor performance and a reduced role of phonological working memory in the nonword repetition task, the NH adults' nonword repetition accuracy ratings were correlated with several speech perception measures and with their forward digit span. The same pattern of correlations observed in the adults listening to the acoustic simulation of the cochlear implant was also observed in the pediatric CI users. This is an important finding because it suggests that the pediatric CI users used the same fundamental component processes to carry out nonword repetition that NH adults use. They do not approach the task in a non-strategic or random manner. This finding may have implications for how pediatric CI users approach other tasks such as novel word learning (Houston, Carter, Pisoni, Kirk, & Ying, 2002). Because nonword repetition requires some of the same basic processing skills that novel word learning makes use of, the present results suggest that pediatric CI users may have more typical word-learning mechanisms than previously thought.

This current set of results is also interesting in light of earlier findings that in clear listening conditions, both NH children and adults demonstrate a relationship between immediate serial recall and nonword repetition (Gupta, MacWhinney, Feldman, & Sacco, 2003). The present study replicates these findings in pediatric CI users and NH adults listening to an acoustic simulation of a CI and suggests that being exposed to degraded auditory stimuli in these tasks does not cause this relationship to be atypical or dysfunctional.

The correlations observed between the accuracy ratings and the measures of speech perception and working memory in both groups indicate that listeners who have better perception and working memory perform better on the nonword repetition task. Previous work has shown that CI children's speech intelligibility scores, as measured by transcriptions of short sentences, also correlates with nonword repetition accuracy (Cleary et al., 2002). In the current study, the measures of speech production that we obtained were based on linguistic analysis and coding of actual productions scoring for both segmental and suprasegmental contrasts.

The lack of a correlation between any of the linguistic measures and the perceived accuracy of the nonword responses for the NH adults may be due to the lack of variability observed for both the linguistic measures and the perceived accuracy ratings. The NH adults were actually rated as having less accurate nonword responses than the children. This may be due to the fact that the adult responses were generally slow, labored, and disfluent, a fact not reflected in the linguistic transcriptions. Thus, the adult perceived accuracy ratings may have exhibited a floor effect and therefore less variability than the child data. We would expect that speakers who display poor articulation of nonwords would be rated as reproducing the target nonword less accurately as we found in the child data. However, there may simply have been insufficient variability in the adult data to capture this.

Another explanation for the difference in the ratings between the children and the adults may be due to the different expectations of speech production for children versus adults. Raters may have been more lenient in rating the children's productions simply because they are children. Furthermore, the adults' nonword productions were often disfluent and therefore the ratings may have been lower since the raters may have attended more to general naturalness than to differences in linguistic accuracy. A future study in which speakers produce multiple repetitions of a nonword stimulus in order to get more fluent imitations may eliminate this problem.

Because the linguistic production measures were not found to be significantly different for the CI children and the NH adults, this suggests that the CI children overall have good speech production skills. This finding is consistent with previous results showing that speech production skills do not independently contribute to the variance observed in the nonword repetition task (Dillon, Burkholder, et al., 2004). The absence of any differences in production accuracy between the CI children and NH adults suggests that perception and working memory are the primary loci for variation observed in the nonword repetition task for CI children. The lack of differences in the nonword responses further suggests that the acoustic simulation used here may sufficiently model the acoustic input heard by CI users. However, if both NH adults and CI children are basing their productions entirely on knowledge of phonotactics, segmental frequencies, and/or transitional segmental probabilities rather than on acoustic or spectral qualities, then other degradations of the input signal should produce similar results in the nonword repetition responses. This remains to be tested in future research.

Conclusions

In the current study, NH adults performed better than the children with CIs on tasks that required more advanced/developed working memory (digit span). The children with CIs, however, performed better on the open-set speech perception task, probably because of their greater experience in listening to degraded input. The atypical and delayed working memory skills of the CI children were also visible in the interaction between word length (number of syllables) and the perceptual accuracy ratings. Children were rated as less accurate when producing longer rather than shorter nonwords. The lack of a syllable-length effect for the adults could either reflect their better working memory skills or a floor effect of the perceptual accuracy ratings. When comparing the production accuracy (as measured by the various linguistic measures) no differences between the two groups of participants emerged, suggesting that their productions are comparable. Taken together, these results suggest that the locus of the difficulty in performing a nonword repetition task in CI children appears to be related to early perception and lies in memory and verbal rehearsal skills needed to maintain a representation in immediate memory. In other words, difficulties in performing this task are due to developmental differences.

If we consider the patterns of performance, the adults and the children often showed similar results. The perceptual accuracy ratings data for the CI children were found to be correlated with measures of speech perception (WIPI and LNT), short-term memory (forward digit span), and speech production (linguistic accuracy). Children with better performance on each of these components were rated as producing more accurate nonword imitations. The adults showed these same patterns of performance for the speech perception and working memory tasks, but not for the linguistic variables. The lack of a correlation with the linguistic measures reflected a floor effect which may be due to disfluent productions which were rated more poorly than the children's. The similarity of the nonword repetition scores for the CI children and NH adults suggests that the same basic component information processing operations are involved in the completion of the nonword repetition task.

References

Bergeson, T.R., Pisoni, D.B., & Davis, R.O. (2003). A longitudinal study of audiovisual speech perception by hearing-impaired children with cochlear implants. *The Volta Review, 103*(monograph), 347-370.

Bergeson, T.R., & Pisoni, D.B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 749-772). Cambridge, MA: MIT Press.

Bishop, D.V.M., North, T., & Donlan, C. (1996). Non-word repetition as a behavioural marker for inherited language impairment: evidence from a twin study. *Journal of Child Psychology and Psychiatry, 37,* 391-403.

Brady, S., Mann, V., & Schmidt, R. (1987). Errors in short-term memory for good and poor readers. *Memory & Cognition, 15,* 444-453.

Briscoe, J., Bishop, D.V.M., & Norbury, C.F. (2001). Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment. *Journal of Child Psychology & Psychiatry, 42,* 329-340.

Burkholder, R.A., & Pisoni, D.B. (2003). Speech timing and working memory in profoundly deaf children after cochlear implantation. *Journal of Experimental Child Psychology, 85*, 63-88.

Burkholder, R.A., Pisoni, D.B., & Svirsky, M.A. (2004). Perceptual learning and nonword repetition using a cochlear implant simulation. *Cochlear Implants: Proceedings of the 8th Annual International Cochlear Implant Conference,* 208-211.

Burkholder, R.A., Pisoni, D. B., & Svirsky, M.A. (2005). Effects of a cochlear implant simulation on immediate memory in normal-hearing adults. *International Journal of Audiology, 44,* 551-558.

Carter, A., Dillon, C., & Pisoni, D. (2002). Imitation of nonwords by hearing impaired children with cochlear implants: suprasegmental analysis. *Clinical Linguistics and Phonetics, 16*, 619-638.

Cleary, M., Dillon, C., & Pisoni, D. (2002). Imitation of nonwords by deaf children following cochlear implantation. *Annals of Otology, Rhinology, and Laryngology, 111*, 91-96.

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

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

Dawson, P.W., Blamey, S.J., Dettman, S.., Rowland, L.C., Barker, E.J., Tobey, E.., Busby, P.A., Cowan, R.C., & Clark, G.M. (1995). A clinical report on speech production of cochlear implant users. *Ear and Hearing, 16,* 551-561.

Dillon, C.M., Burkholder, R.A., Cleary, M., & Pisoni, D.B. (2004). Nonword repetition by children with cochlear implants: accuracy ratings from normal-hearing listeners. *Journal of Speech, Language, and Hearing Research, 47*, 1103-1116.

Dillon, C. M., Cleary, M., Pisoni, D. B., & Carter, A. K. (2004). Imitation of nonwords by hearing-impaired children with cochlear implants: segmental analysis. *Clinical Linguistics and Phonetics, 86*(1), 39-55.

Dillon, C.M., & Pisoni, D.B. (2004). Nonword repetition and reading in deaf children with cochlear implants. *International Congress Series, 1273,* 304-307.

Dillon, C.M., & Pisoni, D.B. (under revision). Nonword repetition and reading skills in children with cochlear implants. *Volta Review.*

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research, 47,* 421-436.

Edwards, J., & Lahey, M. (1998). Nonword repetitions of children with specific language impairment: exploration of some explanations for their inaccuracies. *Applied Psycholinguistics, 19,* 279-309.

Gathercole, S.E., & Baddeley, A.D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28,* 200-213.

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

Gathercole, S., Willis, C., Baddeley, A., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory, 2*, 103-127.

Geers, A.E., & Brenner, C. (2003). Background and educational characteristics of prelingually deaf children implanted by five years of age. *Ear and Hearing, 24,* 2S-14S.

Gupta, P., MacWhinney, B., Feldman, H. M., & Sacco, K. (2003). Phonological memory and vocabulary learning in children with focal lesions. *Brain and Language, 87*, 241-252.

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

Houston, D., Carter, A.K., Pisoni, D.B., Kirk, K.I., & Ying E. (2002). Word learning skills of profoundly deaf children following cochlear implantation: A first report. In *Research on Spoken Language Processing Progress Report No. 25* (pp. 35-61). Bloomington, IN: Speech Research Laboratory.

Kail, R., & Park, Y. (1994). Processing time, articulation time, and memory span. *Journal of Experimental Child Psychology, 57*, 281-291.

Kaiser, A. R., & Svirsky, M.A. (2000, October 15-18). *Using a personal computer to perform real-time signal processing in cochlear implant research.* Paper presented at the Proceedings of the IXth IEEE-DSP Workshop., Hunt, TX.

Kirk, K.I., Eisenberg, L.S., Martinez, A.S., & Hay-McCutcheon, M. (1999). Lexical neighborhood test: Test-retest reliability and interlist equivalency. *Journal of American Academy of Audiology, 10*, 113-123.

Lachs, L., Pisoni, D.B., & Kirk, K.I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report. *Ear and Hearing, 22*, 236-251.

Laws, G. (1998). The use of nonword repetition as a test of phonological working memory in children with Down Syndrome. *Journal of Child Psychology and Psychiatry, 39*(8), 1119-1130

Munson, B., Edwards, J., & Beckman, M. E. (2005). Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Journal of Speech, Language, and Hearing Research, 48*, 61-78.

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

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

Sahlen, B., Reuterskiold-Wagner, C., & Nettelbladt, U., & Radeborg, K. (1999). Language comprehension and nonword repetition in children with language impairment. *Clinical Linguistics and Phonetics, 13*, 369-380.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *Quarterly Journal of Experimental Psychology, 45A,* 21-50.

Tobey, E., Geers, A., & Brenner, C. (1994). Speech production results: speech feature acquisition. *Volta Review, 96,* 109-129.

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale – III.* San Antonio, TX: The Psychological Corporation.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Identification of Bilingual Talkers across Languages[1]

Stephen J. Winters, Susannah V. Levi and David B. Pisoni

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

[1] This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We would like to thank Christina Fonte, Jen Karpicke, and Melissa Troyer for their help in running subjects and editing stimuli.

Identification of Bilingual Talkers across Languages

Abstract. Two groups of monolingual, native English-speaking listeners were trained to identify the voices of ten German-English bilingual talkers. One group of listeners learned to identify the voices from English stimuli only, while the other group learned to identify the talkers from German stimuli only. After four days of training, both groups of listeners were asked to identify the same talkers from novel stimuli in both the language they had been trained on and the language they had not heard during training. No differences were observed in the amount of improvement in talker identification accuracy made by the two groups of listeners during training. In testing generalization across languages, however, the English-trained listeners performed significantly worse on German stimuli than they did on English stimuli, while the German-trained listeners identified talkers just as well from English stimuli as they did from German stimuli. This pattern of generalization across languages suggests that some of the indexical properties of speech are language-specific, while others are language-independent. The English-trained listeners apparently learned to identify talkers by relying on language-specific indexical information, while the German-trained listeners learned to identify talkers through language-independent indexical information. This pattern of results suggests that listeners may follow a mandatory perceptual strategy whereby they make use of language-specific indexical information when they can understand the language that is being spoken; otherwise, they learn to identify voices on the basis of language-independent information alone. This perceptual tendency may result from the influence of automatic linguistic processing on listener performance in a talker identification task that requires conscious control.

Introduction

Traditionally, linguists have distinguished between the linguistic and the indexical properties of speech (Abercrombie, 1967). Indexical properties of speech contain information about personal characteristics of the speaker—such as the speaker's age, gender, sociolinguistic background or personal identity—while linguistic properties carry information about the message the speaker is trying to convey. While both indexical and linguistic information is simultaneously transmitted to listeners in the same speech signal, the extent to which these properties of speech may interact with each other—either in the speech signal itself or in the process of speech perception—has long been a matter of debate. There are two competing models of how these types of information are processed in the perception of speech: the modular view and the integrated view. In brief, the modular view assumes that the indexical properties of speech and the linguistic properties of speech are processed independently of one another and do not interact in speech perception. The integrated view, on the other hand, holds that indexical and linguistic properties are inextricably bound to one another in speech and that they affect each other in language processing and other tasks.

Modular View

In first characterizing the distinction between the linguistic and indexical properties of speech, Abercrombie (1967) described the indexical properties as "extra-linguistic," and argued that information about the "medium" or the "source" of the message is not relevant to linguistic communication. This characterization implies that the perceptual process of recognizing a talker, or identifying some of that talker's personal characteristics, can operate independently of the process of perceiving the linguistic content of an utterance. The listener simply has to identify which properties of the signal derive from the

talker and which derive from the linguistic system of phonological contrasts. Similarly, other researchers have assumed that speech perception involves a process of "talker normalization" (see Pisoni, 1997 for a review) which strips away the talker-specific information in speech and yields linguistic representations that are abstract and talker-independent (Halle, 1985). This "modular" view of speech perception holds that the process of identifying the linguistic content of a spoken utterance essentially involves identifying those linguistic properties of the signal which are independent of the talker.

There is clear evidence from both behavioral and neurological studies that the linguistic and indexical properties of speech can be processed independently of one another. For example, listeners can identify the linguistic content of spoken messages that are largely devoid of talker-specific information. Several studies have shown that listeners can identify talkers from time-reversed samples of speech, the linguistic content of which is unintelligible (Bricker & Pruzanksy, 1968; Clarke, Becker, & Nixon, 1966; Williams, 1964). The same independence of talker and linguistic information has also been found, to a lesser extent, in filtered speech (Compton, 1963; Pollack, Pickett, & Sumby, 1954) and whispered speech (Pollack, Pickett, & Sumby, 1954; Williams, 1964). Phonagnosia, a phenomenon in which neurologically-impaired listeners can comprehend spoken utterances in a language that they know but cannot identify the voices of familiar talkers, also provide converging evidence that the linguistic processing of speech can take place independently of talker recognition (Van Lancker, Cummings, Kreiman, & Dobkin, 1988).

Other behavioral studies have shown that voice and linguistic information appear to be processed in different parts of the brain. In an early study of hemispheric specialization, Landis, Buttet, Assal, and Graves (1982) found that listeners utilize both hemispheres in voice recognition, whereas there was a distinct advantage of the left hemisphere for linguistic tasks (e.g., word recognition). In one experiment, Landis et al. played monosyllabic consonant-vowel words into either the right or the left ear, and asked listeners to press a button every time they heard a specific target word. The listeners' reaction times showed a clear right-ear advantage (REA) for this linguistic task. In a second experiment, listeners were asked to push a button when they heard a particular male or female voice. For this task, no clear advantage for one ear over the other was found. Instead, the results revealed a REA when the target voice was female, but a left-ear advantage (LEA) when the target voice was male. Landis et al. interpreted these results by appealing to the fact that higher frequencies have been shown to elicit a REA and that female voices, with their higher fundamental frequency, may therefore also be processed with a REA. However, the stimuli used in the word recognition task were all presented in a female voice, so the REA found in that condition may have been due to the higher fundamental frequencies inherent to the stimuli, rather than a language-specific processing preference in the brain.

Kreiman and Van Lancker (1988) reported evidence of a dissociation between linguistic and indexical processing using a dichotic listening paradigm. In this paradigm, listeners heard different words played simultaneously in both ears. Each word was spoken in a different voice, selected from a database of fifty different famous male voices that the listeners knew. The listeners were asked to attend only to the stimulus in one ear or the other, and wrote down both the word that was played in that ear and the person who said the word. The listeners showed a clear REA in the word recognition task, but there was no significant advantage for either ear in the voice identification task.

More recent studies have been able to isolate voice processing to more specific brain regions. Glisky, Polster, and Routhieaux (1995) tested elderly listeners' ability to recall either the content or the voice of previously heard sentences. They found that listeners with high frontal lobe function outperformed those with poor frontal lobe function on the voice task, but there were no differences between these two groups in their performance on the sentence content task. Conversely, listeners with

high medial temporal lobe function outperformed listeners with low medial temporal lobe function in the sentence content task, but there were no differences between these two groups of listeners on the voice task. More recently, using functional magnetic resonance imaging (fMRI), Stevens (2004) found that distinct brain regions were involved in voice- and word-discrimination tasks. Stevens presented pairs of words to listeners and asked them to either determine whether the talkers of the words were the same or whether the two words themselves were the same. Stevens found that the voice comparison task resulted primarily in activation in the right fronto-parietal area, whereas lexical processing was associated with the left frontal and bilateral parietal areas. These results indicate that, to some extent, the processing of voice information takes place independently of the processing of linguistic information, in a different part of the brain.

Taken together, these behavioral and neurological findings suggest that there is a double dissociation between linguistic comprehension and talker recognition: both processes can, in certain circumstances, operate independently of one another. Furthermore, when listeners are asked to attend to voice characteristics of a speaker, they appear to utilize different areas of the brain than when they focus on the linguistic content of a spoken message.

Integrated View

Other researchers have proposed that the linguistic and indexical properties of speech are closely coupled, both in the speech signal and in the process of speech perception. Figure 1, reproduced from Hirahara and Kato (1992), illustrates how both sources of information are encoded in an integrated fashion in the speech signal. The spacing between adjacent formants provides information about the vowels a talker has produced, while the absolute values of the same formants provide information about the talker's voice. Global acoustic-phonetic properties of the speech signal, like the values of vowel formants, may therefore be considered both "linguistic" and "indexical."

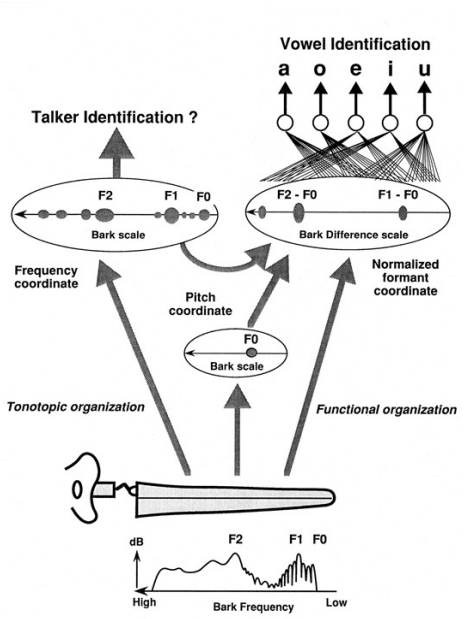


Figure 1. Schematic of acoustic properties which specify both linguistic and indexical information in the speech signal (from Hirahara & Kato, 1992).

More importantly, several studies have shown that the linguistic and indexical properties of speech interact with each other in perception. This interaction is bidirectional in nature: indexical properties affect linguistic processing and linguistic knowledge affects the processing of indexical information.

Indexical Information Affects Linguistic Processing. The influence of indexical information on linguistic processing has been shown in a series of studies that have systematically varied the number and type of voices used to produce the stimuli for linguistic processing tasks. Varying voice information in this way typically results in worse performance on the linguistic processing tasks. Mullennix and Pisoni (1990) first showed this effect by asking listeners to categorize a set of stimuli that varied along two different perceptual dimensions in a speeded classification task. Listeners either had to decide whether a target word began with a "p" or a "b" (i.e., a linguistic distinction) or whether the word was spoken by a male or a female talker (i.e., an indexical distinction). In one condition, Mullennix and Pisoni presented both “p” and “b” words to the listeners in a variety of different voices. In a control condition, all words were presented to listeners in the voice of a single speaker. Mullennix and Pisoni found that reaction times for the linguistic classification task were slower when the stimuli were presented in several different voices than when the stimuli were presented in just one voice. This result indicated that indexical information was not strictly “extra-linguistic” or irrelevant to linguistic processing. Instead, listeners had to take talker-specific voice information into account when performing a phoneme classification task.

In a series of studies, Goldinger (1996) showed that listeners exhibit a same voice advantage when performing a word recognition memory task. Goldinger first asked listeners, in a study phase, to perform a word recognition task, in which they typed a word that they heard presented in noise. In a test phase, the listeners then heard a series of words and were asked to indicate whether or not they had heard that word before in the word recognition task. Half of the words in this test phase were repeated items, and, of these, half were presented in the same voice as they had been presented in the study phase. Goldinger found that listeners more accurately identified test items as being repeated from the study phase when they were presented in the same voice in which they had been presented initially than when they were presented in a different voice.

Other studies have also shown that the indexical and linguistic properties of speech encoded and stored together in representations of spoken words in memory. This results in a “same-voice advantage” effect, whereby spoken word tokens are processed more efficiently and accurately by listeners when they hear those words spoken in the same voice as they have heard in past experiences. For instance, Goldinger, Pisoni, and Logan (1991) found a same-voice advantage effect in a list recall task. In this study, listeners first heard a list of 10 words and were subsequently asked to recall the list. Goldinger et al. varied both the number of voices in which list words were presented and the rate at which stimuli were presented. They found that, at fast presentation rates, lists with multiple talkers were recalled less accurately than lists that were spoken by only a single talker.

Schacter and Church (1992) found a similar same-voice advantage effect in a stem completion task. They initially presented words to listeners in a variety of voices and asked the listeners to rate how pleasant each word token sounded. Later, listeners were presented with the first syllables of those words in noise and were asked to write down the first word that came to mind. Listeners responded more often with words they had heard in the initial phase of the experiment when the words were re-presented in the same voice as the initial presentation than when they were presented in a different voice.

In a continuous recognition memory experiment using spoken words, Palmeri, Goldinger, and Pisoni (1993) played long lists of words to listeners and asked them to determine whether each word was an "old" word (one that had been previously heard) or a "new" word (one that had not been previously heard). In order to assess the effects of voice on the processing of words, half of the old words were presented again in the same voice and half were presented again in a different voice. As in the previous studies, listeners responded more quickly and accurately when old words were repeated in the same voice.

Several studies have also shown that familiarity with a set of talkers' voices can facilitate the processing of the linguistic content of novel messages produced by those talkers. Nygaard, Sommers, and Pisoni (1994) explored how voice familiarity aids linguistic processing by first training listeners to identify ten previously unfamiliar talkers, from individual spoken words, over a period of ten days. After training, listeners were tested on their ability to identify novel words spoken in noise by either the talkers they had learned to identify or by a set of unknown talkers. Nygaard et al. found that the listeners identified a significantly higher percentage of the novel words correctly when they were spoken by familiar talkers. In a follow-up experiment, Nygaard and Pisoni (1998) showed that this advantage of talker familiarity applies not only to individual words, but to sentence-length utterances as well.

Linguistic Knowledge Facilitates Indexical Processing. Not only does knowledge of the indexical properties of speech affect language processing, but linguistic knowledge also affects the processing of indexical information. Several studies have shown that the inability to understand the linguistic content of speech hinders talker identification. Thompson (1987) had native English-speaking participants listen to a paragraph read in either English, Spanish, or Spanish-accented English by a target talker, and then asked the listeners to identify the target talker from among six different voices after a one-week delay. Thompson found that listeners could identify talkers best in the English language condition, followed by the Spanish-accented English condition, and worst in the Spanish language condition. Goggin, Thompson, Strube, and Simental (1991) followed up on this study by presenting Spanish and English stimuli to both monolingual English-speaking and monolingual Spanish-speaking participants in a similar testing paradigm. They found that both groups of listeners were poorer at identifying the voice of the target talker when they did not understand the language.

It has also been shown that the facilitatory effect that knowledge of a language has on the ability to identify talkers extends to a listener's second language, as well. Listeners who have studied a target language as a second language (L2) identify voices in that language better than listeners who have no knowledge of the language (Schiller & Köster, 1996; Köster & Schiller, 1997; Sullivan & Schlichting, 2000). In particular, Schiller and Köster (1996) showed that listeners with no knowledge of German were significantly worse at identifying a target talker, speaking in German, than both L2 listeners and native German listeners. Interestingly, Schiller and Köster found that the L2 and native German listeners did not differ from each other in talker identification accuracy. Sullivan and Schlichting (2000) further showed that the extent to which listeners are familiar with a second language does not affect their ability to identify talkers, so long as they have some knowledge of the language. They found that L2 learners of Swedish all performed significantly better than listeners with no knowledge of Swedish in a talker identification task, but that the amount of exposure the listeners had to the second language (ranging from first year learners to fourth year learners) did not affect their ability to identify Swedish talkers. Sullivan and Schlichting also reported, however, that L2 learners did not reach the same level of proficiency in identifying Swedish talkers as native Swedish listeners did in their earlier study (Schlichting & Sullivan, 1997), though no statistics were presented to corroborate this claim.

Schiller, Köster, and Duckworth (1997) have shown that the facilitatory effect of language knowledge on talker identification disappears when the linguistic content of the signal is eliminated, in reiterate speech. Schiller et al. had German speakers read a passage using only the syllable [ma] and then tested native German listeners, native English listeners, and L2 learners of German attempt to identify the speakers of those passages. They found that the native German listeners did not perform any better at this task than either the L2 learners or the native English listeners, implying that the advantage that native listeners have over non-native listeners in identifying talkers in a given language disappears once of the linguistic content of spoken utterances has been removed.

Summary: Previous Research. The studies reviewed in this section suggest that linguistic and indexical information are closely coupled in the processing of speech. Strong effects of voice were observed in tasks which, on the surface, do not appear to rely on indexical or voice properties—such as word recognition or phoneme discrimination. Familiarity with a talker's voice was also found to facilitate a listener's ability to process the linguistic content of speech. Likewise, listeners can process spoken utterances that they have heard before more efficiently and accurately when they are presented again in the same voice than when they are presented in a different voice. Furthermore, listeners can identify talkers' voices more accurately when they know the language in which an utterance is spoken.

Current Study. The results of previous research showing that language knowledge facilitates the ability of listeners to identify talkers are confounded by the fact that all of these studies changed talkers between language conditions. Since both the linguistic and the indexical properties of the stimulus materials changed between language conditions in these studies, it is not clear whether the listeners' diminished performance in the unfamiliar language condition was due to their lack of knowledge of the linguistic properties of the unknown language or their lack of knowledge of whatever language-specific indexical properties the unfamiliar language might have. It is also unknown whether listeners can identify familiar talkers who are speaking in an unfamiliar language. That is—are the indexical properties that listeners use to identify a familiar voice in one language the same properties of speech that can be used to identify that voice in another language?

In order to investigate these questions, the current study was designed to investigate the ability of listeners to identify bilingual talkers, while they were speaking in two different languages. Listeners were first trained to identify the voices of these bilingual talkers while they were speaking in one language, and then tested on their ability to identify the same talkers while they were speaking in the other language. Any potential change in talker identification accuracy between language conditions would thus be due to the change in language, rather than any change in the specific talkers producing the stimuli. By separating the contributions of language and talker to the spoken test materials in this way, the present experiment provides a much stronger test of the extent to which the linguistic and indexical properties of speech interact with each other in the process of talker identification.

The modular view holds that the indexical properties of speech are extra-linguistic, and do not vary from language to language. If this is the case, then the indexical and linguistic properties of speech should not interact with one another in the process of talker identification. The language that a talker is speaking should not affect the ability of listeners to identify that talker's voice because that talker's indexical contribution to speech will remain constant from one language to another. In this experiment, listeners should therefore be able to generalize all of their knowledge of the bilingual talkers' voices across languages; they should be just as good at identifying voices in the language that they have been trained on as they are at identifying the same voices in a language they have not heard before.

On the other hand, the integrated view holds that the linguistic and indexical properties of speech are closely coupled and interact with one another in the process of talker identification. In this case, the properties of speech that listeners use to identify a talker's voice differ from one language to another. The language that a talker is speaking should therefore affect the ability of listeners to identify that talker's voice. If listeners rely on language-specific indexical properties when learning to identify a talker's voice, they should not be able to identify the same talker's voice as well in an unfamiliar language, which may exhibit a different set of language-specific indexical properties. It should therefore be difficult for listeners to generalize their knowledge of the bilingual talkers' voices completely across languages in the proposed experiment. Instead, the listeners should be able to identify talkers more accurately when they are speaking in the language that they have been trained on than when they are listening to the talkers in an unfamiliar language.

The integrated view of speech perception does not preclude the possibility that some indexical properties might be language-independent, or shared across languages. Thus, some of the listeners' knowledge of talkers' voices should generalize across languages; i.e., their ability to identify a known set of talkers in an unfamiliar language should be better than their ability to identify a set of unknown talkers in a familiar language. It is, of course, possible to take an even stronger view of the extent of integration between the linguistic and indexical properties of speech and propose that all indexical properties are specific to the language which is being spoken. If this is the case, then there should be no generalization of talker knowledge across languages in a talker identification experiment such as this one, since whatever listeners know about what a talker's voice sounds like in one language would not hold for that same talker's voice in a different language. There is, however, little existing evidence or rationale for this strong theoretical standpoint to suggest that such results might emerge from this experiment, but it is worth considering here as a benchmark.

Methods

Stimulus Materials

Twelve female and ten male German L1/English L2 speakers who were living in Bloomington, IN, were recorded in a sound-attenuated IAC booth at the Speech Research Laboratory at Indiana University. Productions were recorded using a SHURE SM98 head-mounted unidirectional (cardioid) condenser microphone with a smooth frequency response from 40 to 20,000 Hz. Productions were digitized into 16-bit stereo recordings via Tucker-Davis Technologies System II hardware at 22050 Hz and saved directly to an IBM-PC Pentium I computer. Each speaker produced a single repetition of 360 English words and 360 German words. Each word was of the form consonant-vowel-consonant (CVC) and was selected from the CELEX English and German databases (Baayen, Piepenbrock, & Gulikers, 1995). German was selected as the second language in the experiment because it not only had a sufficient number of CVC words—which had the same phonotactic structure as the English CVC words—but also because there were uniformly calculated frequency counts for both the English and German sets of words in the CELEX database. Speakers read each word as it was presented to them on a computer monitor. Before each presentation, an asterisk appeared on the screen for 500 ms, signaling to the speaker that the next trial was about to begin. This was followed by a blank screen for 500 ms. After this delay, a recording period began which lasted for 2000 ms. The target word was presented on the screen for the first 1500 ms of this recording period. After the conclusion of the recording period, the screen went blank for 1500 ms, and then an asterisk appeared again to signal the beginning of the next recording cycle. The presentation of production items was blocked by language, but all within-language items were randomized for each speaker. Items that were produced incorrectly or too quietly were noted and re-recorded in the same manner following each recording block. The total recording time for each language

block was approximately one hour for each speaker. Speakers were given the option of recording both sets of language items on either the same day or on two separate days. All speakers elected to record all stimuli in a single recording session. The recording session took approximately two hours, and speakers were paid $10 an hour for their time.

This process yielded recordings which were uniformly 2000 ms long. Since the actual productions of the stimulus word in each recording were always shorter than 2000 ms, the silent portions in the recording before and after each production were removed by hand using Praat sound editing software. All edited tokens were then normalized to have a uniform RMS amplitude of 66.499 dB.

Words from both languages varied in frequency based on counts from the CELEX database. Words varying in frequency of occurrence were included in the stimulus materials because listeners can identify high frequency words more quickly, and from less acoustic information, than low frequency words (Grosjean, 1980). We expected listeners to pay more attention to the acoustic/phonetic details of the low frequency words, and therefore develop a more robust mental representation of the acoustic/phonetic characteristics of the various talkers' voices from these tokens. For the purpose of analysis, the English words were divided into three equal groups of varying frequency. The 120 lowest frequency words all had a CELEX frequency count of less than or equal to 96, while the 120 highest frequency words all had a frequency of greater than or equal to 586. The remaining 120 words thus all had frequency counts between 96 and 586. The frequency count of homophones (e.g., rite, write, right) was taken to be the frequency count of the most frequent homophone; this homophone was also the word that the speakers were presented with during the recording sessions.

Ten speakers were selected as the training voices, based on their native language background and perceived nativeness in English. Speakers with southern German (N= 2), Austrian (N=3) and Romanian German (N=1) dialects were excluded from the set of training voices, along with speakers with self-reported speech or hearing disorders (N=2), and one speaker who did not finish the recording session. Of the remaining speakers, only the five male and five female speakers who were, on average, rated as being the least accented talkers (Levi, Winters, & Pisoni, 2005) were used in the talker identification training study.

Listeners

All listeners were native English-speaking students at Indiana University in Bloomington, Indiana. None reported any knowledge of the German language prior to participation in the study. None of the listeners had ever lived in Germany or had any German-speaking friends or family members. All were right-handed and reported no known speech or hearing impairments at the time of the study. Participants were paid $75 for their participation in the study. A total of 54 listeners participated in the study. Half were trained on English language stimuli, and half were trained on German language stimuli.

The response data from only 40 of these listeners was included in the statistical analysis of the results. Two of the listeners in the English training condition and four listeners in the German training condition did not complete the experiment. The data from listeners who did not correctly identify at least 40% of the talkers in 4 or more evaluation phases during training were also excluded from analysis. We considered 40% correct identification accuracy to be a reasonable level of performance for establishing that listeners had learned the talkers' voices during training, since 30% correct was significantly better than chance performance in each evaluation phase (excluding cross-gender confusions). Four participants did not meet this criterion in the English language group and two did not meet this criterion in the German language group.

There were twenty-one listeners in both language conditions who both completed the experiment and met the criterion for learning during the evaluation phases. In the English training group, 10 of these listeners heard the English language stimuli in the first generalization testing phase, while 11 heard the German language stimuli first in generalization. The data from the last participant who heard the German stimuli first in generalization was excluded from the statistical analysis, in order to balance the numbers between generalization block groups. Similarly, in the German training condition, 11 of the remaining listeners heard the English language stimuli first in generalization testing, while the other 10 heard the German stimuli first in generalization. The data from the last participant who heard the English stimuli first in generalization was excluded from the statistical analysis.

Procedure

Participants were trained and tested in a quiet room. During training, each participant wore Beyer Dynamic DT-100 headphones while sitting in a front of a PowerMac G4. All stimuli were presented to participants over the headphones via a customized SuperCard (version 4.1.1) stack, running on the PowerMac G4.

Training. Participants were trained to identify the ten different bilingual voices, by name, in eight training sessions spanning four days. The methodology used in these training sessions closely followed the methodology first developed by Nygaard, Sommers, and Pisoni (1994). Each training session consisted of seven distinct phases, which are summarized in Table 1.

Phase	Stimuli	Task
Familiarization #1	A set of five words, produced by all ten talkers	Listen and attend to voice/name pair
Re-familiarization #1	One word, produced by all ten talkers	Listen and attend to voice/name pair
Recognition #1	Sets of five different words for each talker, presented twice, in random order	Identify speaker of each word (feedback is provided)
Familiarization #2	A set of five words, produced by all ten talkers	Listen and attend to voice/name pair
Re-familiarization #2	One word, produced by all ten talkers	Listen and attend to voice/name pair
Recognition #2	Sets of five different words for each talker, presented twice, in random order	Identify speaker of each word (feedback is provided)
Evaluation	Sets of ten different words for each talker, presented once, in random order	Identify speaker of each word (no feedback provided)

Table 1. Summary of stimuli and tasks used during each phase in all training sessions.

In the familiarization phases, listeners heard a sequence of five words produced by each of the ten talkers. These words were the same for all ten talkers, and were presented one at a time. There was an inter-stimulus interval of 500 milliseconds between the presentation of each word. As each word was presented to the listener, the name of the talker who had produced the word was shown on the computer screen. Each talker had a name that was a common male or female name in both English and German. Each name was also presented in a unique and consistent color, in a unique and consistent position on the

screen. The layout of all ten names is shown in Figure 2. During this phase of training, participants did not respond to what they heard, but were instructed to pay attention to the names on the computer screen and listen to the sound of each talker's voice.

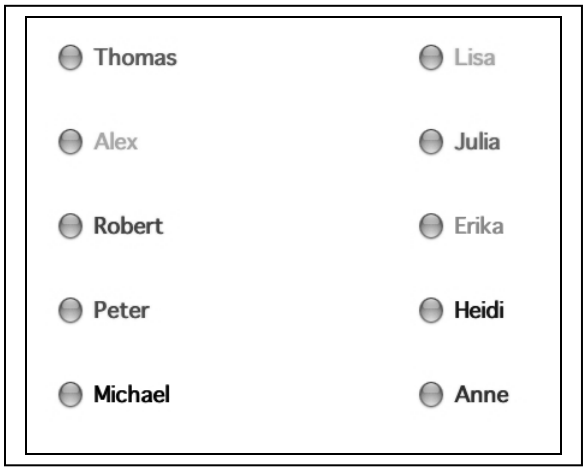


Figure 2. Layout of the ten talker names used in the experiment. (Names were in the following colors: Thomas, light blue; Alex, orange; Robert, red; Peter, purple; Michael, black; Lisa, green; Julia, dark pink; Erika, grey; Heidi, dark blue; Anne, brown).

After each familiarization phase, listeners underwent a brief re-familiarization phase in which they heard only one word spoken by all ten talkers. The same word was spoken by all ten talkers during re-familiarization. The participants did not register any response to the word but, again, were instructed to pay attention to both the name of the talker and the sound of the talker's voice.

In the recognition phases, listeners heard five different tokens, presented twice, from all ten talkers. Each word was presented in isolation to the listeners, whose task was to identify which talker had spoken each word. Participants identified talkers by clicking an on-screen button next to the appropriate talker's name (see Figure 1). After participants registered their responses, they received feedback, after a 333 millisecond interval, by hearing the stimulus token again, while the name of only the correct talker was presented to them on the computer screen. After receiving this feedback information, the listeners clicked an on-screen button to hear the next stimulus. The entire recognition phase was self-paced.

After the first recognition phase, listeners repeated the entire sequence of familiarization, re-familiarization, and recognition phases prior to beginning the evaluation phase. During the evaluation phase, listeners heard ten different words each from all ten talkers. As in the recognition phases, participants heard each word in isolation and were instructed to identify which talker had produced the word immediately after they heard it. Listeners did not, however, receive any feedback during the evaluation phase. Instead, they heard the next stimulus immediately after they registered their response to each stimulus.

The entire sequence of seven phases in each training sessions took most participants approximately 35 minutes to complete. Participants underwent two training sessions on each day of training, over the course of four days. Participants were required to take a short (approximately five minute) break between consecutive sessions on each day of training.

Generalization. After four days of training, all listeners participated in generalization testing on the fifth day of the experiment. Generalization testing began with two brief familiarization phases. In the first familiarization phase, listeners heard the same three words produced by all ten talkers. In the second, re-familiarization phase, the listeners heard the same word produced by all ten of the talkers. All of the words that were presented to the listeners in these familiarization phases were spoken in the same language that the listeners had heard during training. After the re-familiarization phase, the listeners were once again tested on their ability to identify the talkers from individual spoken words, in a series of two testing phases. The procedure used in these testing phases was identical to that used during the evaluation phase of each training session. Listeners heard one word at a time and were instructed to identify which talker had spoken the word. They received no feedback on their responses and were immediately presented with the next stimulus 500 milliseconds after registering their responses. The stimuli presented to the listeners in the two different generalization phases were in different languages. In one phase, the listeners heard words spoken in the language they had been trained on during the first four days of the experiment, while, in the other phase, they heard words spoken in the language they had not been trained on during the first four days of the experiment. Before testing, the listeners were instructed that the talkers might be speaking in an unfamiliar language. The order in which language blocks were presented in these two phases was counterbalanced across participants. For each participant, no more than two days intervened between any successive training days or the generalization test.

Stimulus Selection. The stimuli that were presented during training and generalization were independently selected for each listener from the larger set of individual word tokens in the bilingual talker database. For each listener, 100 words, balanced for frequency in each language, were first selected at random for use in the generalization testing blocks on the final day of the experiment. All 100 words that listeners heard in both generalization testing phases had thus never been presented before to the listeners during training. These 100 words consisted of ten different words spoken by all ten talkers, for both language blocks. No word, that is, was presented to listeners in more than one talker's voice during generalization.

After selecting out 100 words from the bilingual database for use in generalization, another 100 words were selected at random out of the remaining 260 items in the database, for each listener, for use in the familiarization and re-familiarization phases during training. These words were also balanced by frequency. Twelve of these items were presented to the listeners during each training session: five during the first familiarization phase, one during the first re-familiarization phase, five during the second familiarization phase, and one more during the second re-familiarization phase. Ninety-six of these words were thus presented to the listeners over the course of the eight training sessions, with the final four being presented to the listeners during the brief familiarization and re-familiarization phases prior to generalization testing (3 words and 1 word in these phases, respectively). No word that was presented during familiarization or re-familiarization was ever presented during generalization testing or in either the recognition or testing phases of the training sessions. The words selected for familiarization and re-familiarization were always in the same language as those words presented to the listener during the other phases of the training sessions.

The remaining 160 words in the talker database were presented to each listener exclusively during the evaluation and recognition phases of the training sessions. These words were also balanced by frequency. For each training session, 20 words from this collection of 160 were selected at random for each talker for presentation to a particular listener. Five of these words were presented twice during the first recognition phase, while another five were presented twice during the second recognition phase. Talker-specific word tokens were presented more than once during these recognition phases because it

has been found that feedback does not facilitate perceptual learning unless the stimulus items that participants receive feedback on in a training paradigm are presented to them more than once (Winters, Levi & Pisoni, 2005). The remaining set of 10 items in each collection of 20 were then presented to the listeners, without repetition, in the evaluation phase of each training session. Over the course of the eight training sessions, then, listeners heard all 160 words as produced by all ten talkers. Within the evaluation and recognition phases of any given training session, however, listeners heard different sets of words produced by each talker. It was possible, therefore, for there to be overlap between the sets of words produced by each talker in any recognition or evaluation phase. In both the recognition and evaluation phases, all word tokens from all talkers were presented at random to the listeners, with the stipulation that no individual word was ever presented on consecutive trials.

Results

Training

A two-way, repeated measures Analysis of Variance (ANOVA) was run on the response data from the evaluation phases of the eight training sessions. This ANOVA investigated the effects that training session (1, 2, 3, 4, 5, 6, 7, 8) and training language (English, German) had on the percentage of talkers correctly identified in each testing phase. Training session was a within-subjects factor while training language was a between-subjects factor. The ANOVA revealed a significant main effect of training session (F (7,32) = 61.637; $p < .001$), but no effect (at the $p < .05$ level) of training language, nor any interaction between training session and training language.

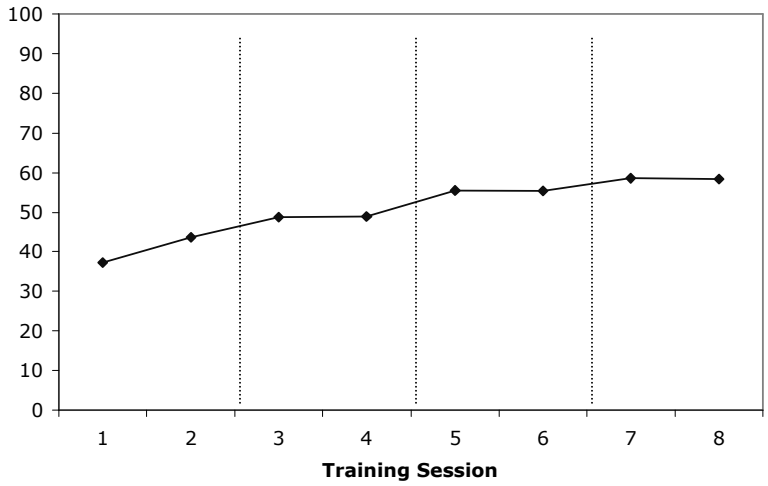


Figure 3. Percentage of talkers correctly identified, by all listeners, in the testing phase of each training session. Dotted lines denote breaks between separate days of the experiment.

Figure 3 shows the percentage of talkers that were correctly identified in the evaluation phases of each training session. Post-hoc, paired samples t-tests indicated that both groups of listeners consistently improved in identification accuracy over the duration of training. This improvement occurred in a step-wise fashion, however. Identification accuracy was significantly higher in training session two than in training session one ($p < .001$). Accuracy was also significantly higher in training session three than in training session two ($p = .002$). After session three, however, significant increases in identification

accuracy were only made between separate days of training. For instance, between sessions four and five—which occurred on days two and three of training, respectively—listeners' average identification accuracy improved from 48.8% to 55.2% ($p < .001$). Likewise, identification accuracy significantly improved between sessions 6 and 7 ($p = .007$), which occurred across days three and four of training. Within a particular day of training, however, listeners did not significantly improve in identification accuracy between sessions ($p > .825$).

Generalization

A three-way, repeated measures Analysis of Variance (ANOVA) was run on the response data from just the generalization testing phases on the final day of the experiment. This ANOVA investigated the effects that testing language (English, German), training language (English, German), and generalization block order (trained language first, trained language second) had on the percentage of talkers correctly identified in each generalization testing phase. Testing language was a within-subjects factor while training language and generalization block order were between-subjects factors. The ANOVA revealed a significant main effect of testing language (F (1,36) = 27.687; $p < .001$), where accuracy was significantly better for English stimuli than German stimuli. There was also a significant interaction between testing language and training language (F (1,36) = 47.864; $p < .001$). All other main effects and interactions did not reach significance at the $p = .05$ level.

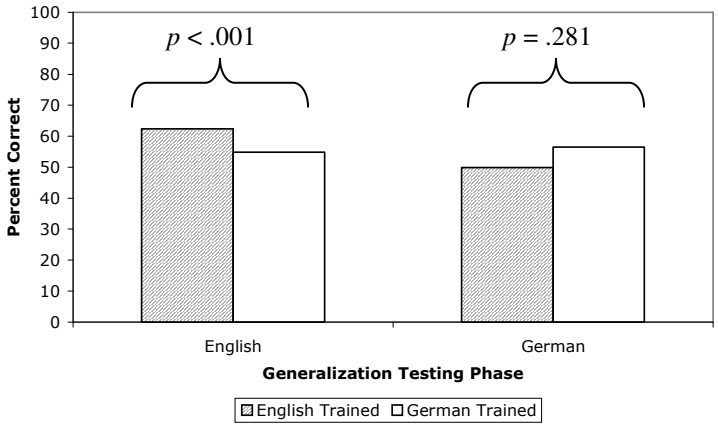


Figure 4. Percentage of talkers correctly identified, by each training group of listeners, in both generalization testing phases.

Figure 4 shows the percentage of talkers correctly identified, by each group of listeners, in the two generalization testing phases. Post-hoc analysis of the significant testing language by training language interaction indicated that the English-trained listeners demonstrated significantly higher talker identification accuracy on the English generalization block than on the German generalization block ($p < .001$). The German-trained listeners, on the other hand, did not perform significantly better on the English generalization block than on the German generalization block ($p = .281$). In comparing results across listener groups, post-hoc tests revealed that the German-trained group performed significantly better than the English-trained group on the German generalization block ($p = .049$), while the English-trained group performed significantly better on the English generalization block ($p = .016$).

Combined Data

In order to assess the extent of generalization from training to novel stimuli, paired samples t-tests were conducted comparing the listeners' level of performance between each training session and the two generalization testing phases. Figure 5 shows the percentage of talkers correctly identified by each training group in both training and generalization.

For the English-trained listeners, there were no significant differences in talker identification accuracy between the English generalization block and the evaluation sessions on the final day of training (p = .095 for session seven and p = .071 for session eight). These listeners' performance on the English generalization block was, however, significantly better than their performance on the first six training sessions (p > .01 in all cases). The English-trained listeners' performance on the German generalization block, on the other hand, was not significantly different from their performance on the third and fourth evaluation sessions, both of which took place on day two of training (p = .779 and p = .826). Their accuracy in German generalization was significantly better than their identification accuracy on day one of training (p < .01, for both sessions), but significantly worse than their identification accuracy on days three and four of training (p < .01, in all cases).

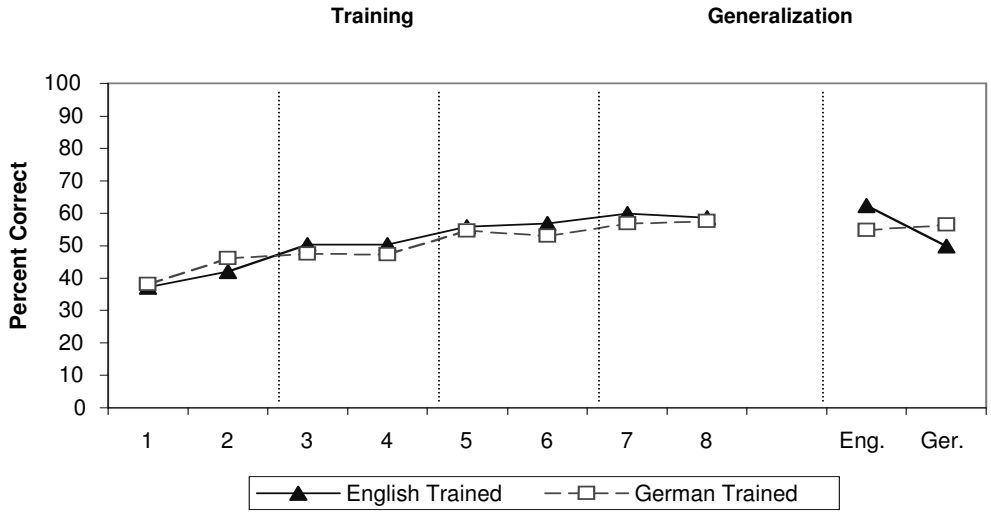


Figure 5. Percentage of talkers correctly identified, by each group of listeners, in the evaluation phase of each training session and in both generalization language blocks. Dotted lines denote breaks between separate days of the experiment.

Paired-samples t-tests also showed that the percentage of talkers correctly identified by the German-trained listeners in both generalization blocks was not significantly different than the percentage of talkers they correctly identified in training sessions five, seven and eight (p > .1 in all cases). The German-trained listeners did identify a significantly higher percentage of talkers in German generalization than they did in training session six (p = .038), but there was no significant difference between their performance in English generalization and in training session six (p > *.1*). Otherwise, their performance in both generalization blocks was significantly better than in all evaluation phases on the first two days of training (p < .001 in all cases).

Lexical Frequency

The words that listeners heard during the evaluation phases were split into three frequency groups of equal size: 120 low frequency words (with frequencies ranging from 0-96), 120 mid-frequency words (97-588) and 120 high frequency words (greater than 588). The lexical frequency of the words presented during the evaluation phases of each training session did not significantly affect the ability of the listeners to identify the talkers who spoke them. The percentage of talkers that listeners identified correctly from low frequency words was 62.5%, while the corresponding percentages for the mid and high frequency groups of words were 59.4% and 64.2%, respectively. Paired samples t-tests revealed that none of these means were significantly different from one another ($p > .08$ in all cases).

Discussion

Perceptual Learning during Training

The initial identification data from the training sessions showed that the training paradigm did, in fact, enable the listeners learn to identify the bilinguals' voices. Although a small minority of participants had some difficulty with the task, the majority of listeners significantly improved between the first and last training sessions in talker identification accuracy. The pattern of improvement in identification accuracy exhibited by these listeners was not consistent from one training session to the next, however. Listeners significantly improved in talker identification accuracy between the first and second training sessions, which were both on the same (first) day of training. After the second session, however, overall talker identification accuracy only improved significantly between sessions that took place on consecutive days. This pattern of learning suggests that some form of consolidation of what the listeners had learned took place in between consecutive training days, probably as a by-product of sleep (cf. Fenn, Nusbaum, & Margoliash, 2003). This pattern of improvement also suggests that listeners reached a learning plateau after the first training session on days two, three and four of the study, since they could no longer improve in their ability to identify talkers in the second training session on each of those days.

Generalization: Effect of Training Language

The identification data from the generalization testing sessions showed that listeners could generalize their knowledge of the talkers' voices to novel stimuli in both languages. Both listener groups demonstrated that they could identify the bilingual talkers speaking in an untrained language at a significantly higher level than they could identify that same group of talkers, in the trained language, on the first day of training. For both groups of listeners, talker identification accuracy on novel stimuli in their trained language also did not decrease from their level of performance in the final training session. Listeners were thus able to generalize all of their knowledge of the talkers' voices to novel words within a language, and at least part of their knowledge of the talkers' voices across languages.

The extent to which the listeners could generalize their knowledge of the talkers' voices across languages depended on the language in which they had been trained to identify those voices. The listeners who had been trained on English stimuli significantly better talker identification accuracy from novel English words in generalization than they did from novel German words. With the novel English stimuli in generalization, these listeners performed just as well as they had on the English stimuli presented to them on the final day of training. With German stimuli, however, their identification accuracy decreased to a level equivalent to their performance on the second day of training (in sessions 3 and 4). Since the English-trained listeners' performance levels were still significantly better than their

accuracy level on the first day of training, their performance on the novel German words indicates that they were able to generalize some of what they had learned about the bilingual talkers' voices to a novel language. The German-trained listeners, on the other hand, showed complete generalization of talker knowledge across languages. These listeners performed just as well on the novel English stimuli in generalization as they had on the German stimuli presented to them in the final day of training. They also performed just as well on the novel German stimuli in generalization as they had on the final day of training. Thus, there was no decrease in performance when these listeners made the transition from training to novel stimuli in either language.

The inability of the English-trained listeners to generalize completely across languages suggests that some of the indexical properties that they used to identify talkers in training were language-specific. Since the German language lacks the language-specific indexical properties of English, the English-trained listeners were not able to use these properties to identify talkers when they heard the set of novel German stimuli in generalization. By the same token, the fact that the English-trained listeners were able to generalize some of their knowledge of the talkers' voices across languages suggests that they attended to and encoded some language-independent indexical properties during training, as well. Taken by itself, the pattern of generalization exhibited by the English-trained listeners thus conforms to the predictions made by the integrated model of speech perception: some indexical properties of speech are language-specific, while others are language-independent.

The generalization data from the German-trained listeners, however, appear to best fit the predictions made by the modular view of speech perception. The German-trained listeners showed complete generalization of their knowledge of the talkers' voices across languages, indicating that they had learned to identify the talkers' voices on the basis of language-independent indexical properties during training. Since this information—whatever it might consist of—does not change between English and German, the listeners did not suffer any decrease in identification performance when they were presented with novel stimuli in the English language in generalization testing.

Training Results: Non-effect of Training Language

Although the language in which the listeners were trained affected how successful they were in generalizing to an untrained language, it did not affect the time course of perceptual learning during training itself; the amount of improvement in identification accuracy made by the participants during training in this study did not differ between the English-trained and German-trained groups of participants. The ability to understand the words that were presented to them in training did not, therefore, seem to provide the English-trained listeners with any additional advantage in the process of learning the talkers' voices. This finding conflicts with the results of earlier studies showing that it is easier for native English listeners to identify non-native talkers of English than talkers who are speaking a language other than English (e.g., Thompson, 1987; Goggin, Thompson, Strube, & Simental, 1991).

An effect of language on talker identification accuracy may not have emerged during the training portion of this study because its materials and methods were fundamentally different from those used in previous research. In voice line-up tasks, listeners are initially exposed to a short passage of speech from one talker and then asked to identify that talker, out of a variety of response options, on later trials. In the present study, listeners were familiarized with a larger set of talkers' voices before being tested on their ability to identify each talker from individual spoken words. The more involved training paradigm used in this study may therefore have reduced the native-language advantage that existed during the testing phase of previous voice identification studies. This advantage may also have been attenuated by the fact that only individual words were presented to listeners in this study. Nygaard and Pisoni (1998) have

shown that it is easier to identify talkers from sentence-long utterances than from individual word stimuli. Listeners can presumably use the higher-level semantic, syntactic and prosodic information in sentence- and paragraph-length utterances to identify talkers more easily when they can understand the language that is being spoken. With sentence stimuli, listeners also receive a longer sample of speech from a talker. Thus, listeners in the earlier voice line-up studies may have performed better on native-language stimuli because they had access to both sentence-length stimuli and higher-level linguistic information. Since the stimuli in this experiment lacked sentential and prosodic cues, however, the language in which the stimuli were spoken may have had less of an effect on talker identification accuracy during training. Finally, it is also possible that differences in talker identification accuracy during training were diminished because the same set of talkers was used in both language conditions. Although the finding of earlier research that talker identifiability is influenced by language has been replicated in several earlier studies, all of these studies consistently used different sets of talkers for different language conditions, and may therefore have confounded differences in the inherent distinctiveness of the voices with language-based difference in talker identification accuracy.

Training Results: Non-effect of Frequency

The lexical frequency of the English words that listeners heard during training also did not affect their ability to identify talkers. The fact that the talker identification task did not require the listeners to access the lexicon may account for the absence of frequency effects on talker identification accuracy. Listeners could simply interpret each stimulus item in acoustic-phonetic terms without relying on higher-level lexical information to help them perform the task. The fact that the English-trained listeners did not need to access lexical information in order to perform the talker identification task may also account for their failure to perform at a higher level than the German-trained group, since they evidently never accessed linguistic information at a more abstract level than what the German-trained listeners could pick up from the acoustic/phonetic surface structure of the speech stimuli alone.

Such perceptual tendencies may actually have been helpful for training purposes, if paying more attention to the acoustic-phonetic properties of speech facilitates the learning of talker identity. Any effect that lexical access might have on a talker identification task could be tested in future research by requiring listeners to write down (or type) each word that they hear, before identifying the person who spoke it. Incorporating these additional processing operations into the talker-learning task could have a variety of effects on listeners' performance. The listeners might find it easier to identify talkers when they are speaking high frequency words, because this would make it easier for the listeners to access the lexical information necessary to do the word identification task. Conversely, listeners might do better when they are listening to low frequency words, because that would require them to pay more attention to the acoustic/phonetic details of the signal, and also require them to do more lexical processing before they are able to identify the word. Finally, the increased processing load may result in a stronger memory trace for both the low-frequency word and the talker who produced it (cf. Luce, Feustel, & Pisoni 1983).

Training Results: Poorer Performance than in Nygaard, Sommers, and Pisoni (1994)

Listeners in this study also did not ultimately reach the same level of performance as the listeners reported in Nygaard, Sommers, and Pisoni (1994) did, even though the talker identification training paradigm in Nygaard et al. served as the basis for the one used in this study. Listeners in Nygaard et al. (1994) had to correctly identify 70% of the talkers on the final day of training in order to be included in the word identification transfer test on the final day of that experiment. About half of Nygaard et al.'s listeners were able to reach this criterion after nine days of training (18 out of 38). In the present study, however, only 10 of the 52 participants (six in the English-trained group, four in the German-trained

group) were able to correctly identify 70% of the talkers correctly in any evaluation session. The criterion for inclusion in this study was therefore reduced to 40% correct performance during testing in at least four different training sessions.

Our criterion was set lower than Nygaard et al.'s because it was not necessary, for the purposes of this study, to establish that the ability to identify talkers could facilitate a linguistic perception task such as word recognition in noise. We were only concerned with the extent to which listeners could generalize what they had learned about talkers' voices in training to novel stimuli, in different languages, on the last day of testing. All that was crucial to the success of this investigation, therefore, was that the listeners demonstrate that they were able to learn to identify the talkers' voices. Improvement to over 40% correct identification seemed to be a minimally satisfying demonstration of each listener's ability to have learned something about the various talkers' voices, since significantly better than chance performance in each training session was 30%. With this reduction in the criterion, only six out of 48 listeners (12.5%) failed to meet it after eight training sessions.

The poorer performance of the listeners in this study may be due to several methodological differences between this study's training paradigm and the one that was used in Nygaard et al.'s study (1994). Nygaard et al. trained listeners in nine sessions over nine separate days. Listeners in this study, however, participated in only eight training sessions, which took place over four days. The pattern of improvement during training in this study suggests that, after the first day of the experiment, it was necessary for listeners to sleep between training sessions in order for them to improve their performance. This result is consistent with the recent finding of Fenn, Nusbaum, and Margoliash (2003) that the perceptual learning of synthetic speech is enhanced by periods of sleep in between training sessions. For this reason, listeners did not show significant gains in talker identification accuracy between training sessions on the same day, after the first day of the experiment. The listeners may have been able to make such advances in identification accuracy between all eight training sessions, however, if those training sessions had all taken place on separate days. Spacing out the training cycles in this way could have enabled their performance to improve to the same level as that of the participants in Nygaard et al.

The listeners in Nygaard et al. (1994) also learned to identify the voices of native English talkers while, in this study, listeners learned to identify the voices of native German talkers who were speaking either English or German. Previous research by Goggin, Thompson, Strube, and Simental (1991) and Thompson (1987) has shown that English listeners have more difficulty identifying non-native speakers of English than native speakers of English. Hence, the native language of the speakers in this study may have contributed to the listeners' comparatively poorer level of performance in the talker identification task. However, some voices may also be simply more perceptually distinctive than others, regardless of their origin. It is thus possible that the voices of the talkers in Nygaard et al. just happened to be more distinctive than the ones used in this study, making the voice identification task easier for their listeners than it was for ours.

Generalization: Alternative Accounts

Although the modular theory of speech perception accounts most gracefully for the German-trained group's generalization data, it is possible to construct an alternative account of this pattern of generalization in which the German-trained talkers learned to identify talkers using German-specific indexical properties. Since the generalization data from the English-trained group indicates that such language-specific information exists in English, similar language-specific information probably exists in the German language, as well. It is possible that the German-trained listeners in this study used such language-specific indexical information in learning to identify the talkers during training, in combination

with the same language-independent indexical information that was available to the English-trained listeners. The German-trained listeners may then have found it easier to generalize their knowledge of the talkers' voices to the English language stimuli because they were already familiar with the language-specific indexical properties that are unique to English (from native language experiences before the experiment). By combining this language-specific information with the language-independent indexical properties they had learned during training, the German-trained listeners could have identified the talkers' voices just as well in the English language generalization condition as they did in the German language condition.

The lack of an effect of language on talker identification accuracy during training argues against this interpretation of the generalization data, however. If the native English-speaking listeners did use German-specific indexical properties to identify talkers in the German language training condition, it should have taken these listeners some time to familiarize themselves with the novel indexical properties of the German language. There should, in other words, have been a gap in performance between the two training groups—at least for the first few training sessions—while the German-trained group learned how to make use of the German-specific indexical information. No such gap was observed in the training results, however, suggesting that the German-trained group used only language-independent information right from the beginning of the experiment to perform the talker identification task.

That the German-trained listeners might not have used German-specific indexical information to identify talkers is not surprising, because they had no knowledge or experience with the German language prior to the experiment. The English-trained listeners did know English before the experiment, however, and apparently relied extensively on what they knew about this language to help them perform the talker identification task in training. It is interesting to note, however, that this information evidently did not help the English-trained listeners perform any better in training than the German-trained listeners, who were using only language-independent information. The use of English-specific indexical information only affected the performance of the English-trained group by making it more difficult for them to generalize their knowledge of the talkers' voices to a novel language. As such, it is possible that the language-specific information English-trained listeners attended to during training did not actually help them perform the talker identification task. Instead, they may simply have been unable to ignore the irrelevant linguistic information in the signal—as long as they could understand it—possibly reflecting a failure of executive function and cognitive control (Schachar & Logan, 1990; Barkley, 1997).

Under this alternative interpretation, listeners engage in the linguistic processing of speech automatically—when they can understand the language that is being spoken—while talker identification is a non-automatic process that requires conscious attention and control. Mandatory linguistic processing may therefore affect the controlled process of talker identification in the same way that, for instance, the automatic process of reading words affects the controlled process of naming colors in the well-known Stroop Effect. Stroop (1935) had participants name the color in which different words were printed. Stroop found that participants named these colors more slowly when the word itself was the name of a different color than the ink in which it was printed. The information that the participants extracted from automatically processing the orthographic representations of the words thus interfered with the slower, controlled process of naming the color of the ink in which the word was printed. It has been shown that this interference effect is reduced, however, when the words are presented to participants in upside-down text, and therefore cannot be read them in an automatic fashion (Liu, 1973).

Analogously, linguistic information may have "interfered" with the process of talker identification in this experiment, when the listeners were presented with words in English and could therefore process them in an automatic fashion. Under these conditions, listeners may have based their

talker identification judgments on irrelevant linguistic information in the training stimuli. This linguistic information may therefore not be "integrated" with indexical information in the speech signal itself. Instead, the two sources of information may only become confused with one another during the process of speech perception. Interference between linguistic and indexical information would not occur when listeners cannot understand the linguistic content of the words automatically, as in the German language training condition. Without interference from linguistic information, the German-trained listeners would be able to process the indexical properties of speech in a more language-independent fashion than the English-trained listeners. The German-trained listeners' representations of the talkers' voices in memory would therefore be more robust and language-independent—and could generalize better across languages—than the English-trained listeners' representations of the same voices. Similar effects of linguistic information interfering with indexical processing have recently been found in a same/different voice discrimination task (see Levi, Winters, & Pisoni, 2006).

Summary of Interpretation

Participants in this study appear to be following a general perceptual strategy in which they make use of language-specific indexical information when it is available to them, regardless of what consequences that strategy might have for the generalizability of their perceptual representations for particular talkers. When listeners are identifying talkers who are speaking in a language they know, those listeners are able to process the indexical properties of speech in an integrated manner. When listeners are identifying talkers who are speaking in a language they do not know, however, those listeners process the indexical properties of speech in a modular, language-independent manner. Learning to identify voices in a modular fashion—on the basis of language-independent information only—makes it easier for listeners to generalize their knowledge of talkers' voices to new languages. Relying on language-specific information to identify talker's voices makes such generalization more difficult, but listeners do it anyway, when that information is available to them. Processing speech in an integrated manner, that is, apparently pre-empts the processing of speech in a modular fashion. Only when linguistic or indexical information is blocked in the speech signal—e.g., when listeners hear an unfamiliar language, are presented with filtered speech, or are suffering from phonagnosia—do listeners revert to a modular form of speech processing, which can operate without both forms of information in the speech signal.

Future Research

By training German-English bilingual listeners in the same voice learning paradigm, it should be possible to test whether a decrease in talker identification accuracy across languages is due to a reliance on language-specific indexical properties or to an unfamiliarity with the language being generalized to. If all listeners automatically rely on language-specific indexical properties to identify talkers who are speaking a language they know, then bilingual listeners should rely on language-specific indexical properties in both the German and English language training conditions. These listeners should therefore have difficulty generalizing their knowledge of the talkers' voices from one language to another, regardless of which language they have been trained in. If incomplete generalization across languages is the result of unfamiliarity with the language being generalized to, however, then bilingual listeners should exhibit no drop-off in talker identification accuracy in going from either English to German or from German to English in generalization, since they are familiar with both languages (and their attendant set of language-specific indexical properties).

Future research might also determine whether integrated linguistic and indexical information might facilitate performance across languages in either linguistic or indexical tasks, as well as hinder it. In this study, evidence for an interaction between the linguistic and indexical properties of speech came

from a significant decrease in performance by the English-trained listeners when they were tested on German stimuli in generalization. In this case, a reliance on language-specific information in training made the talker identification task more difficult when the listeners were required to generalize their knowledge of the talkers' voices to a different language. Past research, however, has indicated that the interaction between the linguistic and indexical properties of speech can also have facilitatory effects on linguistic tasks such as the recognition of words in noise. Nygaard, Sommers, and Pisoni (1994), for instance, found that listeners can identify novel words in noise better when they are spoken by talkers that those listeners have learned to identify, instead of talkers that those listeners have not heard before. This result has been taken as evidence that the linguistic and indexical properties of speech are not only integrated in perception, but that knowledge of language-specific indexical information is stored in memory and can facilitate the ability of listeners to carry out linguistic tasks.

Assuming a modular view of speech perception, however, it is possible that knowledge of the language-independent properties of a talker's voice might facilitate listeners' performance in a linguistic task. The more familiar listeners are with a particular talker's voice—in any given language—the easier it might be for them to filter out the indexical properties of a person's voice when attempting to identify the (talker-independent) linguistic properties of a word that person has spoken. It should be possible to test these alternative views of the relationship between the linguistic and indexical properties of speech in word recognition by training a group of listeners to identify the voices of German-English bilinguals from German stimuli only, and then testing those listeners on their ability to identify words in noise spoken by both the talkers they have learned to identify and an unfamiliar group of talkers. If language-independent knowledge of a talker's voice facilitates word recognition—as in the modular view—then listeners who have learned to identify a talker from German words only should be able to better identify English words spoken by talkers they have learned to identify. If only knowledge of language-specific indexical properties facilitates performance in a linguistic task, however, then the German-trained listeners should not improve in their ability to recognize English words spoken by either familiar or unfamiliar bilinguals. A study of this kind is currently under way in our laboratory.

Conclusions

The present study investigated the extent to which the linguistic and indexical properties of speech are processed independently of one another by testing the ability of listeners to identify bilingual talkers' voices across two different languages. The extent to which listeners were able to generalize their knowledge of the bilinguals' voices from one language to another was considered within the context of two different views of speech perception. On the basis of the modular view of speech perception, which holds that linguistic and indexical information in the speech signal are processed independently of one another, in separate, perceptual channels, we predicted that listeners would be able to completely generalize their knowledge of the talkers' voices from one language to the other. However, on the basis of the integrated view of speech perception, which holds that the indexical properties of speech differ from language to language, we predicted that listeners would only show partial generalization of their knowledge of the talker's voices from one language to the other.

The results of this study suggest that listeners use both language-specific and language-independent indexical properties of speech. Listeners who were trained to identify bilinguals while they were speaking English showed incomplete generalization of their knowledge of the talkers' voices when they were asked to identify the same group of talkers while they were speaking German. In contrast, listeners who were trained to identify bilinguals while they were speaking German showed complete generalization of their knowledge of the talkers' voices when they were asked to identify the same group of talkers while they were speaking English. The English-trained group thus relied, in part, on indexical

properties that were specific to the English language in order to perform the voice identification task, while the German-trained group relied strictly on language-independent indexical information that could generalize across both languages.

Which features of a speaker's voice are language-independent, and which features are language-dependent? It may be assumed that the shape of a talker's vocal tract, nasal cavities and articulators have reliable effects on the acoustic output of that talker's speech, regardless of which language the talker is speaking. Rose (2003) points out, however, that the acoustic consequences of such "compulsory" features of a talker's voice may, in actuality, be very difficult for listeners to distinguish from one talker to another. Rose (2003) suggests, instead, that what makes talker's voices sound perceptually distinctive are the "chosen" features of their speech, which are under the talker's control to manipulate as he or she sees fit. In post-experiment debriefings, the participants in this study cited a number of different acoustic properties that they consciously listened for in attempting to identify each talker's voice. These features included qualities such as the pitch of the speaker's voice or the speed (i.e., the duration) with which a talker produced each word (e.g., some speakers consistently used a low pitch range or a high pitch range, while one female consistently produced each item with a very short duration). Such features of speech—while not necessarily "compulsory" aspects of a person's voice—could easily transfer from one language to another in bilingual talkers. A listener in a study such as this one could therefore identify a talker on the basis of perceiving such low-level acoustic qualities in either English or in a language with which they were not familiar, such as German. These "chosen" features of vocal identity might thus be considered "language-independent", so long as the languages that talkers are speaking do not require them to change acoustic characteristics such as pitch or duration in systematic ways (as in, for example, tone languages).

It is likely that listeners were also able to pick up on certain language-independent features of talkers' voices that were more complex than the basic acoustic properties of the speech signal. For instance, one listener, following the experiment, claimed that she could reliably identify one talker by the way she had "overexaggerated" the pronunciation of each word—in other words, by the fact that she had consistently hyperarticulated (Lindblom, 1990). Another listener claimed that she could consistently identify one of the male speakers by the fact that he sounded "gay." Such broad, phonetic features of a talker's voice may fall under the general rubric of a talker's "articulatory setting" (Rose, 2003). They could provide the listener with reliable, cross-linguistic cues to a talker's identity insofar as talkers are not required to change their articulatory settings by the phonetic rules of any given language.

On the other hand, phonetic markers of social identity (including sexual orientation, gender, class, regional affiliation, etc.) would be expected to change between languages—even two languages which are as phonetically similar as English and German. These phonetic attributes of the speech signal could thus serve as language-specific indexical properties. For instance, one phonetic marker of social identity which would almost certainly <u>not</u> transfer from one language to another is that of having a non-native accent in an L2. Many of the English-trained listeners cited the relative accentedness of each talker's speech as a feature they listened to in trying to identify talkers during training. Knowing how much of an "accent" a non-native talker has while they are speaking English is useless information to have when trying to identify the same talker when they are speaking German. The fact that many of the English-trained listeners claimed to have relied on perceived "accentedness" when identifying talkers in training may therefore account for the difficulty these listeners displayed in transferring their knowledge of the talkers' voices to novel German stimuli. (For a discussion of linguistic information on the perception of accentedness, see Levi, Winters, & Pisoni, 2005).

In the most general terms, the existence of both language-specific and language-independent indexical properties confirms the predictions made by the integrated view of speech perception. However, the results of this study suggest that listeners identify talkers on the basis of more than just language-independent or language-specific indexical properties. Another free parameter in the perceptual system appears to be the listener's strategy for doing the voice identification task. Listeners who understand the language that a talker is speaking will automatically make use of language-specific indexical properties to identify that talker's voice. They may even base their talker identification judgments on irrelevant linguistic information, if the automatic process of word recognition interferes with the controlled process of talker identification. If listeners cannot understand the language that a talker is speaking, however, they will be forced to identify that talker's voice on the basis of language-independent indexical information encoded in the speech waveform. Listeners can thus apparently switch between a modular form of speech perception and an integrated form of speech perception, depending on what information is available to them in the signal. The general perceptual strategy appears to be: make use of the most specific information which is available—including language-specific information—regardless of what consequences this might have for the construction of broadly generalizable perceptual categories for individual talkers.

The ability of listeners to make use of whatever information is available to them in the speech signal in order to perform linguistic and voice identification tasks demonstrates that the perception of speech is a highly robust and adaptive process. The fact that the perceptual system can rapidly adapt to changing listening conditions can also reconcile the apparently conflicting evidence for both the modular and integrated views of speech perception that was presented in the introduction. Speech perception operates in an integrated manner to the extent that listeners can and do use multiple sources of linguistic and indexical information in the speech signal to help them perform both linguistic and voice identification tasks more proficiently. When either linguistic or indexical information is removed from the speech signal, however, the perceptual system is capable of interpreting the linguistic or indexical information that is still available, in an independent and apparently modular fashion. The perception of speech may thus be either integrated or modular, depending on the context in which it operates. The evidence in favor of one view of speech perception does not necessarily invalidate evidence for the other, therefore, as long as the kind of information which is available to listeners in the speech signal is taken into account.

References

Abercrombie, D. (1967). Elements of General Phonetics. Edinburgh: Edinburgh University.

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania [Distributor], Philadelphia, PA.

Barkley, R.A. (1997). Behavioral inhibition, sustained attention, and executive functions constructing a unifying theory of ADHD. *Psychological Bulletin, 121,* 65-94.

Bricker, P.D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40,* 1441-1449.

Clarke, F. R., Becker, R.W., & Nixon, J.C. (1966). Characteristics that determine speaker recognition. *Report ESD-TR-66-638*. Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command.

Compton, A.J. (1963). Effects of filtering and vocal duration upon the identification of speakers, aurally. *Journal of the Acoustical Society of America, 53,* 1741-1743.

Fenn, K.M., Nusbaum, H.C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425,* 614-616.

Glisky, E.L., Polster, M.R., & Routhieaux, B.C. (April, 1995). Double dissociation between item and source memory. *Neuropsychology, 9,* 229-235.

Goggin, J.P., Thompson, C.P., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19 5,* 448-458.

Goldinger, S. D. (1996) Words and Voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166-1183.

Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the locus of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17,* 152-162.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28,* 267-283.

Halle, M. (1985). Speculations about the representations of words in memory. In V. Fromkin (Ed.), *Phonetic Linguistics*. (pp. 101-114). Academic Press: Orlando.

Hirahara, T., & Kato, H. (1992). The effect of F0 on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure.* (pp. 89-112). Tokyo: Ohmsha Publishing.

Köster, O., & Schiller, N.O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics, 4,* 18-28.

Kreiman, J., & Van Lancker, D. (1988). Hemispheric specialization for voice recognition: Evidence from dichotic listening. *Brain and Language, 34,* 246-252.

Landis, T., Buttet, J., Assal, G., and Graves, R. (1982). Dissociation of ear preference in monaural word and voice recognition. *Neuropsychology, 20,* 501-504.

Levi, S.V., Winters, S.J., & Pisoni, D.B. (2005). Speaker-independent factors affecting the perception of foreign accent in a second language. In *Research on Speech Perception Progress Report No. 27* (pp. 49-64). Bloomington, IN: Speech Research Laboratory, Indiana University.

Levi, S.V., Winters, S.J., & Pisoni, D.B. (2006). Perception of the indexical properties of speech: universal or language-dependent? Poster presented at the *Tenth Conference on Laboratory Phonology*, Paris, France, June 30, 2006.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. (pp. 403-439). Dordrecht: Kluwer

Liu, A.-Y. (1973). Decrease in Stroop effect by reducing semantic interference. *Perceptual and Motor Skills, 37,* 263-265.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors, 25,* 17-32.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379-390.

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

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

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19,* 309-328.

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

Pollack, I., Pickett, J.M., & Sumby, W.H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America, 26,* 403-406.

Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.

Schachar, R., & Logan, G.D. (1990). Impulsivity and inhibitory control in normal development and childhood psychopathology. *Developmental Psychology, 26,* 710-720.

Schacter, D.L., & Church, B.A. (1992). Auditory priming: implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18,* 915-930.

Schiller, N.O., & Köster, O. (1996) Evaluation of a foreign speaker in forensic phonetic: a report. *Forensic Linguistics, 3,* 176-185.

Schiller, N.O., Köster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics, 4,* 1-17.

Stevens, A.A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research, 18,* 162-171.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643-662.

Sullivan, K.P.H., & Schlichting, F. (2000). Speaker discrimination in a foreign language: first language environment, second language learners. *Forensic Linguistics, 7,* 95-111.

Thompson, C.P. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1,* 121-131.

Van Lancker, D.R., Cummings, J.L., Kreiman, J., & Dobkin, B.H. (1988). Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195-209.

Williams, C.E. (1964). The effects of selected factors on the aural identification of speakers. Section III, *Report ESD-TDR-65-153*. Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command.

Winters, S.J., Levi, S.V., & Pisoni, D.B. (2005). When and why feedback matters in the perceptual learning of the visual properties of speech. In *Research on Speech Perception Progress Report No. 27* (pp. 107-132). Bloomington, IN: Speech Research Laboratory, Indiana University.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Lip-reading Skills in Bilinguals: Some Effects of L1 on Visual-only Language Identification[1]

Rebecca E. Ronquest and Luis Hernandez

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

[1] This work was supported by NIH T32 Training Grant DC00012 to Indiana University. We would also like to thank David Pisoni and Manuel Diaz-Campos for all of their helpful comments and suggestions.

Lip-reading Skills in Bilinguals: Some effects of L1 on Visual-only Language Identification

Abstract. This study investigated whether observers can identify what language was being spoken in visual-only speech stimuli, and whether or not this ability depends on an observers' prior linguistic experience. Participants watched visual-only speech stimuli and were asked to decide if the talker in the video was speaking English or Spanish. Four groups of participants were studied: monolinguals and bilinguals, who were either native speakers of English or native speakers of Spanish. Results revealed that all subjects were able to identify the language being spoken with about 80% overall accuracy, regardless of their language background. However, the groups of participants differed in terms of their response bias. The L1 English bilinguals were strongly biased toward their native language, whereas the other three groups of participants did not demonstrate a significant bias toward the L1. The results of this experiment and their implications are discussed, and several directions for future research are considered.

Introduction

A large body of research has demonstrated that speech perception is multimodal; that is, the auditory, visual, and tactile properties of speech carry important information that can affect the intelligibility of the speech signal. It is also well established that the visual properties of speech are robust. The pioneering study carried out by Sumby and Pollack (1954) showed that the visual properties of speech carry important information about the linguistic content of the signal. In their study, they found that as the signal-to-noise ratio decreased, the contribution of the visual aspect of speech (i.e. the face of the talker) increased. In other words, when auditory aspects of speech are insufficient to communicate the message, visual information, such as the movements of the talker's face, is often relied upon to fill in the gaps. Both normal-hearing and hearing-impaired listeners take advantage of visual information when perceiving speech.

The findings of Sumby and Pollack have been strengthened by other findings reported by McGurk and MacDonald (1976). They found that when presented with mismatched auditory and visual information, the information carried in those modalities will often become fused together, and the observer will perceive a completely different sound than the one that was presented in either modality. The best example of the McGurk effect occurs when an observer is presented with an auditory /ba/ and a visual /ga/. The perceiver often reports an intermediate version of the two syllables, namely /da/. Thus, the information carried in the visual (gestural) aspects of the signal is great enough to, in a sense, override certain aspects of the auditory signal.

More recently, studies in the field of second language acquisition have shown that the inclusion of visual information, along with auditory information, aids in the acquisition of non-native contrasts. For example, Hazan et al. (2001) reported that visual information facilitates perception of sounds that are constrastive in the L2, but do not contrast in the L1. Another study by Hardison (2003) concluded that facial gestures aid in the perception of L2 targets in difficult phonetic environments and that visual cues to speech can be a source of reliable information for L2 learners.

All of the studies mentioned above suggest that the visual aspects of speech carry information that can contribute substantially to the intelligibility of the signal. However, the amount of information

carried by the visual signal, and whether observers use this information, remain important issues. The goal of the present investigation was to examine these issues by asking several groups of participants to perform a visual-only language identification task. In particular, we asked whether the visual properties of speech are robust enough to allow an observer to extract language-specific information from a visual display.

The present study examined the performance of several groups of native Spanish and English-speaking monolinguals and bilinguals in a visual-only language identification task. The subjects were presented with a series of video clips without sound and were simply asked to decide if the person in the video was speaking English or Spanish. A review of the published literature failed to uncover any other investigations that examined bilingual lip-reading ability. For this reason, it is difficult to make any specific predictions as to how the subjects will perform. Thus, one of the main questions this research addresses is whether the participants can carry out the task successfully. A second question is whether the participant's native language and prior language experience influence their performance in identifying English and Spanish from visual information.

Methods

Participants

A total of 56 participants took part in the present investigation (average age 24.9 years). The participants were from four language groups: Monolingual English speakers (N=16), Monolingual Spanish speakers (N=12), L1 Spanish bilinguals (N=12), and L1 English bilinguals (N=16). The monolingual English speakers were all undergraduate students at Indiana University who reported minimal or no knowledge of Spanish. The monolingual Spanish speakers were all current residents of Caracas, Venezuela, who reported that they did not speak or have knowledge of English.[2] The L1 Spanish bilinguals and L1 English bilinguals were all graduate students in the Department of Spanish and Portuguese at Indiana University. The participants in these two groups reported that they were proficient speakers of both English and Spanish. Age of L2 acquisition ranged from birth to 19 years of age. All participants received $10 for taking part in the study. Each section of the experiment is described in more detail below.

Stimulus Materials and Procedure

The present experiment consisted of three parts: a language history questionnaire, identification of CUNY sentences, and a visual-only language identification task. The stimuli were presented on an Apple Macintosh G4 computer. PsyScript version 5.1 was used for stimulus presentation. Subjects' responses were recorded with the keyboard for the CUNY task, and a button box for the language identification task. The entire experiment took approximately one hour to complete.

Language History Questionnaire. All participants completed a language history questionnaire. The purpose of the questionnaire was to gather demographic information pertaining to the language history of each participant such as the age of L2 acquisition and L2 usage. The monolingual Spanish participants completed a version of the questionnaire that was translated into Spanish. The other three groups of participants completed all paperwork in English.

[2] The data for the monolingual Spanish speakers was collected by Manuel Diaz-Campos in Caracas, Venezuela during July of 2005.

CUNY Sentences. Each participant took part in a CUNY[3] sentences task. The CUNY sentences were presented to each participant in auditory-only, audio-visual, and visual-only modalities. Twelve sentences were presented in each of the three modalities. The participants were asked to type what they thought they heard or saw on each trial. For the visual-only condition, they were told to do their best and guess if they were not able to determine exactly what the person in the video was saying.

Visual-only Language Identification Task. The experimental design of the V-only language identification task consisted of two blocks of 40 V-only video clips of short phrases in Spanish and English. Each block consisted of 20 English phrases and 20 Spanish phrases spoken by either a male or a female talker. One block consisted of 40 phrases presented by the male talker, and the other block consisted of 40 phrases presented by the female talker. The order of the blocks was counterbalanced so that half of the participants were presented with the male speaker first, whereas others saw the female speaker first. After seeing each video clip, the participants were asked to decide if the person in the video was speaking English or Spanish. A button box was used to record the subjects' responses. No feedback was provided. Only the data from the visual-only language identification task will be discussed in this paper.

Results

An initial examination of the data revealed that all subjects were able to successfully complete the visual-only language identification task at accuracy levels that were statistically above chance. The overall mean percent correct score for all subject groups was 78.06 %. A repeated measures ANOVA with stimulus language (English vs. Spanish) and stimulus gender (male vs. female) as within subject variables and participant group (Monolingual English, Monolingual Spanish, L1 English Bilingual, and L1 Spanish Bilingual) as between subject variables revealed a significant main effect for stimulus language ($F(1,49) = 4.107$; $p = .048$) and stimulus gender ($F(1,49) = 4.539$; $p = .038$). The participants performed significantly better on the English stimuli (79.18% English, 76.56% Spanish), and the stimuli spoken by the female talker (79.59 % female, 76.67% male).

The results also revealed a significant interaction between participant group (monolingual, bilingual, L1 English, and L1 Spanish) and stimulus language ($F(3,49) = 5.65$; $p = .002$). No other interactions were significant. Post-hoc paired samples t-tests indicated that, while both groups of monolinguals and the L1 Spanish bilinguals performed no differently on the stimuli presented in English and Spanish, the L1 English bilinguals performed significantly better overall on the English stimuli ($p = .001$). Figure 1 shows the percent correct scores for each group of participants for each presentation condition.

In addition to percent correct scores, non parametric measures of sensitivity (A') and bias (B") were calculated for all subject groups[4]. Both of these measures use the hit and false alarm rates to determine how sensitive the subjects are to the language differences in the signal, and to assess the extent to which they are biased toward one response option over another. A one-way ANOVA of A' score and subject group revealed no significant differences in sensitivity between participant groups. In other words, this indicates that the native language and language experience of the participants did not affect

[3] The group of monolingual Spanish participants did not complete the CUNY sentences because these sentences are in English. At present, there is not a set of CUNY sentences, or equivalent sentences, in Spanish.

[4] Formula for sensitivity (A')= 1/2 + ((P(Hits) - P(FA)) * (1 + P(Hits) - P(FA))/(4 * P(Hits)* (1 - P(FA))) ;
Formula for bias (B") =(P(Hits)*(1-P(Hits)) - P(FA)*(1-P(FA))) / (P(Hits)*(1-P(Hits)) + P(FA)*(1-P(FA)))

their sensitivity to differences in the signal. Figure 2 shows the mean A' scores for each of the four subject groups.

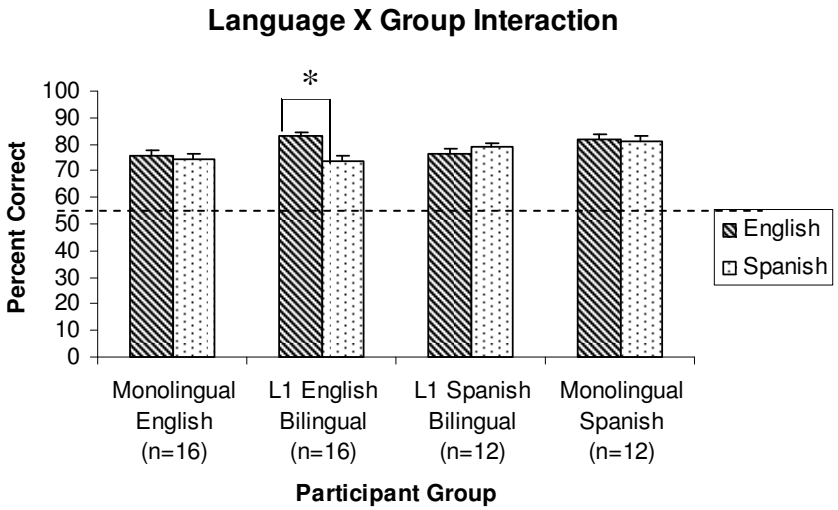


Figure 1. Percent correct scores for four subject groups. The dark bars indicate percent correct score on the English stimuli; the light colored bars represent percent correct scores on Spanish stimuli. Standard error bars are included. The dotted line represents scores significantly above chance using the binomial test.

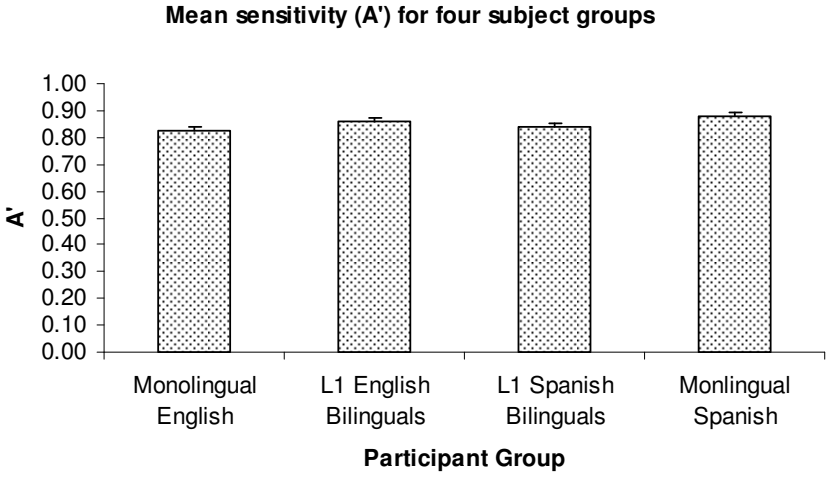


Figure 2. Mean sensitivity A' for all four subject groups.

A one-way ANOVA was also conducted in order to analyze the possible relationship between response bias (B") and participant groups. The results of this analysis revealed that the L1 English bilinguals had a response bias that was significantly different from the other three subject groups. While all participant groups showed at least some kind of response bias towards their native language, this bias was strongest for the group of L1 English bilinguals. This difference is shown in Figure 3.

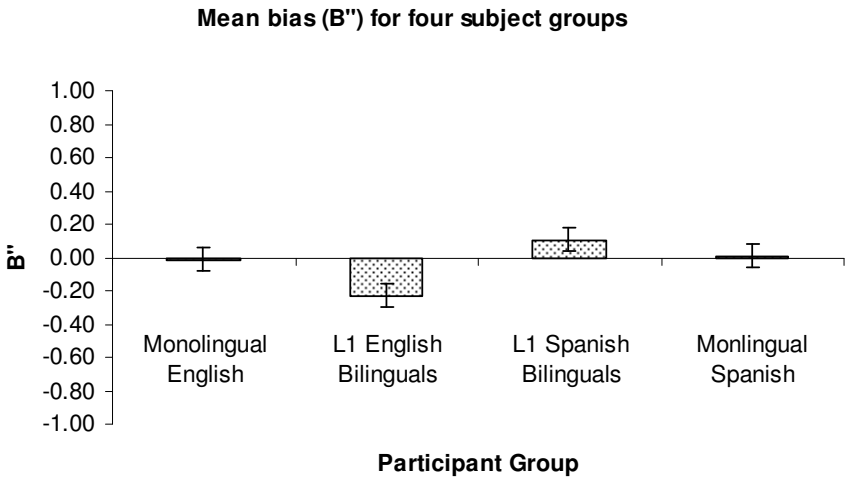


Figure 3. Mean bias (B") for all four subject groups. Negative values indicate a bias to respond more often as English; positive values indicate a bias to respond more as Spanish.

As shown in Figure 3, L1 English bilinguals displayed a strong response bias toward their L1, whereas the other three groups of participants did not appear to favor selecting their native language over their non-native language. The strong response bias toward the L1 explains why the L1 English bilinguals performed much better on the English stimuli.

Conclusions and Directions for Future Research

The results of this preliminary investigation provided some new insights into the robustness of the visual properties of speech. All participants were able to identify the language of a talker in a visual-only stimulus at levels well above chance. This result suggests that the visual signal provides enough information for an observer to correctly select the language being spoken from visual-only displays of speech. We also found that overall, the participants performed significantly better on the English stimuli than on the Spanish stimuli, and that performance was also better with the female talker.

One of the most interesting results was the interaction observed between language group and stimulus language. Our analysis revealed that native English-speaking bilinguals showed an increased level of performance when they were presented with English stimuli. However, the same effect was not found for the native Spanish-speaking bilinguals, or either group of monolinguals. Calculations of sensitivity (A') and bias (B") revealed that, although all four groups had comparable A' scores, the L1 English bilinguals displayed a strong response bias towards their native language. It is possible that the L1 English bilinguals were using a different strategy than the other three groups, which yielded a different result. It is interesting to note, however, that in the preliminary stages of this experiment, the native Spanish-speaking bilinguals showed a similar effect; the first eight subjects showed a higher

percentage correct score on the stimuli in Spanish. However, this effect was attenuated with the addition of more subjects.

One explanation for the lack of bias found in the L1 Spanish bilingual group could be that these participants were "set" in English mode, and thus failed to show the same kind of native language bias as the other group of bilinguals. All paperwork, instructions, and the CUNY sentences were presented to the L1 Spanish bilinguals in their non-native language, whereas the English monolinguals and L1 English bilinguals received instructions and task instructions in their native language. We are planning to present the L1 Spanish bilingual subjects with instructions and materials only in Spanish, as we did with the group of monolingual Spanish speakers from Caracas, Venezuela. Using the native language as the main mode of presentation may produce a native-language bias that is similar to that displayed by the L1 English bilinguals. In order to "set" the L1 Spanish bilinguals in Spanish mode, however, we will need to create a set of CUNY-like sentences in Spanish.

The present experiment measured participants' ability to identify the language of a talker using a visual-only phrase in English or Spanish. In future investigations we plan to examine participants' ability to identify English and Spanish using single words. The stimuli used in the present study varied in length, and on average, the Spanish stimuli were slightly longer and contained more syllables than the English stimuli. Thus, the participants may have used temporal cues to correctly identify the language being spoken. Isolated words, however, are much shorter in length, and may provide the participants with less useful duration information.

In another study we plan to reverse the video clips and present the information backwards in time. As previously mentioned, it is possible that the participants were able to make accurate language identifications based on utterance length or number of syllables. If duration or number of syllables were the major cues to language identity used by the subjects in this task, then temporally reversing the stimulus materials should not have any effect on overall performance. If, however, the participants were making their selections based on articulatory and gestural cues, temporal reversal should produce a decrease in performance on this task.

In conclusion, the present experiment has shown that participants are able to correctly identify the language of a talker when presented with a visual-only stimulus, suggesting that the visual properties of the speech signal are robust even in a language identification task. Future investigations will focus on identifying the particular visual cues that allow subjects to make such reliable judgments.

References

Hardison, D. (2005). Second-language spoken word indentification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics, 26,* 579-596.

Hardison, D. (2005). Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style. *Speech Communication, 46,* 73-93.

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

Hazan, V., Sennema, A., & Faulkner, A. (2001). Audiovisual perception in L2 learners. *Proceedings from the AVSP 2001 Conference on Auditory-Visual Speech Processing,*149-154.

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

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

Cross-modal Priming of Auditory and Visual Lexical Information: A Pilot Study[1]

Adam B. Buchwald and Stephen J. Winters

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

[1] This work was supported by NIH-NIDCD DC00012. The authors would like to thank David Pisoni for helpful advice and comments and Melissa Troyer for editing videos and running subjects.

Cross-modal Priming of Auditory and Visual Lexical Information: A Pilot Study

Abstract. This study assessed whether presenting visual-only stimuli prior to auditory stimuli facilitates the recognition of spoken words in noise. The results of the study indicate that this type of cross-modal priming does occur. Future directions for research in this domain are presented.

Introduction

Psycholinguistic studies often employ priming paradigms to address issues of whether and when certain representations are active in the course of language processing. In priming studies, researchers typically examine changes compared to a baseline level of performance in responding to a 'target' stimulus when the target is preceded by a 'prime' stimulus. The changes in participants' performance on the task that result from presentation of the prime are argued to indicate something about the relationship between the target and prime stimuli in the cognitive processing required for the task. For this reason, primes are generally selected that share some – or all – features with the target stimuli. For instance, in phonological priming, a spoken target and a spoken prime typically share some subset of phonological features. In repetition (or identity), priming, the prime and target are identical and thus share all features with one another.

In this pilot study, we used cross-modal priming and presented spoken word primes in the visual domain only, followed by an auditory-only presentation of the same word, spoken in noise, as the target stimulus. We were interested in using this priming paradigm to find out whether there were enough shared features between sensory modalities in the auditory and visual representations of the spoken word for the visual prime to facilitate the recognition of the auditory target.

There exist two lines of evidence in the psycholinguistic literature that suggest the information in a visual-only stimulus will facilitate lexical processing. Dodd, Oerlemans, and Robinson (1989) found that lexical repetition priming is robust across different modalities of presentation. They presented research participants with lexical primes and targets in three different modalities: orthographic, auditory, and visual. On each critical trial, the lexical item presented as the target and the prime were identical. The research participants were required to make a speeded semantic categorization ('animal' or 'plant') on the target lexical item. Importantly, Dodd et al.'s results indicated that the presentation of visual-only stimuli facilitated processing in the semantic categorization task for all three target types. With respect to the present experiment, it is noteworthy that participants were faster on the semantic categorization task with auditory targets when there was a visual prime than when the priming stimulus was absent. This result suggests that the information present in the visual-only prime facilitated processing of the semantic lexical information present in the auditory target stimulus.

Another line of evidence suggests that observers can integrate information present in auditory and visual signals, even when those signals are presented in separate modalities. Lachs and Pisoni (2004) performed a series of 'cross-modal matching' tasks in which participants were asked to match visual-only stimuli with auditory-only stimuli. Using an XAB task, observers viewed a silent video of a speaker producing a word, followed by two auditory-only stimuli produced by two different speakers (of the same gender) saying the same word. The observers' task was to identify which of the two auditory stimuli

came from the same speech event as the visual stimulus. Participants were able to match the appropriate auditory stimulus to the video at a rate significantly greater than chance. Importantly, when a still image of a speaker was presented as the visual stimulus instead of a dynamic video clip, participants performed at chance levels in matching the stimuli of the two modalities. Lachs and Pisoni argued that the information present in dynamic video clips and their corresponding auditory tracks were perceived as part of an integrated stimulus; that is, they were simply two sources of information about the same perceptual event in the external world.

Current Investigation

The present pilot study was performed to determine whether the presentation of a silent, dynamic video clip of a speaker would facilitate the recognition of spoken words, presented in only the auditory domain. This study tested whether this type of visual-audio cross-modal priming affects spoken word recognition, in addition to semantic categorization (as shown by Dodd et al., 1989). If so, the cross-modal priming paradigm could be used to explore a wide range of additional issues related to the operations and representations that underlie the processes required for spoken word recognition.

Method

Participants

Nine Indiana University undergraduate students, ages 18-20, participated in this study in partial fulfillment of course requirements for introductory psychology. All participants were native speakers of English with no speech or hearing disorders; they all had normal or corrected-to-normal vision at the time of testing.

Materials

All stimuli materials were drawn from the Hoosier multi-talker AV database (Sheffert, Lachs, & Hernandez, 1997). Only monosyllabic, CVC words produced by one female speaker in the database were selected for use in this study. Of the 96 different word tokens included in this study, half were “Easy” words – high frequency words with sparse phonological neighborhood densities (e.g., “fool,” “noise”), while the other half were “Hard” words – low frequency lexical items with high neighborhood densities (e.g., “hag,” “mum”; Luce and Pisoni, 1998).

Auditory Stimuli. In this experiment, we used envelope-shaped or ‘random bit flip’ noise (Horii, House, & Hughes, 1971) to reduce performance on the spoken word recognition task. Presenting the auditory stimulus in noise is necessary to detect the potential facilitatory effects of priming in the spoken word recognition task; performance must be below ceiling for any effects to be detected. To create these stimuli, the acoustic track of each video recording was first saved to an independent .AIFF file, using QuickTime Pro software. These files were then processed through a MATLAB program which randomly changed the sign bit of the amplitude level of 30% of the spectral chunks in the acoustic waveform.

Visual Stimuli. Two kinds of visual primes were used in this study: static and dynamic. Dynamic primes consisted of the original, unedited video clips associated with each target word from the Hoosier Audio-Visual Multi-Talker database. The video track of the static primes, on the other hand, consisted of only a still shot of the first frame of the video associated with the target word in the Audio-Visual database. The duration of the static primes was identical to that of their counterparts in the dynamic prime condition.

Procedure

Participants were tested in groups of four or fewer in a quiet room. During testing, each participant wore Beyer Dynamic DT-100 headphones while sitting in front of a Power Mac G4. A customized SuperCard (version 4.1.1) stack, running on the PowerMac G4, presented the stimuli to each participant. The instructions for the experiment were presented to the participants on the computer screen prior to the first experimental trial and are repeated below:

> In this experiment, you will attempt to identify a series of words that you hear. The words will be difficult to understand. After you hear each word, you should attempt to identify the word that was spoken. You can respond by typing into the computer.
>
> Before each word, you will see either a still image of a speaker or a movie of a person saying a word. Regardless of what you see, your task is to identify the word that you hear. Even if you do not think you understood the word, please make your best attempt to identify what you heard.

On each trial during the experiment, the SuperCard stack first presented participants with either a Static or a Dynamic visual prime. All videos had a 640x480 aspect ratio and filled the entire monitor screen when they were presented to the participants. The sound output from the computer was muted while the videos were presented to the participants. Five hundred milliseconds after the presentation of the visual prime, the participants then heard the auditory target word over the headphones. Following the auditory stimulus, a prompt appeared on the screen asking the participant to type in the word they heard. The prompt remained until the participant pressed the "Enter" key on the keyboard at which point a "Next Trial" prompt appeared. The next trial began after the participant used the mouse to click the "Next Trial" prompt.

Words were presented to participants in random order with Dynamic and Static primes randomly interleaved. Each participant responded to 48 words in each priming condition.

Results

The data were analyzed using a repeated measures Analysis of Variance (ANOVA) with prime condition (Dynamic vs. Static) and target type (Easy vs. Hard) as independent variables and word identification accuracy as the dependent variable. Correct responses were counted for all typographical matches between stimulus and response, as well as homophones (e.g., "peace" and "piece") and obvious typos (e.g., "cheif" for "chief"). The percentage of correct responses in each priming condition, for both target types, is represented graphically in Figure 1. The ANOVA revealed a main effect of prime condition [$F(1,8) = 33.71$, $p < 0.001$]. Participants were significantly more accurate on trials with Dynamic primes ($\bar{x} = 68.1\%$, $\sigma = 7.0\%$) than on trials with Static primes ($\bar{x} = 52.5\%$, $\sigma = 5.4\%$). A main effect of target type was also significant, with participants performing better on Easy targets ($\bar{x} = 68.1\%$, $\sigma = 12.9\%$) than Hard targets ($\bar{x} = 52.1\%$; $\sigma = 8.8\%$; $F(1,8) = 18.14$, $p < .01$). There was no interaction between prime type and target type, although there was a trend, $F(1,8) = 3.46$, $p < .11$), with a larger effect of prime condition for Easy words than for Hard words.

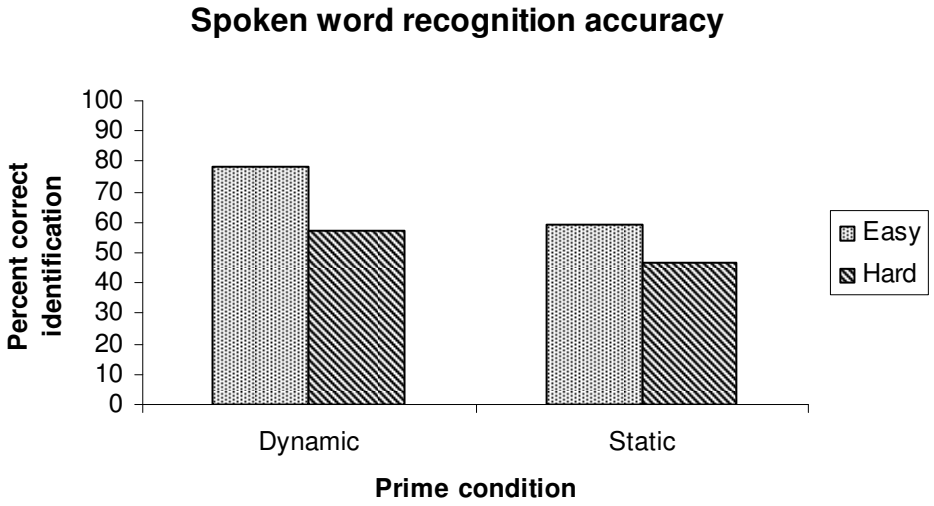


Figure 1. Percentage of words correctly identified as a function of prime condition and target type.

Discussion and Future Directions

The results of this initial study indicate that presenting a visual-only version of a word prior to presenting that word auditorily in noise facilitates the correct identification of the auditory target. This preliminary result has important implications for our understanding of spoken word recognition. In particular, the accuracy benefit for visually primed words suggests that the information that observers receive from the visual presentation of words aids the successful lexical identification of the auditory signals.

Future research building on this pilot study will involve attempting to determine what type of cognitive processing mechanism underlies this pattern of results. Two such proposals are considered here. First, Lachs and Pisoni (2004) have argued that success on the cross-modal matching task was due to observers' integration of the auditory and visual components of the speech act, as each provides information about the same event in the physical world (following the framework of Gibson, 1966). With respect to the study reported here, on trials when observers view the dynamic video clip, they are being presented with information about the speech act in the visual modality and this information allows the observers an additional channel of information with which to directly perceive the speech act. Speech acts are multi-modal by definition, and observers know that the visual-only prime has lawful consequences for the possible speech acts, thus providing information that may be absent in the noise-filtered auditory target.

A second possible explanation holds that the critical property of the relationship between the visual prime and the auditory target is that they share linguistic (or lexical) information, and not that they are from the same physical event in the world. Under this view, the visual prime activates sublexical representations which may also be activated by the auditory target. When these two stimuli contain the same information, they activate representations that enable accurate identification of the auditory signal. Here we are agnostic as to whether these representations are either linked representations of modality-specific information or whether an amodal representational level is activated. Crucially, this second hypothetical mechanism holds that the priming benefit should be maintained even when the visual prime

and the auditory signal come from different speakers, whereas the claim that the priming effect comes from the integration of auditory and visual information does not predict a priming benefit when the auditory and visual information has different sources.

In future work, we plan on replicating this result with a larger population of participants. Additionally, we will use the cross-modal priming paradigm to address these two hypotheses discussed above. In particular, we will investigate the extent to which the word identification accuracy benefit from the visual-only prime comes from the match in lexical information in the visual prime and auditory target as opposed to the match in the entire audiovisual event despite the separation of these two components along a temporal dimension. This issue will be explored by presenting observers with different speakers in the two modalities: speaker A will be seen producing a word and speaker B will be heard in the auditory stimulus component of the trial. If the priming effects observed here are due entirely to the integration of audio-visual information, no priming benefit should be seen in this condition. On the other hand, if the priming effects observed in this pilot study arise solely from the match in lexical information in the two stimulus events, the priming effect should be replicated when the auditory and visual stimuli are produced by different speakers. A third possibility is that a priming effect will be observed, but that the magnitude of the effect will be attenuated when the visual and auditory stimuli are produced by different speakers. This result would suggest that the priming effect observed here relies on both lexical identity and the integration of audio-visual information, such that removing the latter factor diminishes the effect but does not make it disappear altogether.

A second research question will explore the nature of the visual stimuli that can engender this priming effect. The pilot study reported here employs full-face visual stimuli of a speaker producing the given lexical item. In future work, we plan to replace these full-face stimuli with point-light displays that present a relatively impoverished depiction of the speaking event, to determine whether the priming benefit observed here is also obtained when the prime stimulus is a degraded dynamic visual stimulus.

Summary

This study was carried out to determine whether presenting visual-only stimuli prior to auditory stimuli facilitates the recognition of spoken words in noise. The results of the study indicate that this type of cross-modal priming does occur, and may be a useful tool for investigating issues related to spoken word recognition in future work.

References

Dodd, B., Oerlemens, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Language, 22*, 59-77.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.

Horii, Y., House, A.S., & Hughes, G.W. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America, 49*, 1849-1856.

Lachs, L., & Pisoni, D.B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16*, 159-187.

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

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.

III. Publications

ARTICLES PUBLISHED:

Bergeson, T., Pisoni, D.B. & Davis, R.B.O. (2005). Development of audiovisual comprehension skills in prelingually deaf children with cochlear implants. *Ear & Hearing, 26*, 149-164.

Burkholder, R.A., Pisoni, D.B. & Svirsky, M.A. (2005). Effects of a cochlear implant simulation on immediate memory in normal-hearing adults. *International Journal of Audiology, 44,* 551-558.

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

Clopper, C.G., Conrey, B.L. & Pisoni, D.B. (2005). Effects of talker gender on dialect categorization. *Journal of Language and Social Psychology, 24*, 182-206.

Clopper, C.G., Levi, S.V. & Pisoni, D.B. (2006). Perceptual similarity of regional dialects of American English. *Journal of the Acoustical Society of America, 119,* 566-574.

Clopper, C.G. & Pisoni, D.B. (2006). Effects of region of origin and geographic mobility on perceptual dialect categorization. *Language, Variation and Change, 18,* 193-221.

Clopper, C.G. & Pisoni, D.B. (2006). The Nationwide Speech Project: A new corpus of American English Dialects. *Speech Communication, 48,* 633-644.

Clopper, C.G. & Pisoni, D.B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47,* 207-239.

Clopper, C.G., Pisoni, D.B. & deJong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America, 118,* 1661-1676.

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

Conrey, B.L. & Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America, 119,* 4065-4073.

Dillon, C.M., Burkholder, R.A., Cleary, M. & Pisoni, D.B. (2004). Nonword repetition by children with cochlear implants: Accuracy ratings from normal-hearing listeners. *Journal of Speech, Language and Hearing Research, 47,* 1103-1116.

Hay-McCutcheon, M.J., Pisoni, D.B. & Kirk, K.I. (2005). Audiovisual speech perception in elderly cochlear implant recipients. *Laryngoscope, 115*, 1887-1894.

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

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

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

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

Houston, D.M., Carter, A.K., Pisoni, D.B., Kirk, K.I., Ying, E.A. (2005). Name-learning skills of deaf children following cochlear implantation: A first report. *Volta Review, 105*, 39-70.

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

Lachs, L. & Pisoni, D.B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16,* 159-187.

Levi, S.V. (2005). Acoustic correlates of lexical accent in Turkish. *Journal of the International Phonetic Association, 35,* 73-97.

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

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

BOOK CHAPTERS PUBLISHED:

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

Clopper, C.G. & Pisoni, D.B. (2005). Perception of dialect variation. In D.B. Pisoni & R.E. Remez (Eds.), *Handbook of Speech Perception*. Pp. 313-337. Blackwell Publishers.

Clopper, C.G. & Pisoni, D.B. (2005). Speech perception, hearing impairment, and linguistic variation. In M.J. Ball (Ed.), *Clinical Sociolinguistics*. Pp. 207-2187. Blackwell Publishers.

Pisoni, D.B. (2005). Speech perception in deaf children with cochlear implants. In D.B. Pisoni & R.E. Remez (Eds.), *Handbook of Speech Perception*. Pp. 494-523. Blackwell Publishers.

Winters, S.J. & Pisoni, D.B. (2005). Speech synthesis: Perception and comprehension. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics*, Vol. 12, 31-49.

PROCEEDINGS PUBLISHED:

Burkholder, R. & Pisoni, D.B. (2004). Digit span recall error analysis in pediatric cochlear implant users. *International Congress Series, 1273,* 312-315.

Burkholder, R., Pisoni, D.B. & Svirsky, M.A. (2004). Perceptual learning and nonword repetition using a cochlear implant simulation. *International Congress Series, 1273,* 208-211.

Clopper, C.G. & Pisoni, D.B. (2004). Perceptual dialect categorization by an adult cochlear implant user: a case study. *International Congress Series, 1273,* 235-238.

Dillon, C.M. & Pisoni, D.B. (2004). Nonword repetition and reading in deaf children with cochlear implants. *International Congress Series, 1273,* 304-307.

Horn, D.L., Davis, R.A.O, Pisoni, D.B. & Miyamoto, R.T. (2004). Visual attention, behavioral inhibition and speech/language outcomes in deaf children with cochlear implants. *International Congress Series, 1273*, 332-335.

Horn, D.L., Davis, R.A.O, Pisoni, D.B. & Miyamoto, R.T. (2004). Visuomotor integration ability of pre-lingually deaf children predicts audiological outcome with a cochlear implant: a first report. *International Congress Series, 1273*, 356-359.

Pisoni, D.B. (2004). Information processing skills of deaf children with cochlear implants: some new process measures of performance. *International Congress Series, 1273*, 283-287.

Winters, S.J. & Pisoni, D.B. (2005) When and why feedback matters in the perceptual learning of the visual properties of speech. *Proceedings of the International Speech Communication Association Workshop on Plasticity in Speech Perception*, 148-151.

MANUSCRIPTS ACCEPTED FOR PUBLICATION (IN PRESS):

Bent, T. & Pisoni, D.B. (In press). Some comparisons in perception between speech and nonspeech signals. In M. Ball (Ed.), *Handbook of Clinical Linguistics*. Blackwell Publishers.

Clopper, C.G. & Paoillo, J.C. (In press). North American English vowels: A factor analytic perspective. *Literary and Linguistic Computing.*

Clopper, C.G. & Pisoni, D.B. (In press). Some new experiments on perceptual categorization of dialect variation in American English: Acoustic analysis and linguistic experience. In N. Neidzielski (Ed.), *Speech perception in context: Beyond acoustic pattern matching*. Erlbaum.

Clopper, C.G. & Pisoni, D.B. (In press). Free classification of regional dialects of American English. *Journal of Phonetics*.

Levi, S.V. (In press). Reconsidering the variable status of glottals in Nasal Harmony. *Chicago Linguistic Society 41*.

Levi, S.V. & Pisoni, D.B. (In press). Indexical and linguistic channels in speech perception: Some effects of voiceovers on advertising outcomes. In T. Lowery (Ed.), *Psycholinguistic Phenomena in Marketing Communications*. Mahwah, NJ: Lawrence Erlbaum.

Pisoni, D.B. & Levi, S.V. (In press). Representations and representational specificity in speech perception and spoken word recognition. In M.G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. Oxford University Press: UK.

MANUSCRIPTS SUBMITTED:

Buchwald, A.B., Winters, S.J. & Pisoni, D.B. (Submitted). Multimodal speech perception: Evidence form cross-modal priming. *Psychological Science*.

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

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

Dillon, C.M. & Pisoni, D.B. (Submitted). Nonword repetition and reading skills in children with cochlear implants. *Volta Review*.

Lachs, L. & Pisoni, D.B. (Under revision). Visual recognition of spoken words without audition. *Perception & Psychophysics*.

Levi, S.V., Winters, S.J. & Pisoni, D.B. (Under revision). Speaker-independent factors affecting the perception of foreign accent in a second language. *Journal of the Acoustical Society of America*.

McMichael, K.H. & Pisoni, D.B. (Under revision). Effects of talker-specific encoding on recognition memory for spoken sentences. *Memory & Cognition.*

Tierney, A.T. & Pisoni, D.B. (Under revision). Some effects of early musical experience on auditory sequence memory. *Music Perception*.

Winters, S.J., Levi, S.V. & Pisoni, D.B. (Submitted). When and why feedback matters in the perceptual learning of visual displays of speech. *Journal of the Acoustical Society of America*.