RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 20
(1995)

David B. Pisoni, Ph.D.
Principal Investigator

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

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. 20 (1995)

Table of Contents

INTRODUCTION

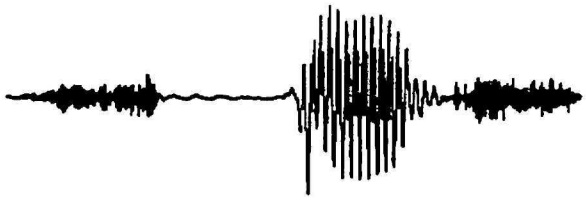
This is the twentieth annual progress report summarizing research activities on speech perception and spoken language processing carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize our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

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

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405-1301
USA
Phone: (812) 855-1155, 855-1768
FAX: (812)855-4691
E-mail: pisoni@indiana.edu

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

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



Speech Research Laboratory Faculty, Staff and Technical Personnel

(1/1/95 - 12/31/95)

Research Personnel:

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

Karen Iler Kirk, Ph.D. ..... Assistant Professor of Otolaryngology-Head and Neck Surgery[2]
Mario Svirsky, Ph.D. ..... Associate Professor of Otolaryngology-Head and Neck Surgery[3]

Ann R. Bradlow, Ph.D. ..... NIH Post-doctoral Trainee
Steven B. Chin, Ph.D. ..... NIH Post-doctoral Trainee
Helena M. Saldaña, Ph.D. ..... NIH Post-doctoral Trainee[4]
Mitchell S. Sommers, Ph.D. ..... NIH Post-doctoral Trainee[5]
Sonya M. Sheffert, Ph.D. ..... NIH Post-doctoral Trainee
Lynne C. Nygaard, Ph.D. ..... NIH Post-doctoral Trainee[6]

Michele Morrisette, B.S. ..... NIH Pre-doctoral Trainee
Amy Neel, M.S. ..... NIH Pre-doctoral Trainee
Carolyn Pytte, B.S. ..... NIH Pre-doctoral Trainee
Gina Torretta, B.A. ..... NIH Pre-doctoral Trainee

John R. Karl, B.A. ..... Graduate Research Assistant
William R. Svec, B.S. ..... Graduate Research Assistant

David Crotzer, B.S. ..... NIH Medical Student Trainee
Donald L. Schilson, B.S. ..... NIH Medical Student Trainee

---

[1] Also Adjunct Professor of Linguistics, Indiana University, Bloomington, IN: also Adjunct Professor of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.
[2] Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.
[3] Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.
[4] Now at House Ear Institute, Los Angeles, CA.
[5] Now at Department of Psychology, Washington University, St. Louis, MO.
[6] Now at Department of Psychology, Emory University, Atlanta, GA.

Technical Support Personnel:

Luis R. Hernandez, B.A. ..... Computer Systems Administrator/Analyst
Darla J. Sallee ..... Administrative Assistant
Jerry C. Forshee, M.A. ..... Computer Systems Analyst[7]
David A. Link ..... Electronics Engineer
Caleb Marcinkovich ..... Programmer
Jon M. D'Haenens ..... Programmer

Melissa Kluck ..... Undergraduate Research Assistant
Jeremy Nation ..... Undergraduate Research Assistant
Nathan Large ..... Undergraduate Research Assistant
Quinn Weaver ..... Undergraduate Research Assistant

E-Mail Addresses

pisoni@indiana.edu
kkirk@indyvax.iupui.edu
svirsky@indyvax.iupui.edu
abradlow@indiana.edu
schin@indiana.edu
saldana@hei.org
msommers@artsci.wustl.edu
ssheffert@psysrl.psych.indiana.edu
nygaard@fs1.psy.emory.edu
mmorrise@indiana.edu
abeardsl@indiana.edu (Amy Neel)
cpytte@indiana.edu
gtorrett@indiana.edu
jkarl@indiana.edu
hernande@indiana.edu
dsallee@indiana.edu
forshee@indiana.edu
cmarcink@indiana.edu
jdhaenen@indiana.edu
mkluck@indiana.edu
jnation@indiana.edu
nlarge@indiana.edu
qweaver@indiana.edu

[7]Also Director of Technical Support, Department of Psychology, Indiana University, Bloomington, IN.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Some Thoughts on "Normalization" in Speech Perception[1]

David B. Pisoni[2]

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

[1]This research was supported by NIH Research Grant DC-00111 to Indiana University in Bloomington, IN. I thank Steve Chin for his editorial help and Darla Sallee for her assistance in preparing the manuscript. This chapter is to appear in K. Johnson and J.W. Mullennix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press, 1996.

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

Abstract

One of the central problems in speech perception concerns stimulus variability, specifically, the mapping relations between acoustic attributes of the signal and linguistic categories resulting from perceptual analysis. Previous accounts of speech perception have treated variability as an undesirable source of noise in the signal that must be reduced or eliminated through a process of perceptual normalization. In this paper, I address what people mean by the term normalization in speech perception and examine the definitions and some of the implicit assumptions that have been made about it in the past. Then I summarize the results of several recent experiments on talker variability and perceptual learning which show that listeners encode fine stimulus details and use indexical attributes of speech in word recognition and sentence perception. Finally, I describe an alternative view of speech perception that is based on ideas from nonanalytic cognition. According to this approach, stimulus variability is "lawful" and "informative" for perceptual analysis. Perceptual normalization, as discussed in past theoretical accounts, may not involve a true "loss" of information but rather may entail the encoding of specific instances and the details of perceptual analysis. The generation of "equivalent forms" may occur at the time of retrieval from memory as a result of computational processes rather than early in perceptual analysis and encoding as previously assumed. This approach to speech perception and word recognition provides a new way of dealing with a number of long-standing problems in the field.

Some Thoughts on "Normalization" in Speech Perception

Introduction

The theoretical problems confronting researchers working in the field of speech perception are, in principle, no different from the problems studied in other areas of perception and cognition. Basically, they involve issues of invariance and variability of the speech signal, the neural representation of speech in the auditory system and the perceptual constancy maintained by human listeners in the face of diverse physical stimulation. These are well-known problems in speech perception that have occupied psychologists, linguists and engineers for close to a half-century since the beginning of modern speech research in the late 1940's (see Klatt, 1989, for a review).

When compared to other perceptual systems, there can be little doubt that speech perception is extremely robust and adaptive over a wide range of environmental conditions which introduce large physical changes and transformations in the acoustic signal. For example, normal hearing listeners can adapt easily to changes in speakers, dialects, speaking rate and speaking style as well as a wide variety of acoustic transformations including the presence of noise, reverberation and the use of different transducers without any noticeable decrease in performance. Even the most sophisticated state-of-the-art speech recognition systems cannot compare to the speed and efficiency of the human listener. It has always seemed that one of the principle tasks of speech science was to explain these remarkable perceptual and cognitive abilities and to model how the nervous system accomplishes this task so quickly and effortlessly in the face of continuous changes in the listener's environment.

For the last several years, I have been interested in the problems of stimulus variability in speech, in particular, the effects of stimulus variability from different talkers and different speaking rates on word recognition performance. Recent findings have suggested to me that some of the long-standing theoretical assumptions that speech researchers have held about the existence of abstract units such as phonemes and words need to be reexamined and substantially revised. More specifically, the assumption of an idealized symbolic representation for spoken language has encouraged researchers to search for simple first-order physical invariants and to ignore the problem of stimulus variability in the listener's environment. Variability is simply treated as a troublesome source of "noise" in the acoustic signal. Following recent accounts of memory and concept-learning by Jacoby and Brooks (1984), we will call this the traditional "abstractionist" or "analytic" approach to speech perception, an approach that places primary emphasis on the search for idealized categories that encode the linguistic content of speech into abstract symbolic units.

Findings from our laboratory on stimulus variability have suggested an alternative view of speech perception that is compatible with a large and growing body of literature in the field of cognitive psychology that deals with categorization. This view of speech perception focuses on the encoding of specific instances and assumes that very detailed stimulus information in the speech signal is processed by the listener and becomes part of the memory representation for spoken language. One of the assumptions of this "non-analytic" approach is that stimulus variability is, in fact, a lawful and highly informative source of information for the perceptual process; it is not simply a source of noise that masks or degrades the idealized symbolic representation of speech in human long-term memory. According to this view, listeners encode particulars rather than generalities. Our research has shown that source information about the talker's voice and detailed information about speaking rate is, in fact, encoded into memory and forms part of the neural representation of speech. Rather than discarding the "indexical" attributes of speech in favor of a highly abstract symbolic code like a string of segments or phonemes, the human perceptual and memory systems appear to encode and retain very fine details of the perceptual event. Our results have a

number of implications for future research on spoken language processing as we continue to gain new insights into the "nonanalytic" aspects of speech and to reconsider the long-standing problems of invariance and variability in speech perception.

In this chapter, I consider the problem of perceptual normalization, specifically, talker normalization in light of some recent findings and new theoretical developments in the field of perceptual learning and categorization. In past theoretical accounts of speech perception, a strict dissociation has been made between the linguistic properties of speech which carry the speaker's intended message and the indexical features of the signal which provide information about the talker's voice (Studdert-Kennedy, 1974). The dissociation between the form and content of the speech signal has a long history in the fields of linguistics and phonetics which has been carried over to theoretical accounts of speech perception despite the obvious fact that both sources of information are carried simultaneously and in parallel by the same acoustic signal. An excellent example of the functional parallelism of the indexical and linguistic properties of speech can be seen in Figure 1 which shows the encoding of speech in the auditory periphery (Hirahara & Kato, 1992). The absolute frequencies of the formants provide cues to speaker identification (left-hand panel) whereas the relative differences among the formants specify information for vowel identification (right-hand panel). Both sets of attributes are carried simultaneously by the same acoustic signal.

Insert Figure 1 about here.

Our recent findings suggest that the indexical attributes of a talker's voice are perceived and encoded in memory by the perceptual system along with the linguistic message, and that information about the taker's voice is not lost or discarded as a consequence of perceptual analysis. A variety of behavioral studies involving perceptual identification, speeded classification and recognition memory tasks have demonstrated that very detailed information about the talker's voice and the specific stimulus tokens is encoded and retained in memory and subsequently affects spoken word recognition performance. Other studies have shown that listeners acquire information about unfamiliar voices that improves the intelligibility of novel words and sentences from these same talkers.

These results imply that the indexical and linguistic attributes of speech are not neatly partitioned into two independent channels of information by the nervous system. Indeed, the mutual dependencies observed in perceptual analysis between these two sources of information suggest very close interactions between the form and content of the linguistic message and the listener's linguistic knowledge. Thus, what a listener learns about a talker's voice-- the acoustic correlates of gender, dialect, speaking rate, and so forth are encoded and subsequently used to facilitate a phonetic interpretation of the linguistic content of the message.

These recent findings raise fundamental questions about the nature of the traditional symbolic representations that have been routinely assumed in previous accounts of speech perception. Current views about the neural representation of speech and the basic perceptual units will need to be revised and substantially enriched to accommodate the additional fine details of stimulus encoding that listeners carry out in speech perception and spoken language processing tasks in the laboratory and real world. We believe these new perceptual findings are important because they suggest an alternative approach to traditional accounts of speech perception and spoken word recognition which have relied heavily on abstract, idealized symbolic representations of the linguistic message with little or no concern about the contribution of the talker's voice and the indexical features of speech to the perceptual process and the neural representation of

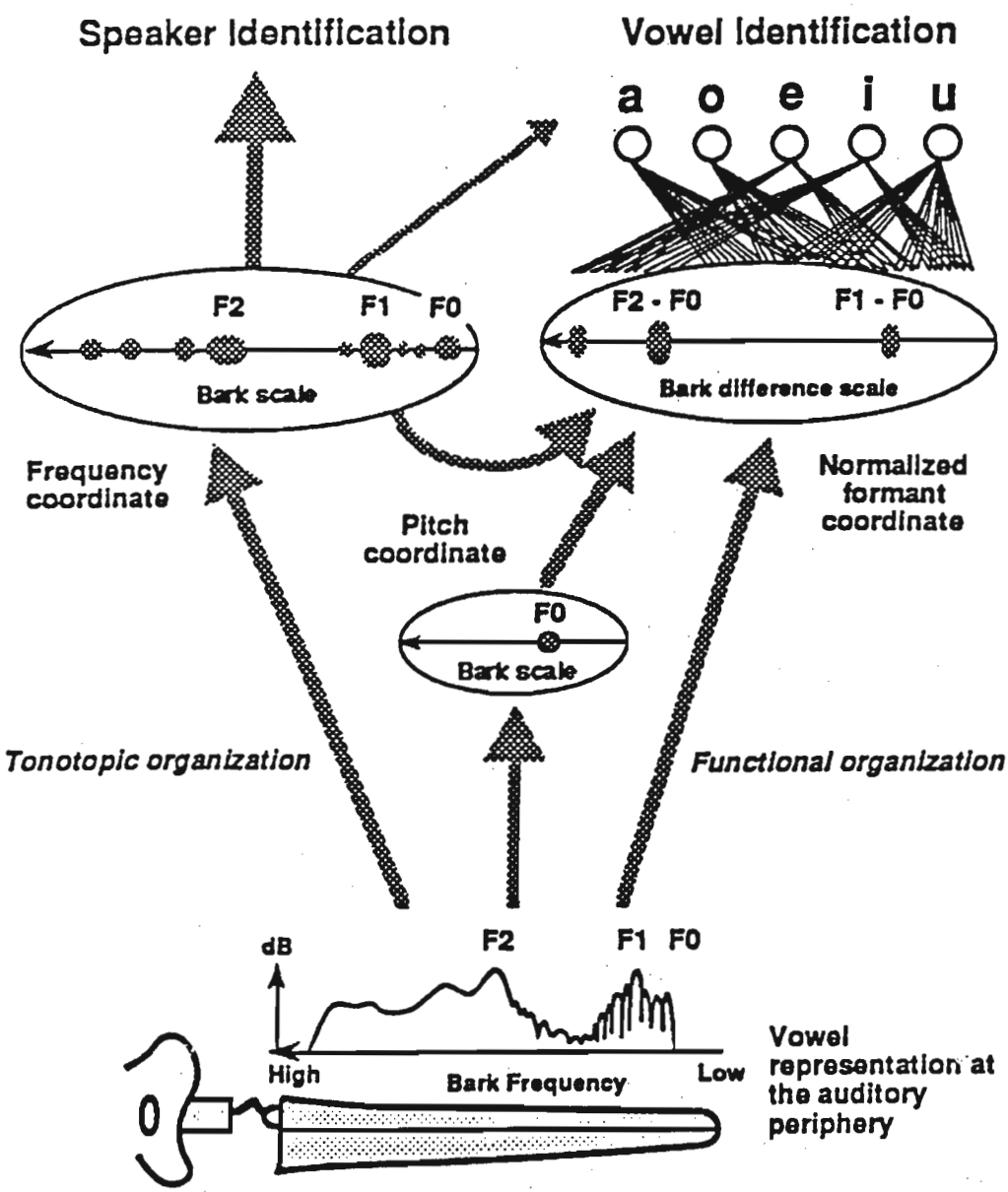


Figure 1. Vowel projections at the auditory periphery show that information for speaker identification and perception of vowel quality is carried simultaneously and in parallel by the same signal. The tonotopic organization of frequencies using a bark scale provides cues to speaker identification whereas the relations among the formant patterns in terms of difference from F0 in barks provide cues to vowel identification. (From Hirahara & Kato, 1992).

speech in memory. The findings from these recent studies also raise several interesting questions about the theoretical assumptions that have been made over the years concerning the process of normalization in speech perception.

This chapter is divided into three sections. First, I begin by reviewing several definitions of the term normalization and consider the implications that follow from adopting the traditional meanings of this term. Then I briefly summarize several recent findings on the contribution of the talker's voice to speech perception and spoken word recognition. Finally, I describe an alternative approach to perceptual normalization that can accommodate these new findings within a theoretical framework that emphasizes the role of non-analytic cognition and the contribution of specific instances to perception, learning and cognition, especially memory and perceptual learning in speech perception and spoken word recognition.

My goal in putting these ideas together at this time is to argue that it is possible to approach some of the old traditional problems in speech perception and spoken word recognition such as invariance, variability and the need for perceptual normalization in novel ways that can lead to new knowledge and insights into the fundamental process of speech perception. I believe this was not possible because of the traditional prevailing metatheory that was adopted by everyone working in the field and because of the experimental methodologies routinely employed in the past to study speech perception. However, this is a new era in the study of brain, behavior and computation. Recent developments in cognitive and neural science, ecological psychology and neural modeling, along with the widespread availability of high-speed computing technology, have provided new tools and powerful methods to study speech perception under much more demanding and robust conditions under which many different sources of variability can be manipulated and studied.

Two key features of the traditional approach to speech perception are quite striking when viewed from the nonanalytic perspective emphasizing the role of specific instances to perception and cognition that I am suggesting here. First is the dissociation, mentioned earlier, between the linguistic and indexical properties of the speech signal. Second is the assumption that stimulus variability is a source of noise and is not informative to the listener. Both of these assumptions have had profound effects on the study of speech perception and have strongly affected the kind of research problems that speech scientists have investigated over the years. These two assumptions have also affected the specific experimental methodologies used to study speech perception. Because stimulus variability was thought to mask or obliterate the underlying idealized symbolic message, factors known to create variability in the speech signal were deliberately reduced or eliminated. These factors were viewed as nuisance variables that needed to be controlled in experiments. For example, the traditional approach to acoustic-phonetic research typically used a small number of talkers, usually only one or two, reading carefully constructed experimental materials in citation format under extremely good recording conditions in the laboratory (Byrd, 1992). And, each experiment also typically addressed only a single specific research issue using a very small sample of highly controlled test materials such as isolated nonsense syllables, words or short sentences. In contrast, the current approach to acoustic-phonetic research relies very heavily on the use of large speech databases such as the TIMIT corpus which was collected with many different talkers producing speech under a wide variety of different conditions. Moreover, the test materials were specifically constructed to sample a larger and more diverse phonetic space, permitting computation of a variety of lexical statistics and the development of different types of models that can be used to characterize the structure and organization of lexical patterns in a given language (see Huttenlocher & Zue, 1984; Shipman & Zue, 1982; Zue, 1985; Pisoni, Nusbaum, Luce & Slowiaczek, 1985).

The consequence of this traditional research strategy in acoustic-phonetics was that little if any effort was made to study the contribution of different sources of variability directly or to try to understand how variability affected speech perception or spoken word recognition performance in human listeners. Based on our recent findings, I believe that we have made some progress in understanding the role of stimulus variability in speech and how it affects performance in a variety of tasks with different populations of listeners. I also feel that a substantial body of new knowledge has been obtained about the interdependencies between the indexical and linguistic attributes of the speech signal. Listeners are sensitive to fine phonetic details in the signal and they encode and represent acoustic changes in their listening environment that are potentially useful.

What Do We Mean By Normalization?

My observations and conclusions about stimulus variability and the encoding of fine phonetic details in speech are not entirely novel and should not be too surprising to people working in the mainstream of speech research. About 15 years ago, Dennis Klatt (1979) made a similar proposal about the need to preserve potentially useful acoustic-phonetic information in the context of his LAFS (Lexical Access From Spectra) model of speech recognition (Klatt, 1979). His proposal was based on the idea that an optimally efficient speech recognition system should be one which can recover gracefully from incorrect labeling or identification without catastrophic effects on performance. Klatt argued that models of speech perception that assume some form of intermediate level of discrete symbolic representation like phonemes or segments discard potentially useful acoustic-phonetic information that might be necessary if backtracking were necessary. Without retaining fine stimulus details, it would be impossible, according to Klatt, to recover from an errorful interpretation based on ambiguous or incorrect information in the signal.

Without remembering very much about Klatt's earlier proposal concerning the importance of retaining fine phonetic details, I initially found our results on talker variability troublesome for traditional accounts of speech perception which assumed that the speech signal is quickly encoded into a sequence of discrete abstract symbolic units like phonemes or phonetic segments. Of course, at that time, there weren't many other viable alternatives to the traditional view of speech or competing accounts of speech perception. Everyone just assumed that phonemes or segments or some kind of symbolic units were perceived during the process of speech perception and constituted the basic perceptual units employed in accessing words from the mental lexicon. These theoretical views about the nature of speech have been around for a long time and have been prominent in discussions of speech perception and speech recognition. One of the best examples of the traditional idealized view of speech is expressed in Charles Hockett's well-known description of speech as a sequence of Easter eggs:

> 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)

Although this description of speech may seem a little simplistic, it does illustrate nicely the prevailing theoretical views at the time and the very strong emphasis on abstract symbolic coding of speech into discrete idealized linguistic units. Hockett was in good company as shown by the approach of two prominent theorists in the field of speech communication, J. C. Licklider and Morris Halle who both endorse the same set of fundamental assumptions about the discrete symbolic nature of speech:

> ...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)

As I went back recently and thought about some of these old views in greater detail, I realized that it might be a useful intellectual exercise to review and reconsider what researchers mean by the term "normalization" in speech perception and what some of the implicit assumptions are that follow from adopting this particular approach to dealing with stimulus variability and perceptual constancy in speech. Before proceeding to this task, however, it should be pointed out here that although the most important and probably the most distinctive property of speech is its inherent physical variability (see Table I), these properties are not what are typically discussed in traditional accounts of speech perception or spoken word recognition. Instead of encouraging research on the study of variability, most theorists have tended to emphasize the discrete symbolic nature of speech and its linguistic function in conveying meaning. Moreover, there has always been a great deal of interest in figuring out ways to eliminate variability from experiments using speech stimuli rather than studying the problem of variability directly.

Insert Table I here

Viewed in this way, the study of variability in speech has historically taken a backseat to the formalist approach to speech derived from linguistic theory, especially the recent views of language promoted by transformational grammar with its emphasis on the linguistic competence of an idealized speaker-hearer as summarized in Chomsky's well-known passages reproduced below:

Table I

Sources of Variability in Speech Acoustics
(after Klatt, 1986)

Type of Variability	Sources of Variability
Ambient Conditions	Background noise, room reverberation, microphone/telephone characteristics
Within-speaker variability	Breathy/creaky voice quality, shifting formants and fundamental frequencies. changing speaking rate, variable degrees of articulatory undershoot, imperfect repetition across tokens of same gesture...
Cross-speaker variability	Differences of dialect, vocal tract length and shapes, detailed articulatory habits
Segment realization variability	Coarticulatory changes, articulatory modification due to stress or duration changes, optional deletions/simplifications in fluent speech...
Word environment variability in continuous speech	Cross-word-boundary coarticulation, phonetic and phonological recoding of words in sentences, changes in word duration due to syntax, pragmatics

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance.
>
> We thus make a fundamental distinction between competence (the speaker-hearer's knowledge of his language) and performance (the actual use of language in concrete situations). Only under the idealization set forth in the preceding paragraph is performance a direct reflection of competence. (Chomsky, 1965, p. 2-3)

To help in this exercise and to find a couple of good working definitions of what "normalization" means, I turned first to *Webster's New Twentieth Century Dictionary* (1983). I also consulted English and English's *Dictionary of Psychological Terms* (1958) and Reber's *Dictionary of Psychology* (1985) for definitions of several other related terms used in perception and cognition. The dictionary definitions of the words "normal," "normalization," and "normalize" are given below:

> ***normal:*** (adj) Conforming with or constituting an accepted standard, model, or pattern; especially, corresponding to the median or average of a large group in type, appearance, achievement, function, development, etc.; natural; standard; regular. (*Webster's*)
>
> ***normal:*** (noun) The usual state, amount, degree, etc.; especially, the median or average. (*Webster's*)
>
> ***normalization:*** (noun) Reduction to a normal or standard state. (*Webster's*)
>
> ***normalize:*** (vt) To bring into conformity with a standard, pattern, model etc. (*Webster's*)

The commonality among all three terms shown above is the proposal that some object, concept or idea is made to conform to a "standard" or a "model" in order to produce or create equivalent forms from diverse inputs. Although not stated explicitly here, there is the further assumption of some loss or reduction of information after normalization is completed.

The definition of the closely related word "standardize," also from *Webster's* is given below:

> ***standardize:*** (vt) To cause to conform to a given standard; to make standard or uniform; to cause to be without variations or irregularities. (*Webster's*)

Here we see an additional property regarding the elimination of variability added to the previous definitions of "normalization" that require conformity with a standard, pattern or model. Thus, at least according to the dictionary definitions considered here, the process of normalization consists of at least two separate

components. The first deals with bringing something into conformity with a preexisting standard so as to produce equivalent forms, while the second has to do with the reduction or elimination of variability thereby creating a standard or normal state.

I believe these definitions of the word "normalization" are fairly good approximations of how this term has been used technically in speech perception over the years to deal with the problems of perceptual invariance and perceptual constancy to non-equivalent forms. Indeed, the intended meanings are consistent with the definitions of the terms "invariance" and "equivalence" given below by Reber:

> ***invariance:*** (noun) Generally, characteristic of that which does not change. The term is most often used with the qualifier relative. That is, few things in this world are truly invariant by some display greater invariance, greater consistency from circumstance to circumstance than others. In general, in the study of perception and learning, those aspects of the stimulus world that display the higher invariances, relative to other aspects, are learned most quickly and easily.
>
> ***equivalence:*** (noun) In general, any relationship between two "things" such that one may be substituted for the other in a particular setting and not alter significantly the situation. The term is often modified so that the particular form of equivalence is specified; e.g., stimulus equivalence refers to two or more stimuli that are sufficiently similar that they evoke the same or nearly the same response, response equivalence refers to similar responses made to similar stimuli, etc.

I have gone through the effort of consulting several dictionaries to make explicit several aspects of the meaning of the term "normalization" that most speech researchers assume more-or-less implicitly in their experimental and theoretical work. In considering these definitions, I believe it is reasonable to conclude that the process of "normalization" produces three consequences in perception. First, there is the generation of equivalent forms from diverse inputs. Normalization is assumed to convert physically different tokens into some common representational format and to store these "standardized" representations in some kind of memory. Second, the process of normalization entails a loss of information and, as a consequence, a reduction in stimulus variability. Finally, although not explicitly stated, I believe that normalization also entails the additional assumption that stimulus variability is an undesirable source of noise in the speech signal which produces perturbations on abstract underlying idealized forms which are assumed to be the true objects of perceptual analysis.

The process of perceptual normalization is, of course, a key component of traditional abstractionist theories of perception, memory and learning (Shankweiler, Strange & Verbrugge, 1977). According to this approach, the stimulus environment is highly impoverished and, as a consequence, the perceiver must rely on top-down knowledge to recognize the intended perceptual object and recover the underlying abstract symbolic form. In the case of speech perception, these objects are the talker's intended linguistic message. Other accounts assume these objects are the talker's intended gestures (Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985). In addition to these two assumptions, there is also the proposal that information processing is selective and results in substantial stimulus reduction because the nervous system cannot maintain all aspects of stimulation in the perceiver's environment. A good summary of the traditional information processing view of perception is given by Haber (1969) below:

> The second assumption, regarding limited information-handling capacities, is also an important one. The problem of limited channel capacity has been clear in the study of perception for most of the history of experimental psychology; witness concepts such as selective attention, and immediate memory span. The nervous system is apparently just not large enough to maintain all aspects of stimulation permanently. What this suggests for information-processing analyses is that we should look for instances in which recoding of information takes place--recoding generally in such a way that some of the content is maintained more explicitly at the expense of the other aspects which are dropped out. The points in time where the recoding occurs should be particularly important ones in the study of information processing, and it is not surprising that most information-processing models refer to these points almost exclusively. (Haber, 1969, p. 4)

Stimulus Variability in Speech Perception

We have completed a number of experiments on the effects of different sources of variability in speech perception and spoken word recognition (Pisoni 1990). Instead of reducing or eliminating variability in the stimulus materials, as most speech researchers have routinely done, we deliberately introduced variability from different talkers to study the effects of these variables on perception (Pisoni, 1992). Our research on this problem began with the observations of Mullennix, Pisoni & 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 same words were produced by 15 different talkers, including male and female voices. Subjects were asked to identify the words using an open-set test format. No feedback was provided. The results were very clear. Across three different signal-to-noise ratios, identification performance was always better for words that were produced by a single talker than for words produced by multiple talkers. Trial-to-trial variability in the speaker's voice affected recognition performance. These findings replicated results reported many years ago by Peters (1955) and Creelman (1957) and suggested that the perceptual system must engage in some form of adjustment or "recalibration" each time a novel voice is encountered during the set of trials using multiple voices.

In a second experiment, we measured naming latencies to the same words presented in both blocked (single-talker) and mixed (multiple-talker) test conditions (Mullennix et al., 1989). We found that subjects were not only *slower* to name words presented in multiple-talker lists but they produced *more errors* when their performance was compared to the same words from single-talker lists. Both sets of findings were surprising at the time because all the test words used in the experiment were highly intelligible when presented in isolation under quiet listening conditions. The intelligibility and naming data from these studies raised a number of additional questions about how the various perceptual dimensions of the speech signal are processed and encoded by the human listener. At the time, we naturally assumed that the acoustic attributes used to perceive voice quality were independent of the more abstract linguistic properties of the signal. However, no one had ever tested this assumption directly.

To assess whether attributes of a talker's voice were perceived independently of the phonetic form of the words, we used a speeded classification task (Mullennix & Pisoni, 1990). Subjects were required to attend selectively to one stimulus dimension (e.g., voice) while simultaneously ignoring another stimulus dimension (e.g., phoneme). Across all conditions, we 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, and subjects could not selectively ignore irrelevant variation on the non-attended dimension. If both perceptual dimensions were processed separately, as we originally assumed, we should have found little, if any, interference from the non-attended dimension. Not only did we find mutual interference, suggesting that the two dimensions, voice and phoneme, 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 to ignore the voice dimension when they had to classify the phonemes.

The results from these perceptual experiments were surprising given our 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 representation of speech in long-term memory. Experiments on serial recall of lists of spoken words by Martin, Mullennix, Pisoni & Summers (1989) and Goldinger, Pisoni & Logan (1991) demonstrated that specific details of a talker's voice are also encoded into long-term memory. Using a continuous recognition memory procedure, Palmeri, Goldinger & 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 another set of experiments, Goldinger (1992) found evidence of implicit memory for attributes of a talker's voice which persists for a relatively long time after perceptual analysis has been completed. He also showed that the degree of perceptual similarity between voices affects the magnitude of the repetition effect in several implicit memory tasks. For example, he found that subjects identified spoken words more accurately when the words were repeated using the same voice in which they had originally been presented than when the words were repeated in a different voice. These findings suggest that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations.

Another series of experiments has been 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, have also shown that the perceptual details associated with differences in speaking rate are not lost as a result of perceptual analysis. In one experiment, Sommers, Nygaard & Pisoni (1994) found that words produced at different speaking rates (i.e., fast, medium and slow) were identified more poorly than the same words produced at only one 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 level.

Other experiments on serial recall have also been completed to examine the encoding and representation of speaking rate in memory. Nygaard, Sommers and Pisoni (1995) found that subjects recall 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. Differences in speaking rate, like those observed for talker variability in our earlier experiments, suggest that perceptual encoding and rehearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by instance-specific sources of variability. If these sources of variability were somehow "filtered out" or "normalized" by the perceptual system at relatively early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments.

Taken together with the earlier results on talker variability, the findings on speaking rate suggest that details of the early perceptual analysis of spoken words are not lost. Instead, they become 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 (Goldinger et al., 1991). Listeners encode speech signals in multiple ways along many perceptual dimensions and the nervous system apparently preserves these perceptual details much more reliably than researchers have believed in the past.

Considering the overall pattern of results, our findings on the effects of talker variability in perception and memory tasks provide support for the proposal that detailed perceptual information about a talker's voice is preserved and that detailed attributes of the listener's stimulus environment are encoded implicitly into long-term memory. At the present time, it is not clear whether there is a single "composite" representation in memory or whether these different attributes are encoded in parallel in separate representations (Eich, 1982; Hintzman, 1986). It is also not clear whether spoken words are encoded and represented in memory as a sequence of abstract symbolic phoneme-like units along with much more detailed episodic information about specific instances and the processing operations used in perceptual analysis. These are important questions for future research on the representation of speech in memory.

Perceptual Learning of Voices

Our findings on talker variability have also encouraged us to examine the 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 speech perception literature despite the obvious relevance to problems of speaker normalization, acoustic-phonetic invariance and the potential application to automatic speech recognition and speaker identification (Kakehi, 1992; Fowler, 1990). Our search of the research literature on talker adaptation revealed only a small number of behavioral studies with human listeners on this topic and all of them appeared in obscure technical reports from the mid 1950's.

To determine how familiarity with a talker's voice affects the perception of spoken words, we had two groups of listeners learn to explicitly identify a set of unfamiliar voices over a nine-day period using common names (e.g., Bill, Joe, Sue, Mary). After the subjects learned to recognize the voices, we presented them with a set of novel words mixed in noise at several signal-to-noise ratios; one group heard the words produced by talkers that they were previously trained on, whereas the other group heard the same words produced by new talkers to whom they had not been previously exposed during the perceptual learning task. In this phase of the experiment, which was designed to measure speech intelligibility, subjects were required to identify the words rather than simply recognize the voices, as they had done in the first phase of the experiment.

The results of the intelligibility experiment are shown in Figure 2 for the two groups of subjects. We found that identification performance for the trained group was reliably better than the control group at each of the signal-to-noise ratios tested. The subjects who had heard novel words produced by familiar voices were able to recognize words more accurately than subjects who received the same novel words produced by unfamiliar voices. Two other groups of subjects were also tested in the intelligibility experiment as controls; however, these subjects did not receive any training in recognizing the voices and were therefore not exposed to any of the stimuli prior to listening to the same set of words in noise. One control group received the set of words presented to the trained experimental group; the other control group

received the words that were presented to the trained control subjects. The performance of these two control groups was not only the same, but was also equivalent to the intelligibility scores obtained by the trained control group. Thus, only the subjects in the experimental group who were explicitly trained on the voices showed an advantage in recognizing novel words produced by familiar talkers.

Insert Figure 2 about here

The findings from this perceptual learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by the same talker. Thus, speech perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that is obtained in an entirely different experimental task-- explicit voice recognition as compared to a speech intelligibility test. Listeners show evidence of tracking changes in their listening environment, specifically, information about a talker's vocal tract transfer function and how it changes over time, and encoding this information in memory for later use when they have to process the linguistic attributes of the signal in order to identify the words in an intelligibility test.

Two additional studies were designed to assess the nature and extent of this kind of perceptual learning (Nygaard & Pisoni, 1995). Subjects in both experiments were trained to recognize a set of ten talkers from sentence-length utterances. In the first experiment, after training was completed, intelligibility was assessed using isolated words produced by familiar and unfamiliar talkers. The aim was to determine if the information learned about a talker's voice from sentences generalizes to the perception of spoken words. The assumption was that training with sentence-length utterances would focus listeners' attention on a different set of acoustic properties than training with isolated words. It was hypothesized that because sentences contain extensive prosodic and rhythmic information in addition to the specific acoustic-phonetic implementation strategies unique to individual talkers, perceptual learning of voices from sentences would require attentional and encoding demands specific to those particular test materials.

In the second experiment, after training on sentences was completed, listeners were given an intelligibility test consisting of sentence-length utterances produced by familiar and unfamiliar talkers. Two issues were addressed here. First, does specific training on sentence-length utterances generalize to similar test materials? Second, are sentence-length utterances which have higher-level semantic and syntactic constraints susceptible to the effects of familiarity with a talker's voice?

All subjects showed continuous improvement over the three days of training. Both groups of subjects identified talkers consistently above chance even on the first day of training and performance rose to nearly 85% correct by the last day of training.

Surprisingly, in the first experiment we found only a small unreliable difference in the word intelligibility test for subjects who heard familiar voices during training compared to the control subjects. These results suggest that perceptual learning of talkers' voices from sentence-length utterances does not generalize to the perception of isolated words.

Insert Figure 3 about here.

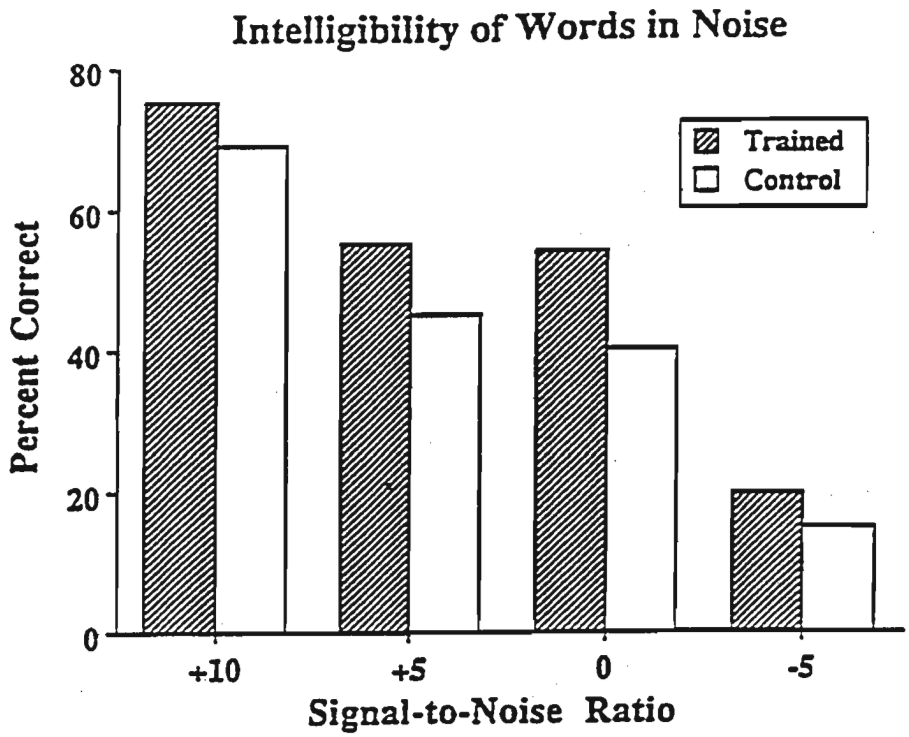


Figure 2. Percent correct word recognition (intelligibility) as a function of signal-to-noise ratio for trained and control subjects on the transfer task administered after voice recognition training was completed. (From Nygaard, Sommers & Pisoni, 1994).

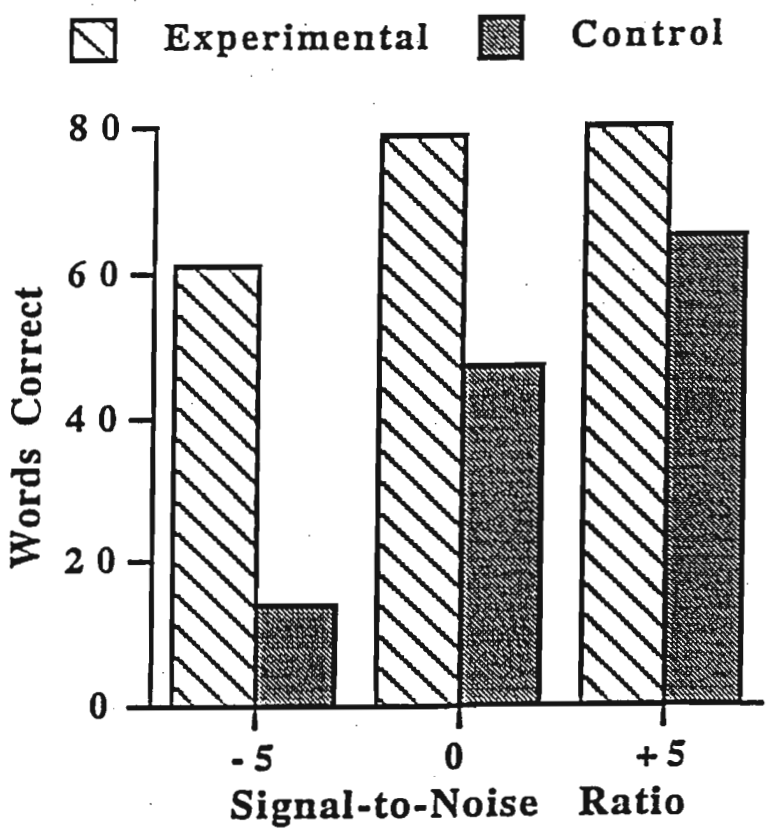


Figure 3. Percent correct recognition of key words in sentences as a function of signal-to-noise ratio for experimental and control groups. (From Nygaard & Pisoni, 1995).

The results obtained in the second experiment which assessed sentence intelligibility at three signal-to-noise ratios were quite different. Figure 3 shows the total number of "key words" that were correctly identified in each sentence. Large differences were observed for the subjects who had heard novel sentences produced by *familiar* talkers. These differences became even larger as the listening conditions became poorer demonstrating transfer of knowledge about the talker's voice from sentence-length utterances.

The results of these experiments suggest that perceptual learning of voices is both talker- and task-specific. Perceptual learning transfers in a task-specific manner suggesting that attention must be directed to learning the specific voice attributes that will be relevant at test. These findings also show that talker-specific effects occur with sentence-length materials which contain higher-level semantic and syntactic constraints. Thus, talker-specific effects operate in a variety of listening situations from isolated words to sentence-length utterances but there are both talker- and task-specific constraints on the transfer of this kind of knowledge.

Familiarity with a talker's voice involves a form of perceptual tuning and adjustment that results in modification of speech and language processing mechanisms. Listeners appear to retain talker-specific information about individual articulatory idiosyncrasies both at the level of acoustic-phonetic implementation in isolated words and at a more global level found in sentence-length utterances. Once again, we see close interactions and dependencies between the indexical properties of speech associated with the talker's voice and the linguistic analysis of the speech signal used in recognizing words in sentences.

What kind of perceptual knowledge does a listener acquire when he or she listens to a speaker's voice and is required to carry out an explicit name-recognition task as our subjects did in these experiments? One possibility is that the analysis procedures or perceptual operations (Kolers, 1973) used to recognize the voices are retained in some type of "procedural" memory system and that these same processing 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 each time a new word was encountered. Another possibility is that the specific instances-- perceptual episodes or exemplars of each talker's voice-- are stored in a composite memory system and then later retrieved during the process of word recognition when new tokens from a familiar talker are presented (Jacoby & Brooks, 1984; Estes, 1994).

Whatever the exact nature of this 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 and sentences produced by the same talker. These findings demonstrate a form of implicit memory for a talker's voice that is distinct from the retention of the individual lexical items and sentences used and the specific perceptual learning task that was employed to familiarize the listeners with the voices (Roediger, 1990). These results provide additional support for the view that the neural representation of spoken words and sentences encompasses both a symbolic description of the utterance in terms of a phonetic representation *and* additional information about the structural description of the source characteristics of the specific talker. Thus, speech perception appears to be carried out in a "talker-contingent" manner; indexical and linguistic properties of the speech signal are apparently closely interrelated and are not dissociated in perceptual analysis (Nygaard et al., 1994; Nygaard & Pisoni, 1995). One of the important discoveries of studies done within the framework of nonanalytic cognition is that many aspects of perception, learning and categorization may not be available to conscious recollection but nevertheless will affect perceptual analysis and memory processes in a variety of ways. Thus, sweeping conclusions like those of Halle (1985)

and Brown (1990) about the retention of voice-specific information in speech will obviously have to be revised in light of the present set of findings on perceptual learning of voices:

> ...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 of the person who taught us the word or the rate at which the word was pronounced. Not only voice quality, speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance are discarded in the course of learning a new word. (Halle, 1985, p. 101)

> Clearly most of the time anyone is listening to English being spoken, he is listening for the meaning of the message-- not to how the message is being pronounced. Indeed if you listen to how the words are spoken it is very unlikely that you can simultaneously understand what it is that is being said. On the whole people do not listen critically to the way the message is pronounced. The odd glottal stop or unusual pronunciation of a word may strike the listener, but most of the time he is busy abstracting the meaning of the message, and preparing his own mental comments on it. This is why most people are quite unaware of how English is actually spoken. (Brown, 1990, p. 3)

Exemplar-based Approach to Speech Perception

The approach to speech perception proposed here relies heavily on earlier work by Jacoby and Brooks on non-analytic concept formation. Their studies on categorization and memory have provided evidence for the encoding and retention of episodic information and the details of perceptual analysis (Jacoby & Brooks, 1984; Brooks, 1978; Tulving & Schacter, 1990). According to this approach, stimulus variability is informative for perceptual analysis. Memory involves encoding both specific instances, and the specific processing operations used during recognition (Kolers, 1973; Kolers, 1976). The major emphasis of this view of cognition is on particulars, rather than abstract generalizations or symbolic coding of the stimulus input into idealized categories. We believe that the problems of variability and invariance found in speech perception can be approached in a fundamentally different way by nonanalytic or instance-based accounts of perception and memory. Moreover, this view of cognition provides insights into several long-standing theoretical issues in speech and offers a potentially useful way of dealing with findings that show that listeners encode very fine details of their stimulus environment.

The findings from studies on non-analytic cognition are directly relevant to theoretical questions about the nature of perception and memory for speech and to assumptions about abstractionist representations based on formal linguistic analyses. When the criteria used for postulating episodic or nonanalytic representations are examined carefully, it becomes clear that speech signals display a number of distinctive properties that make them excellent candidates for this approach (Jacoby & Brooks, 1984; Brooks, 1978). These criteria are summarized in the sections below.

High Stimulus Variability.

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,

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 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.

Complex Category Relations.

The 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 meaning contrast, phonetic similarity and distributional characteristics. In traditional taxonomic linguistics, for example, the concept of a phoneme is used in a number of different ways, as shown by the following definitions from Gleason (1961) given in Table II:

Insert Table II about here.

Thus, the perceptual categories used in speech display complex relations that place a number of strong constraints on the class of models that can account for these operating principles. These categories cannot be defined clearly and easily by simple rules and must be identified via a set of relations involving sound contrast and lack of contrast in meaning within a specific phonological system.

Incomplete Information.

Spoken language is a highly redundant symbolic system which has evolved to maximize transmission of linguistic information. In the case of speech perception, research has demonstrated the existence of multiple speech cues for almost every phonetic contrast. While these speech cues are, for the most part, highly context-dependent, they also provide reliable information that can facilitate comprehension of the intended message when the signal is presented under degraded conditions. This feature of speech perception permits very high rates of information transmission even under poor listening conditions.

High Analytic Difficulty.

Speech is inherently multidimensional in nature. As a consequence, many quasi-independent articulatory attributes can be mapped onto the phonological categories of a specific language. 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 formalize a set of explicit rules that can successfully map speech cues onto discrete phoneme categories. 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.

Perceptual Spaces for Perception and Production.

Among category systems, speech appears to be unique because of the close interrelations between speech production and speech perception. Speech exists simultaneously in several very different domains: the acoustic domain, the articulatory domain and the perceptual domain. Although the relations among these domains are complex, they are not arbitrary and reflect properties of a unitary articulatory event. And, even within the domain of production and articulation, speech is encoded simultaneously in the optical display of the talker's face and the acoustic signal generated by the vocal tract. The sound contrasts used in a given language function within a common linguistic system that is shared by both production and

Table II

Definitional Characteristics of the Phoneme

(from Gleason, 1961)

- The phoneme is the minimum feature of the expression system of a spoken language by which one thing that may be said is distinguished from any other thing which might have been said.

- A phoneme is a class of sounds...There is no English phoneme which is the same in all environments, though in many phonemes the variation can easily be overlooked, particularly by a native speaker.

- A phoneme is a class of sounds which: (1) are phonetically similar and (2) show certain characteristic patterns of distribution in the language or dialect under consideration.

- A phoneme is one element in the sound system of a language having a characteristic set of interrelationships with each of the other elements in that system.

- The phoneme cannot, therefore, be acoustically defined. The phoneme is instead a feature of language structure. That is, it is an abstraction from the psychological and acoustical patterns which enables a linguist to describe the observed repetitions of things that seem to function within the system as identical in spite of obvious differences...The phonemes of a language are a set of abstractions...

perception. Thus, the phonetic contrasts generated in speech production by the vocal tract are precisely the same acoustic differences that are distinctive in perceptual analysis (Stevens, 1972). As a result, any theoretical account of speech perception must also take into consideration aspects of speech production and acoustics, as well as the multimodal relations between the auditory and visual correlates of speech.

In learning the sound system of a language, children not only develop abilities to discriminate and identify sounds, but they also learn to control the motor mechanisms used in articulation to generate precisely the same phonetic contrasts in speech production to which they have become attuned in perception. One reason that the developing perceptual system might preserve very fine phonetic details, 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). This perceptuomotor skill would provide children with an enormous benefit in acquiring the phonology of the local dialect from speakers they are exposed to early in life.

In short, when properties of speech signals are examined more closely and when the criteria for nonanalytic cognition are applied and evaluated, it becomes plausible to assume that very detailed information about specific instances in speech perception might be stored in memory. In contrast to a traditional symbolic rule-based representational approach, listeners may store a very large number of specific instances or perceptual episodes and then use them in an analogical rather than analytic way to perceive and categorize novel stimuli (Brooks, 1978; Whittlesea, 1987). Recent findings from studies on talker variability in speech perception provide support for this conclusion.

The traditional view of "normalization" assumes that perceptual constancy is achieved via a process or set of procedures involving a reduction in variability and a "loss" of detailed stimulus information. But this view is wrong. Human listeners preserve fine phonetic details and appear to encode other aspects of their listening environment, including indexical information about the talker's voice, gender, dialect, speaking rate, affect and speaking style, among other nonlinguistic attributes in their surroundings. We believe these are important new insights into speech perception and spoken language processing that will open up new areas of research on the study of sources of variability in speech perception.

Summary and Conclusions

This paper began with a review and discussion of the term "normalization" as it has been used in the field of speech perception. An examination of dictionary definitions revealed a number of distinctive properties of normalization that were generally consistent with the way this term is used technically in the field of speech perception to deal with perceptual constancy and invariance. Three consequences of normalization were addressed in greater detail. First, we discussed the assumption that normalization produces equivalent forms. Second, we examined the proposal that normalization entails a loss or reduction of information. And third, we considered the property of normalization that involves elimination of variability among specific instances. All three attributes of "normalization" are assumed in traditional abstractionist accounts of speech perception and spoken word recognition, and all three of these attributes have played important roles in theorizing about the nature of the perceptual and neural mechanisms used in speech perception.

In the second section of this paper we summarized findings from a series of recent experiments on the contribution of stimulus variability to speech perception and spoken word recognition. Other results on learning novel voices and the contribution of voice information to word recognition and sentence

intelligibility were also described. These studies with normal-hearing listeners demonstrate that fine details of the listener's perceptual environment are encoded and stored in memory and used at a later time during the perceptual analysis of novel stimuli. The results on perceptual learning of voices demonstrate that the linguistic and indexical properties of speech are not maintained independently as separate channels of information by the nervous system. Interactions between these two properties of the speech signal occur early on in perceptual analysis and produce mutual dependencies that affect the efficiency of subsequent perceptual operations.

In the third section of this paper, these observations and recent findings were considered together within the framework of recent developments in perceptual learning, memory and categorization, specifically, the nonanalytic or instance-based approach to cognition which emphasizes the episodic encoding of specific details of the stimulus environment. The studies on talker- and rate-variability provide important information about speech perception and spoken word recognition and have raised a set of new questions for future research. These findings encourage us to look at several of the long-standing problems in speech perception in very different ways than we have been able to do in the past.

The present research findings and our re-examination of the term normalization suggests several new directions for research on speech perception. Exemplar-based or episodic models of categorization provide novel solutions to the problems of invariance, variability and perceptual normalization that have been difficult to resolve with traditional models of speech perception that were motivated by formal linguistic analyses of spoken language. These problems can now be approached in quite different ways when viewed within the general framework of nonanalytic or instance-based models of cognition which provide alternative ways of dealing with stimulus variability which has been one of the most difficult problems in the field of speech perception and spoken language processing.

References

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

Brown, G. (1990). *Listening to Spoken English*. Second Edition. New York: Longman.

Bryd, D. (1992). Sex, dialects, and reduction. *ICSLP 92 Proceedings*, pp. 827-830.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

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

Eich, J.E. (1982). A composite holographic associative memory model. *Psychological Review*, **89**, 627-661.

English, H.B. & English, A.C. (1958). *A Comprehensive Dictionary of Psychological and Psychoanalytical Terms: A Guide to Usage*. New York: Longmans, Green and Co.

Estes, W.K. (1994). *Classification and Cognition*. New York: Oxford University Press.

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

Fowler, C.A. & Rosenblum, L.D. (1991). The perception of phonetic gestures. In I.G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33-59). Hillsdale, NJ: Erlbaum.

Gleason, H.A. (1961). *An Introduction to Descriptive Linguistics*. New York: Holt, Rinehart & Winston.

Goldinger, S.D. (1992). Words and voices: Implicit and explicit memory for spoken words. *Research on Speech Perception Technical Report No. 7*, Indiana University, Bloomington, IN.

Goldinger, S.D., Pisoni, D.B. & 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.

Haber, R.N. (1969). *Information-Processing Approaches to Visual Perception*. New York: Holt, Rinehart & Winston.

Halle, M. (1956). *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday, 11 October 1956*. The Hague: Mouton.

Halle, M. (1985). Speculations about the representation of words in memory. In V.A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 101-104). New York: Academic Press.

Hintzman, D.L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, **93**, 411-423.

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.

Hockett, C.F. (1955). *A Manual of Phonology*. Baltimore: Waverly Press.

Huttenlocher, D.P. & Zue, V.W. (1984). A model of lexical access based on partial phonetic information. *Proceedings ICASSP-84*, 1-4.

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). New York: Academic Press, pp. 1-47.

Kakehi, K. (1992). Adaptability to differences between talkers in japanese monosyllabic perception. In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 135-142). Tokyo, Japan: Ohmsha.

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

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

Klatt, D.H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process* (pp. 169-226). Cambridge, MA: MIT Press.

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

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

Liberman, A.M. & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

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

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.

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

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

Nygaard, L.C. & Pisoni, D.B. (1995). Talker- and task-specific perceptual learning in speech perception. *Proceedings of the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden, August, 1995*, pp. 194-197.

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. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics*, **57**, 989-1001.

Palmeri, T.J., Goldinger, S.D. & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition, **19***, 1-20.

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

Pisoni, D.B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, pp. 1399-1407.

Pisoni, D.B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, pp. 587-590.

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.

Reber, A.S. (1985). *The Penguin Dictionary of Psychology*. Harmondsworth, UK: Penguin Books.

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

Shankweiler, D., Strange, W. & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (pp. 315-345). Hillsdale, NJ: Erlbaum.

Shipman, D.W. & Zue, V.W. (1982). Properties of large lexicons; Implications for advanced isolated word recognition systems. *Proceedings ICASSP-82*, pp. 546-549.

Sommers, M.S., Nygaard, L.C. & Pisoni, D.B.(1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, 1994, **96**, 1314-1324.

Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory acoustic data. In E.E. David, Jr. and P.B. Denes, (Eds.) *Human communication: A unified view* (pp. 51-66). McGraw-Hill, New York.

Studdert-Kennedy, M. (1974). The perception of speech. In T.A. Sebeok (Ed.) *Current Trends in Linguistics*, The Hague: Mouton, pp. 2349-2385.

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

Tulving, E. & Schacter, D.L. (1990). Priming and human memory systems. *Science*, **247**, 301-306. *Webster's New Twentieth Century Dictionary*. (1983). New York: Simon & Schuster.

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

Zue, V.W. (1985). The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, **73**, 11, 1602-1615.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Some Considerations in Evaluating Spoken Word Recognition by Normal-Hearing and Cochlear Implant Listeners I: The Effects of Response Format[1]

Mitchell S. Sommers,[2] Karen I. Kirk,[3] and David B. Pisoni[3]

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

[1]This research was supported by NIH-NIDCD Grants DC-00064 to Indiana University School of Medicine and DC-00111 to Indiana University at Bloomington.

[2]Now at Department of Psychology, Washington University, St. Louis, MO.

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

Abstract

The purpose of the present investigation was to assess the validity of using closed-set response formats to measure two cognitive processes essential for recognizing spoken words--perceptual normalization (the ability to accommodate acoustic-phonetic variability) and lexical discrimination (the ability to isolate words in the mental lexicon). In addition, the experiments were designed to examine the effects of response format on evaluation of these two abilities in subject populations with differing degrees of sensory impairment. Speech recognition performance of normal-hearing (NH), noise-masked normal hearing (NMNH), and cochlear implant (CI) listeners was measured using both open- and closed-set response formats under a number of experimental conditions. To assess talker normalization abilities, identification scores for words produced by a single talker were compared with recognition performance for items produced by multiple talkers. To examine lexical discrimination, performance for words that are phonetically similar to many other words (hard words) was compared with scores for items with few phonetically-similar competitors (easy words). Open-set word identification for all subjects was significantly poorer when stimuli were produced in lists with multiple talkers, compared with conditions in which all of the words were spoken by a single talker. Open-set word recognition was also reduced for "easy" compared with "hard" words. Closed-set tests, in contrast, failed to reveal the effects of either talker variability or lexical difficulty even when the response alternatives provided were systematically selected to maximize confusability with target items. These findings suggest that, although closed-set tests may provide important information for use in clinical assessment of speech perception abilities, they may not adequately evaluate a number of cognitive processes that are necessary for recognizing spoken words. The parallel results obtained across all subject groups indicate that normal-hearing, hearing-impaired, and cochlear implant listeners engage similar perceptual operations to identify spoken words. Implications of these findings for the design of new test batteries that can provide comprehensive evaluations of the individual capacities needed for processing spoken language are discussed.

Some Considerations in Evaluating Spoken Word Recognition by Normal-Hearing and Cochlear Implant Listeners I: The Effects of Response Format

Introduction

The ability to accurately recognize spoken words is critically dependent upon the integration of a number of sensory, perceptual and cognitive capacities (Klatt, 1989; Pisoni, 1985). Identification and discrimination of phonetic features, segmentation of the speech waveform, compensation for talker differences and accessing words from the mental lexicon are just some of the component operations necessary for transforming acoustic speech sounds into meaningful linguistic perceptions. Although a number of instruments are currently available for both clinical and experimental assessment of speech perception abilities (Dorman, 1993; Dowell, Brown, & Mecklenberg, 1990; Geers & Brenner, 1994; Rosen et al., 1985), relatively little systematic empirical research has been directed at establishing the specific perceptual and cognitive capacities that each of these tests measure. Consequently, little is known about either the utility or limitations of different assessment procedures for evaluating the individual abilities necessary to recognize spoken words.

For example, one of the most common methods for measuring speech discrimination is the closed-set test format. Closed-set tests such as the Modified Rhyme Test (MRT) (House, Williams, Hecker, & Kryter, 1965) and portions of the Minimum Auditory Capabilities test (Owens, Kessler, & Schubert, 1981) require listeners to select one of several provided response alternatives that best matches a presented stimulus. This response format has been extremely useful for assessing certain perceptual capacities necessary to understand spoken language, especially in clinical populations with severely impaired auditory functions. Closed-set tests have been used to provide information about the discrimination of phonetic features, prosodic characteristics, and timing aspects of speech signals (Blamey, Dowell, Brown, Clark, & Seligman, 1987; Geers & Brenner, 1994, Tyler, Lowder, Otto, Preece, Gantz, & McCabe, 1984). In addition, closed-set response formats have been shown to be sensitive to word frequency (Elliott, Clifton, & Servi, 1983) and to differences in linguistic background (Garstecki & Wilkin, 1976).

One potential disadvantage of closed-set measures, however, is that they may not adequately simulate the same cognitive demands that individuals confront in natural listening environments. Closed-set speech perception tests typically contain highly articulated tokens of words produced by a single talker and provide listeners with a restricted set of response alternatives. In contrast, real-world listening environments often contain poorly articulated stimuli spoken by multiple talkers with only minimal restrictions on potential response candidates. These differences between natural listening conditions and closed-set tests may limit the ability of closed-set formats to predict speech perception performance in everyday conversational situations. The purpose of the present studies, therefore, was to establish the validity of using closed-set tests to assess two cognitive abilities necessary for recognizing spoken words--normalizing for talker differences and isolating words in the mental lexicon. In addition, the experiments were designed to determine whether evaluation of these two capacities with closed-set formats differed in listeners with varying degrees of simulated or actual sensory impairment.

To achieve these goals, normal-hearing (NH), noise-masked normal hearing (NMNH), and cochlear-implant (CI) listeners were tested using both open- and closed-set formats on their ability to accommodate changes in talker characteristics and to isolate words in long-term lexical memory. These two

capacities were selected as the focus of our investigations because they have been shown to be critical components of the earliest stages of spoken language processing (Luce, 1986; Luce, Pisoni & Goldinger, 1990; Mullennix, Pisoni & Martin, 1989; Sommers, Nygaard & Pisoni, 1994). Results indicating that open- and closed-set tests are equally effective in their ability to evaluate these two processes would provide empirical support for additional perceptual capacities that can be assessed with closed-set formats. In contrast, differential results using the two response formats would suggest that closed-set measures of spoken word recognition may be limited in their ability to assess one or more perceptual capacities used in speech perception.

The Importance of Talker Normalization and Lexical Discrimination for Spoken Word Recognition

The ability to rapidly adapt to changes in talker characteristics is an essential aspect of processing spoken language because differences in the size and shape of vocal-tracts results in a many-to-one mapping between acoustic speech signals and phonetic perceptions. Thus, the same word produced by a man, a woman and a child will have dramatically different acoustic properties due to differences in the physical characteristics of the talkers' vocal tracts (Peterson & Barney, 1952). Normal-hearing listeners, however, generally have little difficulty recognizing these distinct speech signals as phonetically equivalent (i.e., as instances of the same word). This ability to maintain perceptual constancy in the face of extensive acoustic-phonetic variability has traditionally been attributed to a stage of processing, referred to as perceptual normalization, during which listeners derive standardized phonetic representations that can then be matched to canonical forms stored in long-term memory (Johnson, 1990; Joos, 1948; Nearey, 1989; Pisoni, 1993).

The importance of talker normalization for spoken word recognition has now been well established (Mullennix & Pisoni, 1990; Mullennix et al., 1989; Sommers et al., 1994). For example, several investigators (Mullennix et al., 1989; Sommers et al., 1994) have examined the effects of requiring listeners to compensate or normalize for talker differences by comparing speech recognition performance for word lists produced by single and multiple talkers. The general finding from these studies is that multiple-talker contexts produce a 10-15 percent reduction in identification scores relative to the identical words produced by a single talker. One hypothesis that has been proposed to account for this finding is that the greater demand for talker normalization in the multiple-talker contexts diverts limited processing resources from perceptual operations used in phonetic identification (Martin, Mullennix, Pisoni & Summers, 1989; Mullennix et al., 1989; Sommers et al., 1994). That is, the mixed-talker condition requires more processing to maintain perceptual constancy and, consequently, listeners have fewer cognitive resources available for identifying spoken words. If assessment instruments employing closed-set formats fail to engage all of the resources needed for speech perception under more natural (open-set) conditions then they may not be sensitive to the effects of talker variability or other factors that affect the acoustic-phonetic properties of speech signals.

Once a standardized phonetic representation has been obtained through the normalization process, it must be matched to idealized representations stored in the mental lexicon. However, matching every representation derived from incoming speech signals to one of the tens of thousands of representations stored in long-term memory would place considerable, and almost certainly excessive, demands on the speech perception system. Speech researchers have therefore proposed several mechanisms that function to restrict the number of items within the mental lexicon that are compared with incoming speech waveforms (Luce et al., 1990; Marslen-Wilson, 1987). One such proposal, the Neighborhood Activation Model of spoken word recognition (NAM) (Luce et al., 1990), assumes that words in the mental lexicon are organized into similarity neighborhoods. A similarity neighborhood, according to the model, consists of a

target word and all other words that can be created from that item by adding, deleting, or substituting a single phoneme. Thus, the neighborhood for the word "CAT" would include the words (referred to as neighbors) "COT," "KIT," "CAB," and "SCAT" (and all other words differing from CAT by a single phoneme). Luce et al. (1990) suggested that speech signals activate only items within a single similarity neighborhood and that word recognition occurs by selecting among this restricted set of activated neighbors.

One prediction from the model that has received considerable empirical support (Kirk, Pisoni, & Osberger, 1995; Luce, et. al., 1990; Sommers, in press) is that the difficulty of isolating a target word from its neighbors will be determined by both the number of words within the neighborhood (neighborhood density) and the average frequency of those neighbors (neighborhood frequency) as determined by word frequency norms (Kucera & Francis, 1967). Specifically, the NAM predicts that words from high-density, high-frequency neighborhoods should be identified less accurately than words from low-density, low-frequency neighborhoods. That is, words with many similar sounding, high-frequency neighbors (lexically hard words) will be more difficult to isolate than those that have only a few, low-frequency competitors (lexically easy words). Consistent with this prediction, a number of investigations have reported that both the speed and accuracy of processing spoken words is reduced for lexically hard, compared with lexically easy items (Cluff & Luce, 1990; Luce et al., 1990; Sommers, in press).

Taken together, the results of previous studies suggest that both talker variability and lexical difficulty can significantly affect spoken word recognition performance. However, these results have been obtained exclusively using open-set formats. Given the extensive use of closed-set measures in clinical assessments, it is essential to establish whether closed-set tests are also sensitive to the effects of talker variability and lexical difficulty. Differential effects of stimulus variability and lexical difficulty as a function of test format would indicate that the two types of assessment procedures measure distinct perceptual capacities in speech perception. Furthermore, comparing the effects of response format in both hearing-impaired and normal-hearing subject populations will provide an indication of whether these groups engage similar processing mechanisms in recognizing spoken words.

Experiment 1A

The purpose of Experiment 1A was to examine the effects of talker variability and lexical difficulty on perceptual identification in normal-hearing (NH), noise-masked normal-hearing (NM), and cochlear implant (CI) subjects using both open- and closed-set response formats. The rationale for this approach is that it provided a methodology for examining the independent effects of hearing loss and cochlear implants on the perceptual operations used to recognize spoken words in open- and closed-set tests. Differences between the normal-hearing and noise-masked normal groups as a function of test format would suggest that reduced absolute sensitivity alters the mechanisms that listeners engage to recognize spoken words in open- and closed-set measures. Similarly, qualitative performance differences between the noise-masked normals and CI listeners would suggest that, independent of hearing loss, processing limitations imposed by cochlear implants change the perceptual operations used for speech perception.

Method

Subjects

Four groups of subjects were tested in Experiment 1A. Group 1 consisted of eight adult cochlear implant patients who were seen at Indiana University Medical Center as part of their regularly scheduled

postimplant appointments. Seven of the CI participants used the Nucleus 22-channel implant and one used the Clarion implant. Etiologies for the profound deafness in the CI patients were: meningitis (2), Meniere's disease (3), and unknown (3). Mean age at implantation was approximately 40 years and mean length of device use was 3.6 years.

The remaining three groups of subjects all consisted of normal-hearing adult listeners tested under different signal-to-noise (S/N) ratios. Subjects in groups 2-4 were recruited from the Washington University student population and surrounding community. All had pure-tone air conduction thresholds of less than 20 dB HL for octave frequencies between 250 and 8000 Hz. Group 2 consisted of 11 listeners tested in quiet. Groups 3 and 4 were composed of 12 subjects each and stimuli were presented at S/N ratios of +5 and -5 dB, respectively. These two S/N ratios were designed to simulate moderate and severe hearing losses in the normal-hearing listeners.

Stimulus Materials

Stimuli for the experiment were taken from a digital database containing 300 monosyllabic words from the Modified Rhyme Test (House et al., 1965) recorded by 20 different talkers (10 male and 10 female). A total of 200 different words were selected from this database for use in Experiment 1A. Half of the items were produced by one of the 20 talkers and constituted the stimuli for the single-talker conditions. The remaining 100 items were used in the multiple talker conditions and consisted of words produced by 10 different talkers (5 male and 5 female) with each talker contributing 10 items. Previous studies with stimuli from this database have reported that the intelligibility of the twenty talkers did not differ significantly (Martin et al., 1989; Mullennix et al., 1989).

In addition to dividing the 200 stimuli into single- and multiple-talker conditions, the words were further divided on the basis of lexical difficulty. Half of the words for the single- and multiple-talker conditions were lexically easy and half were lexically hard, as defined by the Neighborhood Activation Model. The lexically easy words had an average neighborhood density of 11.3 (i.e., on average each word had approximately 11 neighbors) and an average neighborhood frequency of 43.4 (occurrences per million words). The corresponding values for the lexically hard words were a mean neighborhood density of 26 and a mean neighborhood frequency of 255.8. Thus, the easy words were selected from low-density, low-frequency neighborhoods, while the hard words were from high-density, high-frequency neighborhoods. Each of the ten talkers used in the multiple-talker conditions contributed 5 easy and 5 hard words.

Procedure

All subjects were tested using a repeated-measures design. Listeners received 100 items in both the open- and closed-set tests. Within each response format, two blocks of stimuli were presented. In one 50-item block, all of the words were produced by a single male talker. Half of these (25 items) were lexically easy and half were lexically difficult. In the other 50-item block for each format, the words were produced by all ten talkers and the voice presented on a given trial was selected randomly. Again, half of the mixed-talker words were lexically easy and half were lexically difficult. Order of presentation for the two blocks in each format was counterbalanced. Two versions of the test were constructed such that words presented in the open-set format on one form were presented in the closed-set format on the second form. Approximately half of the subjects in each group received each form of the test.

Open-set tests

For the CI listeners in Group 1, the digitized stimuli were converted to analog signals using a 12-bit D/A converter and a 10-kHz sampling rate. The signals were low-pass filtered at 4.5 kHz and recorded on audio tape. The stimuli were presented free field to CI patients sitting in a double-walled sound attenuating booth using a tape recorder (Nakamichi, CR-1A). Subjects sat approximately 1.5 m from the transducer and wrote their responses on answer sheets provided by the experimenter. The inter-stimulus-interval was 5 seconds which was generally sufficient for listeners to complete writing their responses. If the experimenter noticed a subject taking longer than normal to complete a response, the recorder was stopped until the listener completed their answer.

For the normal-hearing listeners (Groups 2-4) stimuli were converted to analog signals (12-bit D/A converter, 10-kHz sampling rate) and presented binaurally over matched and calibrated TDH-39 headphones. For the two subject groups tested with noise, the noise was generated using a noise generator (Grason Stadler, 901B) and was gated on and off coincident with presentation of the speech stimuli. The three groups of normal hearing listeners all responded by typing their responses on a keyboard connected to a CRT terminal.

Closed set tests

The procedures and equipment for testing listeners with closed-set response formats were identical to those used with the open-set tests except that all participants responded by circling one of six alternatives on a response sheet. The six response choices included the target item and five foils that differed from the target by a single phoneme. These were the same response alternatives used in the standard MRT test.

Results and Discussion

A mixed-design analysis of variance (ANOVA) was conducted on the identification scores for all four subject groups with subject group and test version (version 1 or 2) as between-subjects factors and response format (open vs. closed), lexical difficulty (easy vs. hard) and stimulus variability (single vs. multiple talker) as repeated measures. No significant main effects or interactions were observed for test version. Therefore, the remaining analyses were conducted with data from the two versions combined. Significant main effects were obtained for subject group ($F[3, 39] = 991.1$, $p < .001$), and response format ($F[1, 39] = 1358.5$, $p < .001$). As expected, performance was poorer for open- than for closed-set tests and the differential hearing loss in the three impaired groups (actual loss in the case of the CI listeners and simulated losses in the case of the NH(+5) and NH(-5) listeners) produced systematic decrements in identification performance. Consistent with previous findings (Luce et al., 1990; Mullennix et al., 1989; Sommers et al., 1994) identification scores were also affected by lexical difficulty ($F[1, 39] = 24.4$, $p < .001$) and stimulus variability ($F[1, 39] = 77.3$, $p < .001$); easy words were identified with greater accuracy than hard words and identification scores were higher for single-talker word lists than for multiple talker word lists.

To determine whether open- and closed-set response formats were differentially sensitive to the effects of stimulus variability and lexical difficulty, several of the interactions obtained in the overall ANOVA were examined. First, a significant two-way interaction between stimulus variability and response format was observed ($F[1, 39] = 30.2$, $p < .001$). Figure 1 displays identification scores for single and multiple talkers (collapsed across lexical difficulty) as a function of test format. Tukey HSD post-hoc

analyses indicated that all of the subject groups except the normal-hearing listeners tested in quiet[4] exhibited significantly poorer identification scores for multiple-talker lists in the open-, but not the closed-set response format ($p < .05$ for all open-set comparisons; $p > .2$ for all closed-set comparisons). It is important to note that this pattern of results was obtained despite significant differences in overall performance levels among the subject groups. Thus, although the NH(-5) and CI groups had significantly lower identification scores than the NH(+5) listeners, they nevertheless exhibited effects of talker variability in open-, but not closed-set tests. These results demonstrate that an important limitation on closed-set test formats is that they may not be sensitive to the effects of at least one source of acoustic-phonetic variability present in many natural listening environments, namely the variability that results from changes in talker (vocal-tract) characteristics.

Insert Figure 1 about here

A second important finding from this study was a reliable response format x lexical difficulty ($F[1, 39] = 111.1$, $p < .001$) interaction. Figure 2 displays identification scores for lexically easy and lexically hard words (collapsed across single- and multiple-talker conditions) as a function of test format. Tukey HSD post-hoc analyses indicated that, with the exception of the normal-hearing listeners tested in quiet (see footnote 4), identification performance in open-set formats was significantly poorer for lexically hard compared with lexically easy words ($p < .05$ for all comparisons except the NH(Q) group). In contrast, examination of the data for the closed-set response formats revealed that none of the groups exhibited differences between easy and hard words when response alternatives were provided in the closed-set test (the effects of lexical difficulty for the NH(+5) group did approach significance ($p < .1$) in the closed-set format). Thus, the findings are similar to the results obtained with talker variability in that lexical difficulty influenced identification performance in open- but not closed-set tests.

Insert Figure 2 about here

One unanticipated finding revealed in the ANOVA was a significant subject group x format ($F[3, 39] = 100.8$, $p < .001$) interaction; the effects of changing from open- to closed-set formats differed significantly across the four subject groups. To further explore this result, ratios of recognition performance in open and closed-set formats were computed for all listeners except those in the NH(Q) group (performance for this group was already close to ceiling in the open-set test). Post-hoc analyses computed on these ratios indicated that the CI patients exhibited the greatest benefit of changing from open- to closed-set measures ($p < .05$). One explanation for this finding is that the impoverished acoustic signal CI patients receive from their devices may force them to rely on deriving broad phonetic categories (Shipman & Zue, 1982) rather than obtaining detailed phonetic information in recognizing spoken words. Such a strategy will be most beneficial in closed-set tests where the response alternatives are already given. Koch, Carrell, Tremblay and Kraus (1996) have recently provided evidence to suggest that the ability to group speech sounds into broad phonetic classes is highly predictive of CI patients' ability to understand everyday speech and may, in part, explain why implant patients can have relatively good word intelligibility despite demonstrating poor phonetic perception. Thus, closed-set tests may be particularly beneficial to CI

[4]The absence of a significant difference in the open-set format for this group is most likely due to subjects approaching ceiling-level performance.

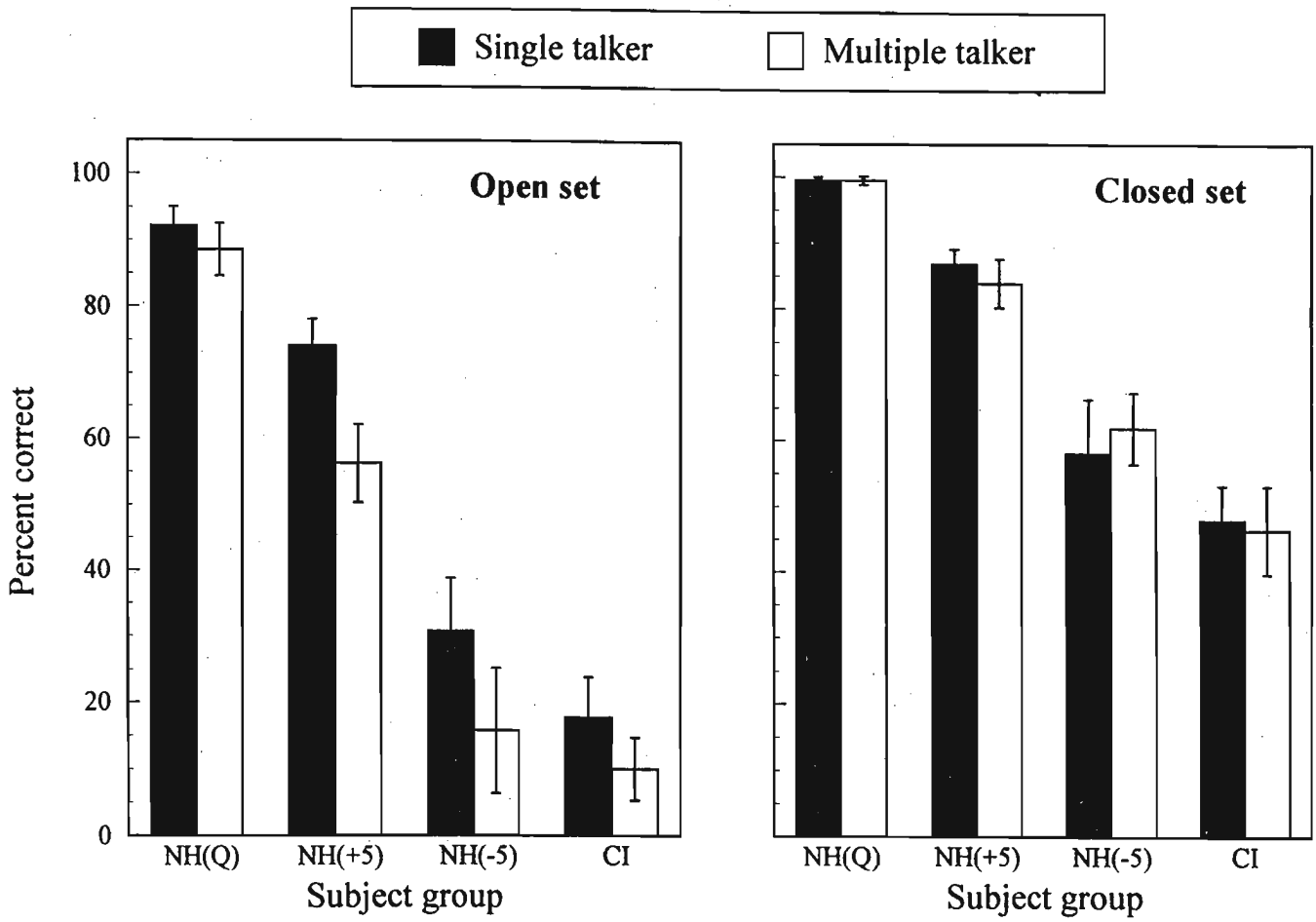


Figure 1: Comparison of identification scores for words produced by single (dark bars) and multiple (open bars) talkers collapsed across lexical difficulty. The left side of the figure displays data obtained with an open-set response formant and the right side shows results obtained with a closed-set test. The four subject groups are normal-hearing tested in quiet (NH), normal-hearing tested at +5 (NH(+5)) and -5 (NH(-5)) signal-to noise ratios and cochlear implant patients (CI).

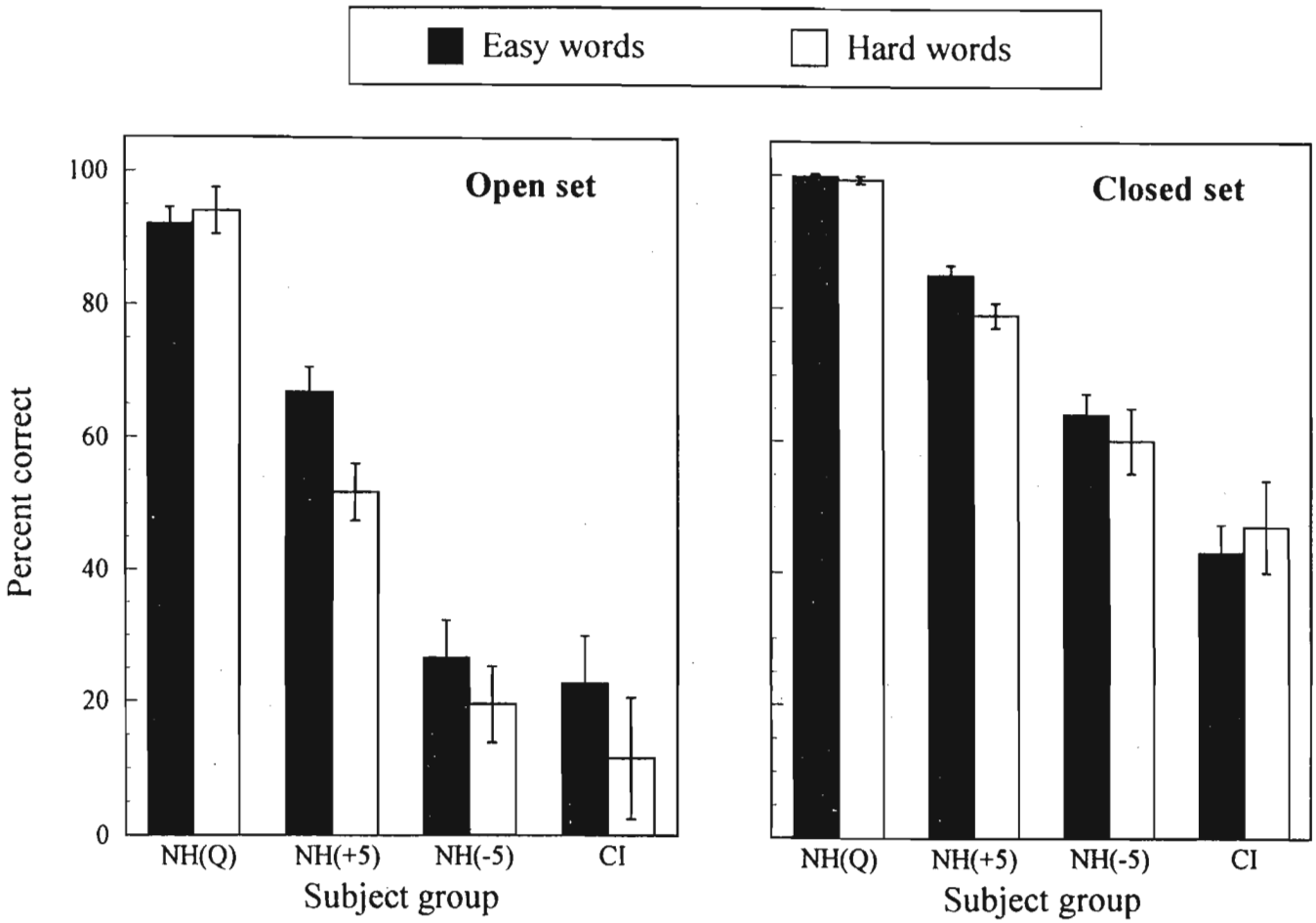


Figure 2: Same as Figure 1 except the data show the effects of lexical difficulty (easy vs. hard words) collapsed across single and multiple talkers.

patients because the response alternatives provided allow them to map their broad phonetic classifications onto individual spoken words.

Examination of the remaining two-way interactions revealed a significant subject group x variability ($F[3, 39] = 3.8$, $p < .05$) effect; the reduction in identification scores resulting from increased stimulus variability was significantly larger for the two simulated loss groups than for either the CI or NH(Q) listeners (as noted, the small effects for the NH(Q) group are probably due to listeners approaching ceiling level performance). In addition, a reliable three-way subject group x format x variability ($F[3, 39] = 3.6$, $p < .05$) interaction was obtained indicating that the differential effects of stimulus variability across subject groups was limited to the open-set format. None of the remaining effects were statistically reliable.

Although the present study failed to find effects of lexical difficulty and talker variability in closed-set formats, this result may have been due, in part, to the specific response alternatives used in the MRT. If the five foils used with each item in the closed-set tests were not sufficiently confusable with the target word, then the absence of significant lexical difficulty and stimulus variability effects may have been due to the relative ease of discriminating target words from response alternatives. Under conditions in which the target is easily distinguished from the foils, the effects of variables such as lexical difficulty and stimulus variability may be obscured. Therefore, Experiment 1B was designed to examine the effects of lexical difficulty and talker variability in closed-set tests containing response alternatives that were systematically selected to be most confusable with the target items. Results similar to those of Experiment 1A would suggest that the failure to find effects of stimulus variability and lexical difficulty was not due to the nature of the response foils but to changes in task demands with closed-set test formats.

Experiment 1B

Method

Subjects

Implant patients were not available for testing in Experiment 1B. Therefore, only the three normal-hearing subject groups (NH(Q), NH(+5), and NH(-5)) were examined. Each group consisted of 15 listeners with pure-tone air-conduction thresholds less than 20 dB HL for octave frequencies between 250 and 8000 Hz. All participants were native speakers of English and reported no history of hearing loss or other auditory dysfunction at the time of testing.

Selection of response alternatives

As noted, the primary goal of Experiment 1B was to determine if the effects of lexical difficulty and stimulus variability could be obtained in a closed-set format when response alternatives were systematically selected to maximize confusability with the target word. To determine the five most confusable foils for each item of the MRT, phoneme confusion matrices derived by Luce (1986) were examined. Luce (1986) measured consonant and vowel confusions for each position (initial consonant, medial vowel, final consonant) in consonant-vowel-consonant (CVC) stimuli at a number of signal-to-noise ratios. To determine the relative confusability of a specific response alternative on the MRT, the probability of misidentifying each of the individual phonemes in the target with the corresponding phoneme in the alternative was determined. The product of these individual probabilities was then multiplied by a log transform of word frequency to obtain a measure of the overall confusability of the alternative.

For example, to calculate the probability of confusing the target word CAT with the response alternative KIT, the separate probabilities of /k/|/k/ (i.e., the probability of saying /k/ when /k/ was presented as the initial phoneme), /ɪ/|/æ/ (the probability of saying /ɪ/ when /æ/ was presented as the middle phoneme) and /t/|/t/ (saying /t/ given /t/ in the final phoneme position) were obtained from the confusion matrices. The product of these individual probabilities was then multiplied by a measure of word frequency to give a combined index of overall confusability with the target item.

In Experiment 1B, an on-line lexical database was first used to identify all of the words that could be created from a given target item on the MRT by adding, deleting or substituting a single phoneme. This provided the set of possible alternatives for that target stimulus. The phoneme confusion matrices were then used to determine which five of these alternatives were most confusable with the target and these items were selected as the response foils for that item. Thus, each of the 100 MRT items tested in the closed-set format were presented with the target word and the five most confusable foils as response alternatives.

Stimulus materials and procedure

The stimuli and procedures were identical to Experiment 1A with the following exceptions. First, only three groups of listeners (NH(Q), NH(+5), and NH(-5)) were tested. Second, only the closed-set format was used. Third, the response foils provided with the MRT were replaced by the five most confusable items for that target (as identified by the procedure for selecting response alternatives described above). As in the closed-set tests of Experiment 1A, two 50-item blocks (one single talker, one multiple-talker) were presented. Within each block, half of the items were lexically easy and half were lexically hard.

Results and Discussion

Figure 3 displays results for both talker variability (top panel) and lexical difficulty (bottom panel). A mixed-design ANOVA with lexical difficulty and talker variability as repeated-measures variables and subject group as a between-subjects factor revealed only a significant main effect of subject group ($F[2,42] = 357.2$; $p < .001$). To examine the effects of using the most confusable response alternatives, the closed-set data from Experiments 1A and 1B were combined and analyzed together. Experiment (1A or 1B) was treated as a between-subjects factor and lexical difficulty and talker variability served as repeated-measures variables. The only statistically reliable finding revealed in the analysis was a main effect of experiment ($F[1,74] = 4.2$; $p < .05$); as expected, overall performance was poorer in Experiment 1B than in Experiment 1A. Thus, although increasing the difficulty of the closed-set test, by systematically selecting the most confusable response alternatives, reduced identification performance, it did not increase sensitivity to the specific effects of talker variability or lexical difficulty. This finding suggests that even difficult closed-set tests may fail to engage perceptual operations that are used in recognizing spoken words under more natural (open-set) listening conditions. The implications of these findings are discussed in more detail below.

Insert Figure 3 about here

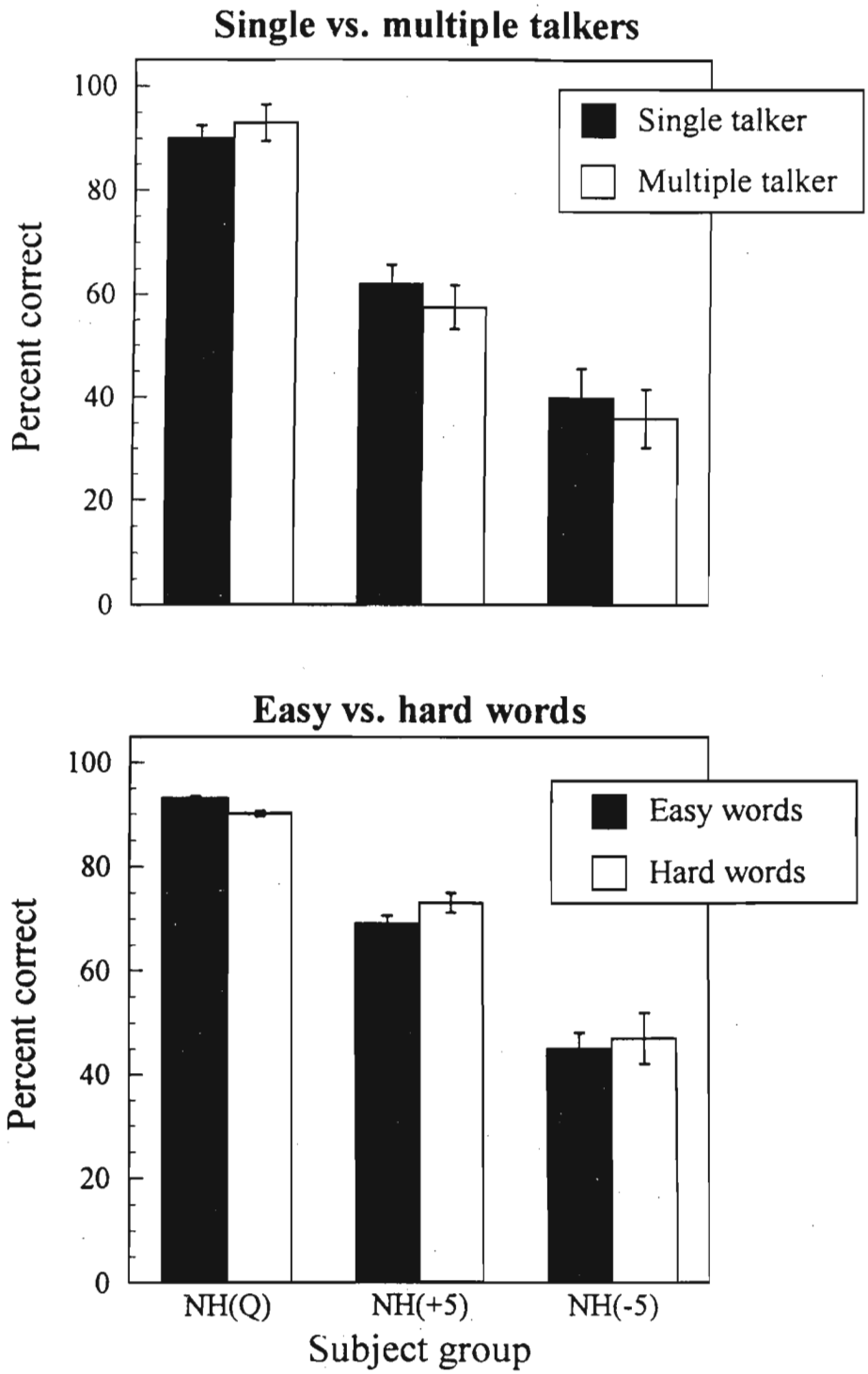


Figure 3: Effects of talker variability (top) and lexical difficulty (bottom) as measured in a closed-set test designed to maximize confusability between response alternatives and the target item. The subject groups are the same as in Figure 1 except the CI patients were not tested.

General Discussion

Taken together, the results of Experiments 1A and 1B suggest that with open-set response formats, normal-hearing, noise-masked normal and cochlear implant listeners all exhibit reduced spoken word recognition as a function of increased stimulus variability and greater lexical difficulty. In contrast, these same variables produced no effects on recognition performance when a closed-set response format was used. One implication of the parallel findings across the different subject groups is that, despite considerable differences in absolute sensitivity, these listeners engaged qualitatively similar mechanisms in recognizing spoken words. For example, the significant reduction in open-set identification performance for lexically hard words suggests a similar structural organization of the mental lexicon and comparable use of this organization across subject groups. Normal-hearing, cochlear implant, and noise-masked normal hearing listeners appear to organize words into lexical neighborhoods based on acoustic-phonetic similarity and their word recognition performance is affected by differences in the number (neighborhood density) and frequency (neighborhood frequency) of the phonetically similar items in those neighborhoods.

The findings from this study are potentially quite important for clinical evaluation of speech perception abilities because they suggest that assessment instruments focusing exclusively on listeners' ability to extract phoneme information are necessary but not sufficient for predicting spoken word recognition performance. If speech perception required only the sequential identification of individual phonemes, then word identification and phoneme recognition scores in CI patients should be highly correlated. However, the results of several recent studies provide evidence against this hypothesis (Kirk et al., 1995; Koch et al., 1996, although see Rabinowitz, Eddington, Delhorne, & Cuneo, 1992). Kirk et al. (1995), for example, reported that phoneme recognition scores were not highly predictive of pediatric cochlear implant patients' word identification performance on lexically easy and hard stimuli presented in open-set formats. One factor that may have contributed to the low correlation between phoneme and word recognition scores is that, in addition to extracting phonetic information from the speech signal, spoken word recognition requires listeners to isolate (i.e., discriminate) individual words from phonetically similar sound patterns stored in long-term lexical memory. Thus, any comprehensive evaluation of spoken word recognition abilities in normal or clinical populations could benefit from incorporating tests designed to measure both phonetic and lexical discrimination as a means of assessing identification performance at several levels of lexical difficulty.

A second implication of the present findings is that measures of speech perception obtained with tests that minimize stimulus variability, by using highly articulated stimuli produced by a single talker, may fail to generalize to more natural listening situations where many different factors combine to produce extensive acoustic-phonetic variability. The significant reduction in identification scores that was observed following a change from single- to multiple-talker word lists indicates that the ability to adjust or normalize for acoustic-phonetic variations due to changes in talker characteristics is an integral aspect of the speech perception system even for CI patients who receive highly impoverished acoustic information. The ability to predict "real-world" speech perception abilities will therefore require the development of new assessment procedures that can evaluate listeners' ability to rapidly accommodate changes in vocal-tract characteristics. The current findings indicate that closed-set formats are likely to be inadequate for this purpose because they fundamentally change the task demands imposed on listeners during spoken word recognition.

One reason that closed-set formats may not be sensitive to the effects of either stimulus variability or lexical difficulty is that providing listeners with a set of response alternatives alters the perceptual strategies used to recognize spoken words. For example, in open-set speech discrimination tests listeners

must derive a best match between a phonetic representation obtained from the incoming speech signal and patterns stored in long-term lexical memory. Therefore, the organization of words within the mental lexicon can have a significant influence on the dynamics of the recognition process. In closed-set formats, however, listeners can effectively eliminate the need to access or consult long-term lexical memory by limiting their search to the response alternatives provided on a particular trial. That is, in closed-set formats the set of potential response candidates is no longer determined by the structure of the mental lexicon. Instead, listeners can restrict their lexical search to the response alternatives that accompany each stimulus. Under such conditions, words are no longer recognized in the context of other phonetically-similar words in memory and the processes of word recognition and lexical discrimination are changed substantially. The differential nature of word recognition under the two response formats may, in part, explain the absence of lexical difficulty and stimulus variability effects in closed-set tests.

Although the preceding account of the influence of response format on spoken word recognition must be considered preliminary until additional evidence is obtained, it nevertheless raises the important theoretical issue that demand characteristics of an assessment instrument may alter the processes used in spoken word recognition. One difference between open- and closed-set response formats that may partially account for the differential results obtained with talker variability and lexical difficulty is that closed-set tests may fail to mimic the cognitive demands associated with rapid, on-line recognition of spoken words. Consistent with this explanation, Mullennix et al. (1989) reported that reducing cognitive demands, by increasing signal-to-noise ratios, reduced the effects of talker variability. In fact, at artificially high S/N ratios, Mullennix et al. failed to observe any effects of talker variability. Considered with the earlier suggestion that providing response alternatives fundamentally alters the normal processes of lexical search and access, the present findings with closed-set formats may be attributable to a combination of reduced task demands and altered perceptual strategies that result from providing listeners with a set of response alternatives.

It could be argued that similar reductions in cognitive demands are achieved under natural listening conditions by using semantic context. That is, listeners may be able to limit the number of possible lexical alternatives that they consider by using semantic information. Although increasing semantic predictability has been shown to raise overall identification scores (Nittrouer & Boothroyd, 1990), recent evidence (Karl & Pisoni, 1994) indicates that the effects of talker variability are observed even with highly constrained semantic contexts. Karl and Pisoni (1994) had listeners transcribe Harvard sentences (Egan, 1948) that were produced in either single or multiple talker contexts. Their results indicated that multiple-talker transcription performance was significantly poorer than for the single-talker condition. Thus, although semantic context can reduce overall task difficulty, the change in cognitive demands resulting from the addition of semantic information is qualitatively different from that produced by providing listeners with response alternatives. This proposal indicates that not all sources of context are equivalent and we believe it is important to distinguish how different kinds of contextual information affect listeners' performance in a variety of speech perception tasks.

In summary, traditional instruments for assessing speech perception that rely on closed-set formats and single-talker productions can provide invaluable information about the perceptual capacities of clinical populations. These protocols are important for evaluating a number of individual abilities, such as identification and discrimination of phonetic features, that may be necessary for accurate speech perception. In addition, closed-set tests have been useful for examining speech processing strategies in listeners with sensory aids. However, the present findings comparing open- and closed-set tests of spoken word recognition suggest that closed-set formats may not be effective at assessing other operations, such as perceptual normalization of talker differences and isolating words in long-term memory, that are also

critical for spoken word recognition in natural listening environments. The results of the present study therefore represent an initial step in systematically evaluating and understanding the limitations of individual assessment instruments used to measure speech intelligibility in different populations. The goal of this research should be to develop comprehensive test batteries that include a variety of instruments designed to evaluate the broad spectrum of abilities necessary for understanding spoken language under a variety of listening conditions.

Speech perception is an extremely robust process that can quickly adapt to changing listening conditions. To understand the perceptual and neural mechanisms that are responsible for these abilities, we will need to develop a new generation of theoretically motivated tests that assess spoken word recognition across a range of task requirements and listening populations. The results of the present study demonstrate the potential value of this research strategy for gaining a more detailed understanding of speech perception and spoken-language processing.

References

Blamey, P.J., Dowell, R.C., Brown, A.M., Clark, G.M., & Seligman, P.M. (1987). Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors. *Journal of the Acoustical Society of America*, **82**, 48-57.

Cluff, M.S., & Luce, P A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 551-563.

Dorman, M.F. (1993). Speech perception by adults. In R. S. Tyler (Ed.), *Cochlear implants: Audiological foundations* (pp. 145-190). San Diego: Singular Publishing.

Dowell, R.C., Brown, A. M. & Mecklenburg, D. (1990). Clinical assessment of implanted deaf adults. In G. Clark, Y. Tong, & J. Patrick (Eds.), *Cochlear prostheses* (pp. 193-206), Edinburgh: Chuchill Livingstone.

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

Elliott, L.L., Clifton, L.A., & Servi, D.G. (1983). Word frequency effects for a closed-set word identification task. *Audiology*, **22**, 229-240.

Garstecki, D.C. & Wilkin, M.K. (1976). Linguistic background and test material considerations in assessing sentence identification ability in English- and Spanish-English-speaking adolescents. *Journal of the American Audiological Society*, **1**, 263-268.

Geers, A. E. & Brenner, C. (1994). Speech perception results: Audition and lipreading enhancement. In A. E. Geers & J. S. Moog (Eds.), *The Volta Review Vol. 96*, Effectiveness of cochlear implants and tactile aids for deaf children: The sensory aids study at Central Institute for the Deaf (pp. 97-108), Washington DC: A.G. Bell Assoc. for the Deaf.

House, A.S., Williams, C.E., Hecker, M.H.L. & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, **88**, 642-654.

Joos, M.A. (1948). Acoustic phonetics. *Language*, **24**, Suppl. 2, 1-136.

Karl, J. & Pisoni, D.B. (1994). The role of talker-specific information in memory for spoken sentences. *Journal of the Acoustical Society of America*, **95**, 2873.

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.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169-226). Cambridge, MA: MIT Press.

Koch, D.B., Carrell, T.D., Tremblay, K., & Kraus, N. (1996). Perception of synthetic syllables by cochlear-implant users: Relation to other measures of speech perception. Poster presented at the mid-winter meeting of the Association for Research in Otolaryngology, St. Petersburg Beach, FL.

Kucera, F. & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 122-147). Cambridge, MA: MIT Press.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71-102.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 676-684.

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

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

Nearey, T.M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088-2113.

Nittrouer, S., & Boothroyd, A. (1990). Context effects in phoneme and word recognition by young children and older adults. *Journal of the Acoustical Society of America*, **87**, 2705-2715.

Owens, E., Kessler, D. & Schubert, E. (1981). The minimal auditory capabilities (MAC) battery. *Hearing Aid Journal*, **34**, 9-34.

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.

Pisoni, D.B. (1985. Speech perception: Some new directions in research and theory. *Journal of the Acoustical Society of America*, **78**, 381-388.

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.

Rabinowitz, W.M., Eddington, D.K., Delhorne, L.A., & Cuneo, P.A. (1992). Relations among different measures of speech reception in subjects using a cochlear implant. *Journal of the Acoustical Society of America*, **92**, 1869-1881.

Rosen, S., Fourcin, A. J., Abberton, E., Walliker, J.R., Howard, D M., Moore, B.C.J., Douek, E.E., & Frampton, S. (1985). Assessing Assessment. In R.A. Schindler & M.M. Merzenich (Eds.), *Cochlear implants* (pp. 479-498). New York: Raven Press.

Shipman, D.W. & Zue, V.W. (1982). Properties of large lexicons: Implications for advanced isolated word recognition systems. Paper presented at the IEEE Conference on Acoustics of Speech and Signal Processing, Paris.

Sommers, M.S. (in press). The structural organization of the mental lexicon and its contribution to age-related deficits in spoken word recognition. *Psychology & Aging*.

Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-24.

Tyler, R.S., Lowder, M.W., Otto, S.R., Preece, J.P., Gantz, B.J., & McCabe, B.F. (1984). Initial Iowa results with the multichannel cochlear implant from Melbourne. *Journal of Speech & Hearing Research*, **27**, 596-604.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Training Japanese Listeners to Identify English /r/ and /l/ IV: Some Effects of Perceptual Learning on Speech Production[1]

Ann R. Bradlow, David B. Pisoni,[2] Reiko Akahane-Yamada[3] and Yoh'ichi Tohkura[4]

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

1 This work was supported by NIH-NIDCD Training Grant DC-00012 and NIH-NIDCD Research Grant DC-00111 to Indiana University. The authors are grateful to Luis Hernandez and Takahiro Adachi for technical support, and to Fernando Vanegas for subject running.
2 Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.
3 ATR Human Information Processing Research Laboratories, Kyoto 619-02, Japan
4 ATR Human Information Processing Research Laboratories, Kyoto 619-02, Japan

Abstract

This study investigated the effects of training in /r/-/l/ perceptual identification on /r/-/l/ production by adult Japanese speakers. Subjects were recorded producing English words that contrast /r/ and /l/ before and after participating in an extended period of /r/-/l/ identification training. All subjects showed significant perceptual learning as a result of the training program, and this perceptual learning generalized to novel items spoken by new talkers. Improvement in the Japanese trainees' /r/-/l/ productions as a consequence of training in perception was evaluated by two separate tests with native English listeners. First, a direct comparison of the pretest and post-test tokens showed significant improvement in /r/ and /l/ production as a consequence of perceptual learning. Second, the post-test productions were more accurately identified by English listeners than the pretest productions in a two-alternative minimal-pair identification procedure. These results indicate that the knowledge gained during perceptual learning of /r/ and /l/ transferred to the production domain, implying a close link between perception and production.

Training Japanese Listeners to Identify English /r/ and /l/: IV. Some Effects of Perceptual Learning on Speech Production

Introduction

The acquisition of novel phonetic categories by non-native speakers has been a long-standing issue in speech science and experimental phonetics. Previous research has shown that foreign accents persist even for highly proficient speakers of a non-native language (e.g., Tahta, Wood, and Loewenthal, 1981; Flege and Hillenbrand, 1987), and that non-native speakers have extreme difficulty with both the perception and production of certain non-native phonetic contrasts (e.g., Flege, 1988; Goto, 1971). These findings raised the possibility that adult foreign language-learners are incapable of modifying their existing phonetic systems to accommodate all possible non-native contrasts. For instance, summarizing work on the acquisition of foreign language perceptual contrasts by adults, Strange and Dittmann (1984) state,

> ...although intensive conversational instruction (with native speakers) is correlated with improved perception of the foreign contrasts, perception is still not native-like, even for the most advanced students. Thus, for adults learning a foreign language, modification of phonetic perception appears to be slow and effortful, and it is characterized by considerable variability among individuals. (page 132)

Similarly, in an attempt to develop a training technique that would successfully promote the acquisition of the English /θ/-/ð/ contrast by francophone speakers, Jamieson and Morosan (1986) reviewed a number of studies that failed to train non-native contrasts. Summarizing their review, they state,

> These difficulties with non-native speech contrasts may indicate that certain distinctions are extremely difficult for adults to learn, or even that adults cannot learn to make certain distinctions in a linguistically meaningful manner. (page 206)

In response to these pessimistic views of adult acquisition of non-native perceptual contrasts, several studies have been carried out to identify appropriate perceptual training techniques. Strange and Dittmann (1984) employed a psychophysical training task and attempted to modify the perception of word-initial /r/ and /l/ in adult Japanese learners of English. They used a fixed-standard AX discrimination procedure with a synthetic "rock-lock" 10-point continuum. Although the subjects in this study showed more categorical perception on the synthetic /r/-/l/ series as a consequence of training, this perceptual learning did not generalize to the /r/-/l/ contrast in a variety of phonetic environments using natural speech produced by a native American English speaker. They concluded that the perceptual reorganization required for non-native phonetic contrast learning "...requires intensive instruction and considerable time and effort," and that "we do not yet know which methods will best produce generalization of training" (page 142).

Jamieson and Morosan (1986) used a perceptual fading technique to train francophone speakers on the English /θ/-/ð/ contrast. In this procedure, subjects performed an identification task with feedback. The training stimuli were synthetic CV tokens that varied in duration of the initial voiced or voiceless fricative. During training, the stimulus set increased in variability as the training progressed. At the start of the training program, only the tokens with the longest fricative durations were presented for identification. Gradually, the intermediate stimuli were introduced, until finally the entire set of stimuli was presented for identification. The results of this study showed improved identification and intercategory discrimination of

both synthetic and naturally produced /θ/ and /ð/ tokens as a consequence of the training study. However, in a follow-up study, Morosan and Jamieson (1989) found that, although this perceptual learning generalized across male and female voices, it did not generalize to /θ/ and /ð/ in word-medial and word-final position, nor did it generalize to the /ð/-/d/ contrast.

In an attempt to develop a perceptual learning procedure that would facilitate the acquisition of robust, highly-generalized non-native speech contrasts, Logan, Lively, and Pisoni (1991) introduced stimulus variability into their training stimuli by using tokens from several talkers producing the contrast in different phonetic environments. It is well known that the English /r/-/l/ contrast is a particularly difficult phonetic contrast for native speakers of Japanese (Miyawaki, Strange, Verbrugge, Liberman, Jenkins and Fujimura, 1975; Mochizuki, 1981; MacKain, Best, and Strange, 1981; Yamada and Tohkura, 1992), and that this contrast presents difficulties to Japanese speakers in both perception and production (Goto, 1971; Sheldon and Strange, 1982). Thus, the goal of this study was to train native speakers of Japanese to identify American English /r/ and /l/ by developing robust, highly generalizable phonetic categories. The key features of Logan et al.'s high-variability training procedure were the training task and the training stimuli. Following Jamieson and Morosan (1986), they used a two-alternative forced-choice identification task with feedback rather than the AX fixed-standard discrimination task used by Strange and Dittmann (1984). This identification task encourages classification of different stimuli into categories rather than promoting discrimination of fine within-category differences. The stimuli used in this training procedure were naturally produced English words that contrasted /r/ and /l/ in five phonetic environments as produced by five different talkers. The principle motivating the development of this stimulus set was that, in order to develop robust phonetic categories, the trainees should be exposed to the full range of stimulus variability that they can expect to encounter in natural English /r/ and /l/ categories.

In contrast to the finding reported by Strange and Dittmann, the result of this study showed consistent improvement in /r/-/l/ minimal pair identification, as well as a reliable decrease in response times from pretest to post-test. Thus, this study provided the first demonstration that, given exposure to a wide range of stimuli, adult Japanese speakers can learn to identify naturally produced English /r/ and /l/ tokens in multiple phonetic environments.

In a second study using this training procedure, Lively, Logan, and Pisoni (1993) directly investigated the role of variability in phonetic environment and talker in the acquisition of the English /r/-/l/ contrast by Japanese speakers. In one experiment, they modified the training stimulus set to include only tokens that contrasted /r/ and /l/ in the three most difficult phonetic environments: they omitted tokens with /r/ and /l/ in word-final and final cluster positions (Atkinson, 1972). They hypothesized that training on only the three most difficult environments would still provide sufficient variability in phonetic context to promote the acquisition of robust phonetic categories. The results of this experiment indicated that, even with the reduced set of phonetic contexts in the training stimuli, subjects improved in /r/-/l/ identification from pretest to post-test, and this improved level of performance generalized to new tokens and a new talker. Furthermore, response times decreased from pretest to post-test across all environments, even for the word-final environment, which was not included in the original training stimulus set.

In a second experiment, Lively et al. investigated the effect of multiple talkers by comparing the effects of training with a single talker to training with multiple talkers. Subjects trained with stimuli from only one talker showed improvement in /r/-/l/ identification from pretest to post-test for some, but not all, phonetic environments, and this learning showed only minimal generalization to a new talker and to new tokens. This result suggested that training with stimuli produced by multiple talkers is crucial to the development of robust, highly generalizable non-native phonetic categories. The role of talker variability in

non-native phoneme training was further investigated by Magnuson et al. (1995). In this study, five groups of subjects were trained with stimuli from five different talkers: each group was trained with stimuli from a single talker. Only two of the five groups showed robust perceptual learning that generalized to novel tokens and talkers; the remaining three groups failed to show any significant perceptual learning from pretest to post-test.

Taken together, the results of these investigations into the role of variability in talker and phonetic context in non-native phoneme training lead to two main conclusions. First, it is possible to achieve robust non-native phoneme acquisition by training with stimuli from a limited set of the most difficult phonetic environments. However, based on the more limited success of training techniques that use training stimuli with only one environment (e.g., Strange and Dittmann, 1984; Morosan and Jamieson, 1989), it is still most likely that some degree of variability in phonetic environment is crucial to the acquisition of robust novel phonetic categories. Second, training with certain talkers in a single-talker training paradigm may lead to robust novel category acquisition; whereas, training with other talkers fails to promote such acquisition. In contrast, training with multiple talkers consistently leads to robust phonetic category acquisition.

In another investigation using this "high-variability" training technique, Lively, Pisoni, Yamada, Tohkura, and Yamada (1994) found that the perceptual learning was retained in long-term memory for a period of up to 6-months after training. This study also provided the first replication of the original Logan et al. (1991) result with a group of monolingual Japanese speakers who had never lived in an English-speaking country. In a follow-up study of the "high-variability" training program, Yamada (1993) investigated the effect of extended training on /r/ and /l/ identification by native speakers of Japanese. This study continued training and testing beyond the 15 sessions used in the previous studies, and showed that subjects' performance continued to improve until 45 sessions, at which point the learning curve leveled off at about 85% correct. Taken together, this series of studies showed that, given appropriate perceptual training, the adult nervous system does indeed show the "plasticity" necessary for the perceptual identification of new, very difficult phonetic contrasts.

The present project builds on this earlier research by investigating the effect of the acquisition of a new perceptual contrast on the production of that contrast. We wanted to know if a listener's control over speech production would be affected by learning a new phonetic contrast. By directly examining the effect of perceptual learning on speech production, this project also addresses the relationship between speech perception and production. Any transfer of learning in perception to the production domain would provide new evidence for a direct perception-production link; and, would therefore suggest that perceptual identification training can facilitate the acquisition of non-native production categories.

Previous studies that have investigated the relationship between perception and production of a non-native phonetic contrast, have generally focused on the subjects' performance in perception and production at a single point in time. For example, studies by Goto (1971) and by Sheldon and Strange (1982) showed that some Japanese subjects were able to produce identifiable /r/ and /l/ tokens even though they were unable to reliably identify native English /r/ and /l/ tokens. This finding lead these researchers to conclude that production can precede perception in the acquisition of a non-native contrast. A similar result was reported by Yamada et al. (1994), who investigated the intelligibility of /r, l, and w/ productions by a large number of Japanese speakers with varying degrees of exposure to English. They found a positive correlation between performance in an /r, l, w/ identification task and intelligibility of /r, l, w/ productions by the Japanese subjects as judged by American English listeners. Furthermore, the results showed that for low performers in the perceptual identification task, the intelligibility scores for production varied from low

to high; whereas for higher performers in the perception task, the intelligibility scores tended to be high. This indicates that for some of the subjects, production abilities exceeded perception abilities, but not vice versa.

These studies are informative about the relationship between perception and production in adult second-language learners; however, they do not provide direct information about how the changes in one domain (i.e. perception) affect performance in the other domain (i.e. production). Whereas the information provided by the studies discussed above is correlational, the major goal of the present study was to investigate the possibility of a perception-production link to the extent that success in a task of perceptual learning leads to an improvement in speech production by adult second-language learners.

Two recent studies provide some indication that transfer of perceptual learning to speech production can occur. Rochet (1995) reported that after perceptual identification training with a synthetic French /bu/-/pu/ continuum, Mandarin speakers displayed more French-like VOT perceptual categorization. Furthermore, production data from the Mandarin subjects showed a change in VOT durations in the direction of native French VOT durations. In recent studies with children with articulation disorders, Jamieson and Rvachew (Jamieson and Rvachew, 1992; Rvachew, 1994; Jamieson and Rvachew, 1994) found that speech perception training can facilitate sound production learning in children who exhibit both perception and production deficits. For instance, Rvachew (1992) examined the role of perception training, in conjunction with traditional speech production therapy, in the remediation of production errors for 27 phonologically impaired preschoolers who misarticulated /ʃ/. Subjects were randomly assigned to three different perception training groups: Group 1 listened to various correct and incorrect productions of the word "shoe," Group 2 listened to correct productions of "shoe" and "moo," Group 3 listened to the words "Pete" and "cat." At the end of six weeks of training, subjects in Groups 1 and 2 both showed greater improvement in /ʃ/ production than Group 3, indicating that perception training can enhance the effectiveness of speech production therapy. In the present study, we examine this perception-production link further by investigating the effects of perceptual learning on production of the /r/-/l/ contrast by monolingual Japanese in the absence of any explicit production training, and across a wide range of phonetic contexts.

The general design of the study had four phases: a pretest phase, a perceptual training phase, a post-test phase, and a production assessment phase. During the pretest phase, both perception and production data were collected from a group of monolingual Japanese subjects. In the perceptual training phase, the subjects were trained to identify English /r/ and /l/ minimal pairs using the high-variability training technique developed in earlier work (Logan et al., 1991; Lively et al, 1993, 1994; Yamada, 1993). In the post-test phase, both perception and production data were once again collected from the Japanese listeners. Finally, during the production assessment phase, the pre- and post-test utterances were evaluated by a group of native American English speakers. Thus, in this study, we investigated the effect of perceptual learning on subsequent performance in both perception and production. If we find transfer of training from perception of /r/ and /l/ to control over production of this contrast, the results would have a number of theoretical implications for issues surrounding the neural representation of speech and the processing units employed in speech perception and production.

Perceptual Learning

Method

Subjects

The subjects were 11 monolingual, native speakers of Japanese (5 females and 6 males), ranging in age from 19-22 years. None had lived abroad or had special training in English conversation. However, as is typical in Japan, all of the subjects had studied English since junior high school (from about age 12). The subjects were recruited from Doshisha University, Kyoto prefecture, Japan. A comparable group of 12 Japanese speakers (6 females and 6 males) served as control subjects. These control subjects were drawn from the same population as the experimental subjects. None of the subjects reported any history of speech or hearing impairment at the time of testing. A hearing screening performed at 15 dB HL for the frequencies 250 through 8000 Hz showed all subjects to have normal bilateral acuity.

Procedure

The perceptual training program followed the procedures first developed by Logan et al. (1991), and later extended by Yamada (1993). This procedure consisted of a pretest phase, a training phase, and post-test phase. The pretest phase consisted of a minimal pair identification task with naturally produced English /r/-/l/ minimal word pairs produced by a native speaker of General American English. The perceptual learning phase involved 45 sessions (over a period of 3-4 weeks) of perceptual identification with feedback. The training stimuli consisted of a wide range of naturally spoken /r/-/l/ minimal word pairs produced by five native speakers of General American English. Finally, the post-test phase included a perceptual identification post-test (identical to the pretest), and two tests of generalization. The tests of perceptual generalization consisted of a minimal word pair identification task with novel words spoken by a new speaker (Test of Generalization 1), and with novel words produced by one of the speakers used in creating the training stimuli (Test of Generalization 2). Control subjects performed the pretest, post-test, and two tests of generalization; however, these subjects did not go through the perceptual identification training program. For the control group, the time lag between the pretest and post-test phase was equal to the time it took for the trained subjects to go through the entire 45-session training program (i.e., 3-4 weeks).

All perception training and testing was carried out at ATR Human Information Processing Research Laboratories in Kyoto, Japan. For all four perception tests, (pretest, post-test, two tests of generalization) the same two-alternative forced choice minimal word pair identification procedure was used. Subjects were tested individually in a sound-treated room where they sat in a cubicle equipped with headphones (STAX-SR-Lambda Signature) and a NeXT workstation. Each trial began with a 500 ms presentation on the computer monitor of the standard English orthographies for an /r/-/l/ minimal pair. One member of the minimal pair appeared in the lower left corner of the screen; the other appeared in the lower right corner. The spoken test word was then presented at a comfortable listening level through the subjects' headphones. Subjects had 10 seconds to respond by pressing "1" to identify the spoken word as the orthographic word on the left of the screen, or "2" for the word on the right of the screen. For half the trials, a response of "1" corresponded to an /r/ identification label and a response of "2" corresponded to an /l/ identification label; for the other half of the trials the order of identification labels was reversed. During the training trials, feedback was given in the form of a chime signaling a correct response and a buzzer signaling an incorrect response. After the buzzer for an incorrect response, the test word was repeated. As an additional motivation, each correct response received a 1 yen (approximately 1 cent) reward over and

above the regular subject payment. There was no feedback in the pretest, post-test, or tests of generalization. Feedback was provided only during the perceptual learning phase.

Stimuli

A large digital database of spoken words for the perception tests was originally recorded and compiled in the Speech Research Laboratory at Indiana University (see Logan et al., 1991 for additional details). All stimuli were recorded in an IAC sound-attenuated booth. The utterances were low-pass filtered at 4.8 kHz and digitized at 10 kHz using a 12-bit analog-to-digital converter. The waveform files were then equated for rms amplitude using software developed in the Speech Research Laboratory. The files were then digitally transferred to ATR Human Information Processing Research Laboratories where they were upsampled to 22.05 kHz and rescaled to 16-bit resolution for presentation on the NeXT workstations.

The pretest and post-test stimuli were the same words as those used by Strange and Dittmann (1984). This set of stimuli consisted of 16 minimal pairs that contrasted /r/ and /l/ in 4 phonetic environments: initial singleton, initial cluster, intervocalic, and final cluster. There were also 4 minimal pairs that contrasted other English phonemes. These stimuli were recorded by a male speaker of General American English. The training stimuli consisted of 68 minimal pairs that contrasted /r/ and /l/ in five phonetic environments: 12 initial singleton pairs, 25 initial cluster pairs, 5 intervocalic pairs, 15 final singleton pairs, and 11 final cluster pairs. These stimuli were recorded by 5 speakers of General American English (3 males and 2 females). The stimuli for the first test of generalization consisted of 96 words with /r/ or /l/ in 5 phonetic environments spoken by a new talker. The stimuli for the second test of generalization consisted of 99 /r/-/l/ words (5 environments) spoken by one of the talkers in the training set. All lists of words are provided in the Appendix.

Results of Perceptual Learning

Figure 1 shows the results of perceptual identification training for the experimental (left panel) and the control (right panel) groups. This figure displays the percentage of correct identifications for all four of the perceptual tests: the pretest, post-test, and the two tests of generalization. As shown in the left panel, the experimental (trained) group of subjects showed improved identification scores from pretest to post-test, and this improved level of performance was maintained for the two tests of generalization. A one-factor ANOVA with test as the factor (Pre, Post, Gen1, Gen2) and the trained group's accuracy scores as the dependent variable showed a significant main effect of test ($F(3,40)=4.077$, $p=0.013$). Post-hoc pairwise comparisons (Tukey HSD) showed a significant difference between Pretest and Post-test ($p=0.036$), and between Pretest and Gen1 ($p=0.019$). The difference between Pretest and Gen2 approached significance ($p=0.065$). None of the pairwise differences between the post-test and the two tests of generalization were significant. This pattern of results may be contrasted with the performance of the control group of subjects who did not go through the training program. Their data, displayed in the right hand panel, showed no difference in performance across all four tests. A one-factor ANOVA revealed no main effect of test on accuracy score ($F(3,44)=0.962$), $p=0.419$), and none of the pairwise comparisons showed significant differences. These results replicate the findings of the earlier studies that used the high-variability perceptual training program (Logan et al., 1991; Lively et al., 1993; Lively et al., 1994; Yamada, 1993), and show the effectiveness of the training program for facilitating the acquisition of the /r/-/l/ perceptual contrast by adult, monolingual Japanese speakers. After exposure to a large set of /r/-/l/ minimal pairs with the target phoneme in a variety of phonetic contexts as produced by multiple speakers, the experimental

subjects learned to identify this non-native contrast. Furthermore, they were able to generalize this perceptual learning to novel tokens produced by a new talker.

Insert Figure 1 about here

In order to gain some insight into the perceptual reorganization that resulted from the perceptual learning task, Figure 2 shows the distribution of /r/ and /l/ responses to /r/ and /l/ stimuli at the pretest (left panel) and post-test (right panel) phases for the experimental group of subjects. Table I gives the accuracy scores represented in this figure. In each panel, the five columns represent the stimuli with /r/ and /l/ in the four different phonetic environments that were included in the pre- and post-test stimulus sets, as well as averaged across all environments. The bottom half of each column represents the /r/ stimuli, and the top half of each column represents the /l/ stimuli. The middle (bold) portion of each column represents the proportion of stimuli that received an incorrect response label. A 3-factor repeated measures ANOVA was performed with test (pre or post) as the repeated measure, and phoneme (/r/ or /l/) and environment (four levels) as within-groups factors.

Insert Table I and Figure 2 about here

This analysis revealed three main findings. First, the main effect of test was highly significant ($F(1,80)=40.369$, $p=0.0001$), due to the overall improvement in identification accuracy from pretest to post-test. This is indicated in Figure 2 by a consistent decrease in the area of the bold portion of each column in the right panel relative to the area of the bold portion in the corresponding column in the left panel. Second, the main effect of phoneme was also significant ($F(1,80)=4.215$, $p=0.043$). This result demonstrates that /r/ was generally more accurately identified than /l/. The effect of phoneme is seen most clearly in Figure 2 by the asymmetrical distribution of incorrect responses (the bold portion) at pretest (left panel). Overall, as seen in the average column, there were fewer incorrect responses for the /r/ stimuli than for the /l/ stimuli at pretest. The third finding revealed by this analysis, is the significant interaction between test and phoneme ($F(1,80)=5.644$, $p=0.012$): /l/ responses showed more overall improvement from pretest to post-test than /r/ responses. This is shown in Figure 2 by the more symmetrical distribution of incorrect responses at post-test (right panel) than at pretest (left panel). This change from an asymmetrical distribution to a more symmetrical distribution suggests that, after training, subjects have established two separate perceptual categories which correspond more closely to the target English /r/ and /l/ categories.

There was also a main effect of phonetic environment ($F(3,80)=4.603$, $p=0.005$), indicating that the accuracy for the various phonetic environments decreased from final to initial singleton to medial to initial cluster positions. (This dependency on phonetic context replicates earlier findings, e.g., Gillette (1980), Mochizuki (1981), Sheldon and Strange (1982).) The interaction between phoneme and environment was significant ($F(3,80)=4.647$, $p<0.005$), due to the high identification accuracy of /l/ in initial cluster position relative to /r/ in that environment. Finally, the three-way interaction (test x phoneme x environment) ($F(3,80)=7.272$, $p=0.0002$) was also significant indicating that the perceptual learning of /r/ and /l/ was highly context-dependent.

It is clear from an examination of these data that the Japanese listeners displayed significant perceptual learning as a result of the extended, high-variability training program (Figure 1). Furthermore,

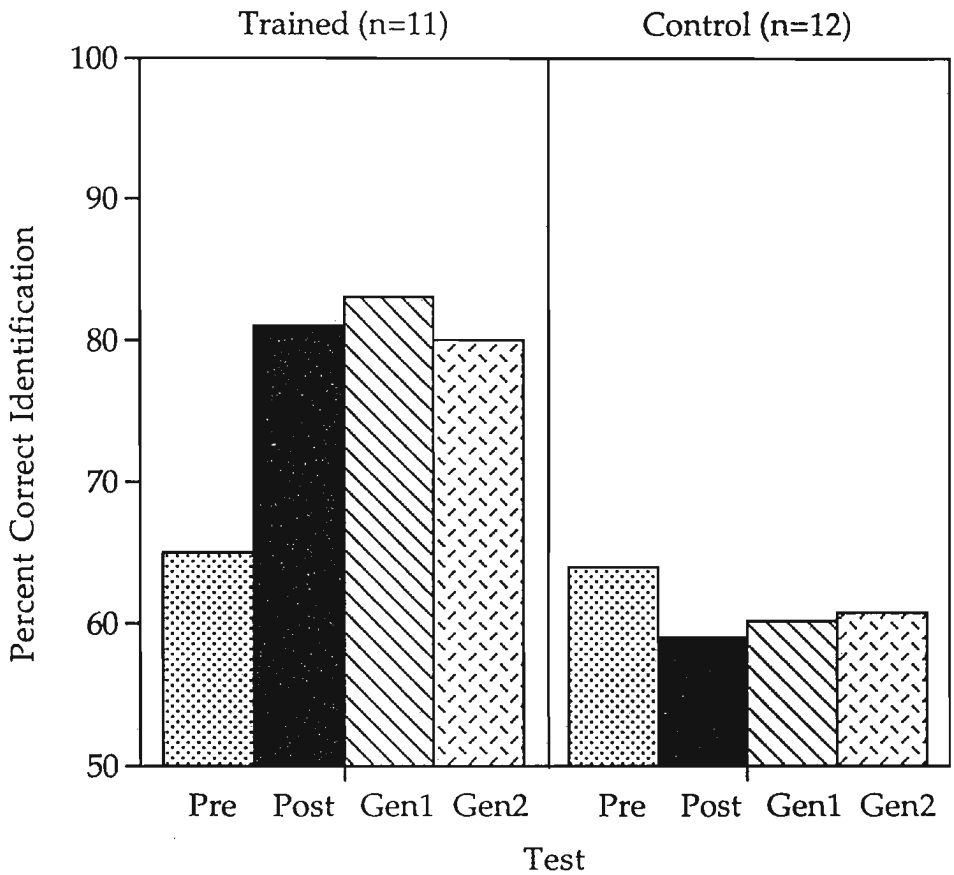


Figure 1. Percent correct perceptual identification performance for trained (left panel) and control (right panel) subjects at pretest, post-test, and the two tests of generalization.

Table I.

Japanese trainees' perceptual accuracy at pretest and post-test by phonetic environment.

	/r/		/l/	
Environment	Pretest	Post-test	Pretest	Post-test
initial	84.09	75.00	44.32	85.23
initial cluster	50.00	70.46	70.46	80.68
medial	62.50	84.09	55.68	70.46
final	90.91	98.86	62.50	86.36
Totals	71.88	82.10	58.24	80.68

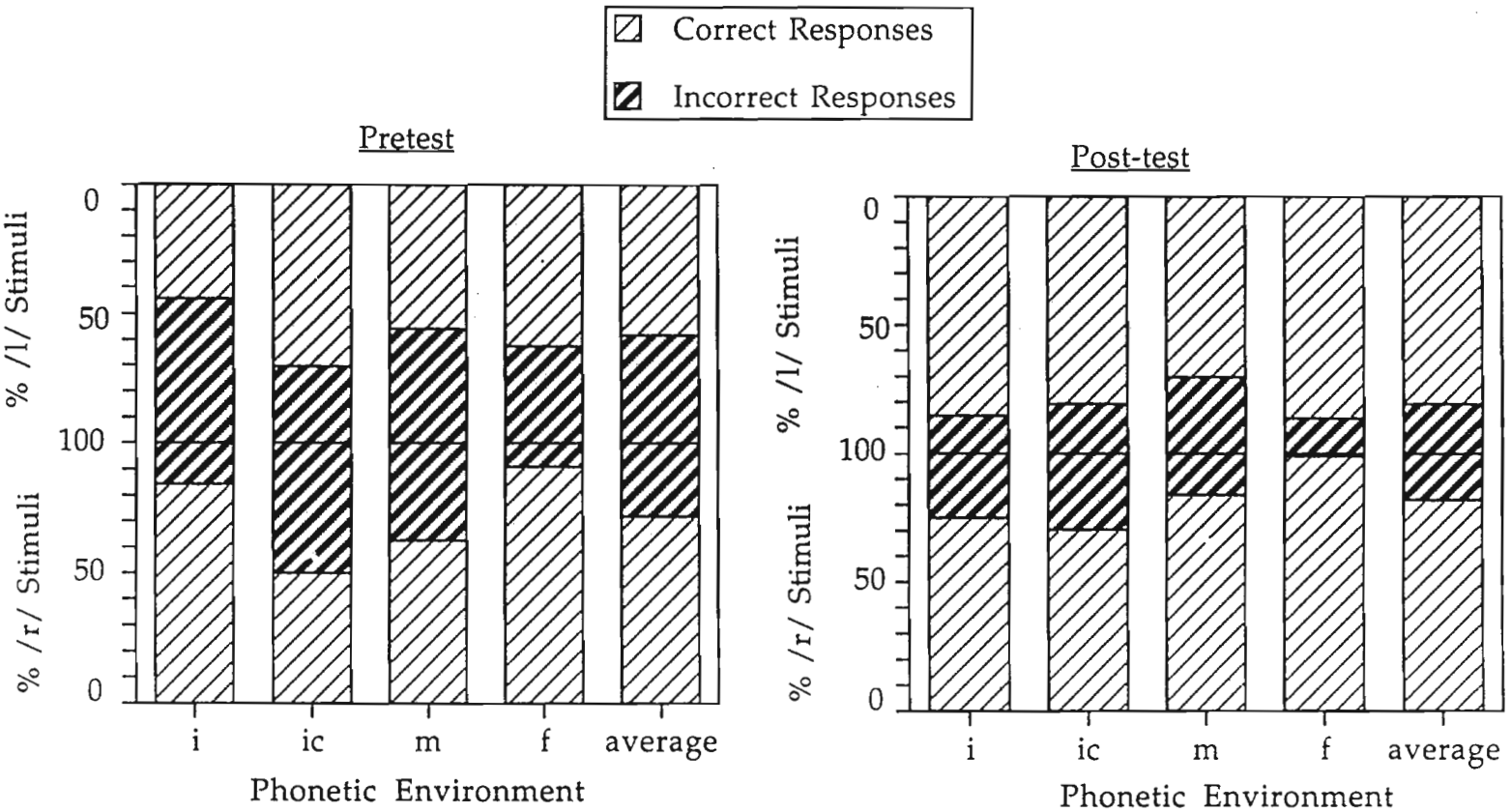


Figure 2. Distribution of the Japanese trainees' identification responses at pretest (left panel) and at post-test (right panel) by phonetic environment. The bottom half of each panel represents the /r/ stimuli, and the top half represents the /l/ stimuli. The unbold portion of each column represents the proportion of correct responses, and the bold portion represents the proportion of incorrect responses. The environments are: i=initial, ic=initial cluster, m=medial, f=final.

by looking at their performance on the /r/ and /l/ stimuli separately across the various phonetic environments (Figure 2 and Table I), we were able to gain some insights into the complex perceptual reorganization that occurs in their categorization behavior over the course of the perceptual training program.

In summary, these perceptual learning data provide a further replication of the findings of the earlier studies that used the high-variability perceptual training procedure to promote the acquisition of the English /r/-/l/ contrast by monolingual, adult Japanese (Logan et al., 1991; Lively et al., 1993; Lively et al., 1994; Yamada, 1993). Having demonstrated significant perceptual learning for these Japanese adults, we now turn to the main concern of this study, that is, an assessment of the effects of perceptual learning on the production of English /r/-/l/ minimal word pairs by these subjects. Our goal here was to assess the extent to which the knowledge acquired during perceptual learning transferred to the production domain.

Effects of Perceptual Learning on Speech Production

Japanese /r/-/l/ Productions

Procedure

During the pretest and post-test phases, audio recordings were made of the Japanese subjects producing English words that contrast /r/ and /l/. The pre- and post-test recordings were made directly before and after the perception pretest and post-test, respectively. The speech production task used a repetition procedure in which the subject simply read a set of English /r/-/l/ minimal pairs from a list of randomly ordered words. The subjects were provided with both visual and auditory prompts. The visual prompts consisted of the list of words written in standard English orthography. The auditory prompts consisted of a recording of the words spoken by a male speaker of General American English. This speaker was not one of the speakers that produced the stimuli for the perceptual identification tasks. The purpose of the auditory prompt was to provide the speakers with a model of how to pronounce the entire word aside from the target /r/ or /l/ segments. The recordings were made in an anechoic chamber at ATR Human Information Processing Research Laboratories. The recordings were digitized at a sampling rate of 22.05 kHz with 16-bit resolution through DAT (Sony PCM-2500 or 2600) and a DAT interface, DAT-Link+ (Townshend Computer Tools., Inc.). The speech files were stored on the hard disk of a Sun Sparc workstation. They were then digitally transferred to the Speech Research Laboratory at Indiana University where they were rescaled to 12-bit resolution for later presentation to native-speakers of English using a PDP-11 laboratory computer.

Stimuli

The stimuli obtained from the pretest and post-test recordings consisted of 55 English words containing /r/ and /l/, giving a total of 110 words. These stimuli for the production tests included /r/-/l/ minimal pairs with the target phoneme in seven phonetic environments. The breakdown of the word-pairs by phonetic environment was as follows: 10 with initial singletons, 10 with initial clusters, 10 with medial singletons, 10 with final singletons, 10 with final clusters, 2 with medial clusters, and 3 with initial triple clusters (e.g., "splint-sprint"). The full set of words is provided in Appendix A. Of these minimal pairs, half in each of the first 5 environments listed above came from the set of minimal pairs that were included in the earlier perceptual training stimuli, the other half were "new," that is, they were not used in any of the perception tests. None of the perceptual stimulus sets included minimal pairs with /r/ and /l/ in the last two

environments listed above: these were all "new" word pairs for our subjects. As noted below, the inclusion of both "new" and "old" words in the pre- and post-test recordings, allowed us to assess the extent to which the Japanese subjects could transfer knowledge gained during perceptual learning to production of novel items.

American English Listeners' Preference Judgments

To assess the transfer of the Japanese trainees' perceptual learning to production, a group of native speakers of American English performed a paired comparison task using each Japanese trainee's pretest and post-test productions. The purpose of this procedure was to assess whether native speakers of American English could reliably tell the difference between the trainees' pre- and post-test productions. If the perceptual learning procedure is effective in producing changes in control over production, then the native English listeners should display a consistent preference for the post-test tokens over the pretest tokens. This paired-comparison method of judging the trainees' improvement in production was selected as an initial test because it was expected to be sensitive to small differences in articulation. Our rationale was that if the Japanese trainees' post-test productions were indeed reliably preferred over the pretest productions, then we would have a reason to submit the pre- and post-test productions to additional tests of perceptual analysis using speakers of American English as listeners. The initial paired-comparison task provides information about the degree and direction of change between the pretest and post-test tokens. A subsequent minimal pair identification task would provide information about a change in speech intelligibility.

Procedure

The procedure for the paired-comparison task was as follows. Each trial began with a visual presentation of the target word in standard English orthography centered on a CRT monitor. The listeners then heard a single Japanese trainee's pretest and post-test productions of this word over headphones. The two versions of the target word were separated by 500 milliseconds of silence. The listeners then had to decide which version of the target word was "better," that is, which version was "a clearer and more intelligible pronunciation of the word shown on the screen." The judges responded on a 7-button response box which was labeled using a 7-point scale where "1" indicated that the first version was "much better" than the second version, "4" indicated no noticeable difference between the two versions, and "7" indicated that the second version was much better than the first. Each pair of utterances was presented twice: once with the pretest version first and the post-test version second, and once in the reverse order. There were 110 pre-post pairs in each of the two presentation orders, plus 10 practice trials at the start of the session, for a total of 230 trials per session. The initial practice trials were excluded from the final data analysis. Each listener judged the full set of pre- and post-test productions from a single Japanese speaker. Ten listeners were assigned to each of the Japanese speakers. Because there were 11 Japanese trained subjects and 12 Japanese control subjects, a total of 230 native English speakers participated as subjects.

In the final data analysis, the responses were collapsed and recoded so that a response of "5" or higher corresponded to a preference for the post-test version, and a response of "3" or lower corresponded to a preference for the pre-test version. This recoding simply takes into account the counter-balanced order of stimulus presentation.

Subjects

The American English listeners were all students at Indiana University. None reported any history of speech or hearing impairment at the time of testing, and all were monolingual native speakers of General American English. All received one hour of course credit for their participation.

Results

The data from the paired-comparison task addressed two specific questions. First, can American English listeners reliably tell the difference between the pretest and post-test tokens? Second, are the post-test tokens more often judged better than the pretest tokens? In order to answer these questions, we examined the percentage of trials that received a rating of "4," indicating no noticeable difference between the pretest and post-test tokens ("post=pre"), the percentage of trials that received a recoded rating of less than 4, indicating a preference for the pretest token ("post<pre"), and the percentage of trials that received a recoded rating of greater than 4, indicating a preference for the post-test token ("post>pre"). Figure 3 shows a summary of the mean scores for the trained subjects (left panel) and for the control subjects (right panel). Table II provides the proportion of responses for each of the 7 response categories.

Insert Table II and Figure 3 about here

As shown in the left panel of Figure 3, the native English speakers gave more "post>pre" ratings than "post<pre" ratings to the trained subjects' tokens. A one-factor ANOVA with response category ("post=pre", "post<pre", "post>pre") as the within-groups factor showed a significant effect of response category ($F(2,30)=42.256$, $p<0.001$). All pairwise comparisons were significantly different at the $p<0.001$ level by Tukey HSD post-hoc comparisons. The right panel of Figure 3 shows the results for the control subjects. The distribution of responses across the three response categories is quite different when compared to the distribution shown in the left panel of this figure. An ANOVA revealed a significant effect of response category ($F(2,30)=15.201$, $p<0.001$) for the control subjects. However, the pairwise comparisons showed no difference between the "post<pre" and "post>pre" response categories for these subjects. (The difference between the "post=pre" category and the other two response categories was significant at the $p<0.001$ level). Thus, response distributions were skewed in favor of a preference for the post-test tokens over the pretest tokens only for the trained subjects. In contrast, the responses were distributed uniformly over the "post<pre" and "post>pre" response categories for the control subjects, indicating no systematic preference for either the pre- or the post-test tokens. Furthermore, the frequency of "post=pre" responses was lower for the trained subjects than for the control subjects ($t(11)=-2.901$, $p=0.014$), also indicating the increased discriminability of the trained subjects' pre- and post-test productions relative to those of the control subjects .

Taken together, these preference data for the trained and control Japanese subjects demonstrate reliable transfer of learning from perception to production. Native-speakers of English were able to reliably detect and discriminate an improvement from the pretest to the post-test productions for the trained subjects; whereas, no reliable difference was observed for the tokens produced by the control subjects. Given that native speakers can discriminate the trained subjects' pre- and post-test tokens, our next step was to assess whether the improvement in these utterances was due to improved intelligibility; that is, whether the Japanese /r/-/l/ productions can be identified more accurately by American English listeners in a minimal pair identification task.

Table II.

Frequency in percentage of total trials of the American English listeners' reponses to the Japanese trained and control subjects' productions on the seven point preference rating scale. Responses have been recoded to take into account the counter-balanced presentation order: 1 indicates a strong preference for the pretest token, 4 indicates no noticeable difference, 7 indicates a strong preference for the post-test token.

Response	Trained	Control
1	3.65	4.12
2	12.0	11.8
3	17.6	20.5
4	20.8	26.6
5	20.4	20.2
6	17.3	11.8
7	7.05	3.93

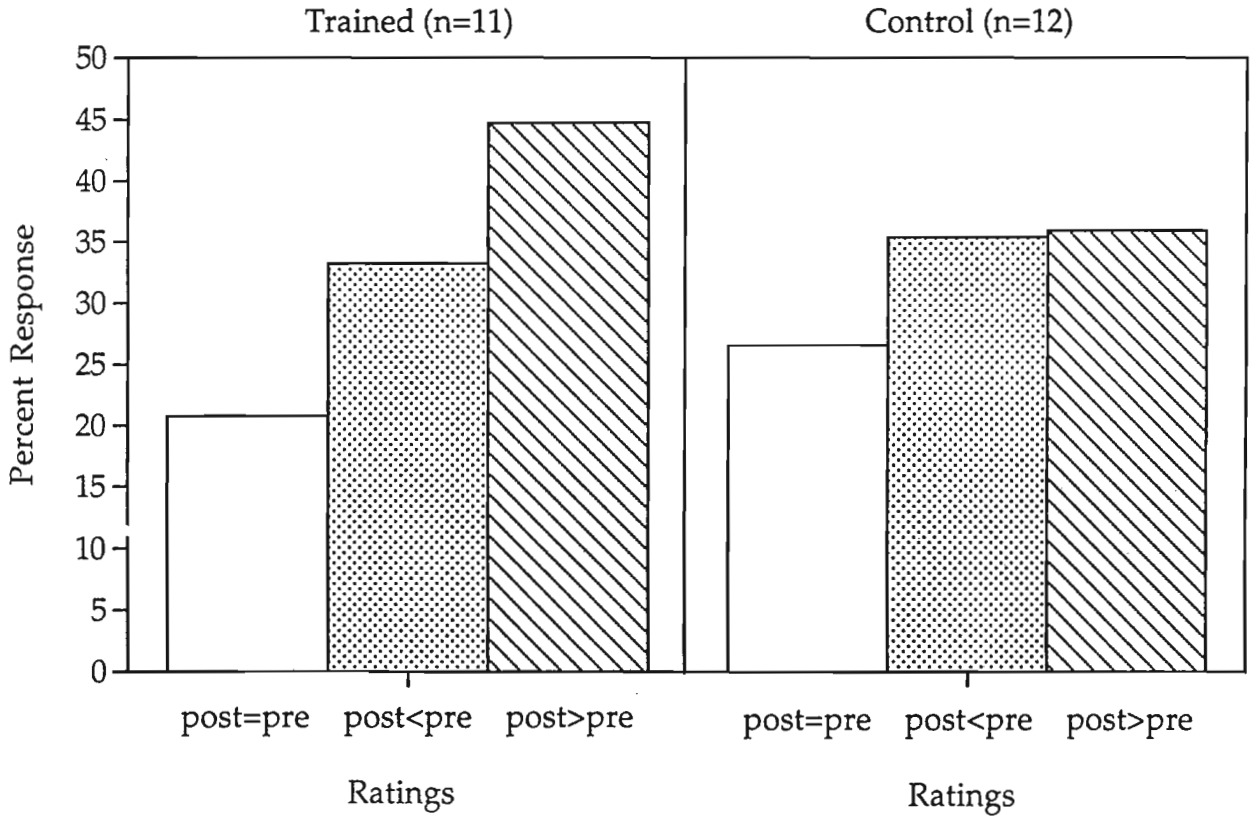


Figure 3. Distribution of preference judgments by the American English listeners for the Japanese trained (left panel) and control (right panel) subjects' pre- and post-test productions. The open bars represent the proportion of trials for which the listeners judged no difference between the pre- and post-test productions, the dotted bars represent the trials for which the pretest token was judged better than the post-test token, and the slashed bars represent the trials for which the post-test token was judged better than the pretest token.

American English Listeners' Identification Data

Procedure

The procedure for the minimal pair identification task was closely modeled after the task that the Japanese trainees performed during perceptual testing and training. In each experimental session, English listeners identified the full set of pre- and post-test productions from a single Japanese trainee. Each trial began with the two members of an English /r/-/l/ minimal pair appearing in standard English orthography on a CRT monitor in front of the subjects. One member of the minimal pair produced by a Japanese trainee was then played out over headphones. The listeners identified the word by pushing the left button on a 2-button response box to select the word on the left of the CRT monitor, or the right button for the word on the right. Within a single experimental session, the complete set of pre- and post-test production tokens from a single Japanese trainee were presented in random order with each word presented twice, once with the correct response as a left button and once with the correct response as a right button. This arrangement resulted in a total of 440 experimental trials plus 10 practice trials at the beginning of the session, for a total of 450 trials. Each of the 11 Japanese trainee's productions were presented in this manner to 10 English listeners for a total of 110 listeners. In addition, each of the 12 Japanese control subjects productions were presented to another group 120 American English listeners (10 listeners for each of the 12 Japanese control subjects).

Subjects

The English listeners were recruited from the university community in Bloomington, Indiana. None reported any history of speech or hearing impairment at the time of testing, and all were monolingual speakers of General American English. All were paid for their participation.

Results

Figure 4 shows percent correct identification of tokens from the Japanese trained (left panel) and control (right panel) subjects' productions as judged by the English listeners. Each panel shows the pretest level of performance along with the identification accuracy of the Japanese productions at the post-test phase for the "old" words (words that were included in the perceptual training stimulus set) and for the "new" words (novel words that the Japanese subjects had not been exposed to in any of the perceptual identification tests). The data shown here are averaged across the five phonetic environments that were included in the perceptual training stimulus set. The remaining two phonetic environments that were included in the production pre- and post-test set (medial clusters and initial triple clusters) were omitted from this "old" versus "new" analysis because there were no "old" stimuli for these two environments.

Insert Figure 4 about here

As shown in Figure 4, utterances from the trained subjects displayed significant improvement in identification from pretest to post-test. Moreover, this improvement was consistent across both the "old" and the "new" items. A one-factor repeated measures ANOVA showed a significant effect of test ($F(2,22)=12.946$, $p<0.001$). Paired t-tests established a significant difference between pretest and "old" post-test items ($t(11)=-3.947$, $p=0.002$), between pretest and "new" post-test items($t(11)=-3.584$, $p=0.004$), but no difference between "old" post-test and "new" post-test items ($t(11)=1.621$, $p=0.133$). In contrast, for

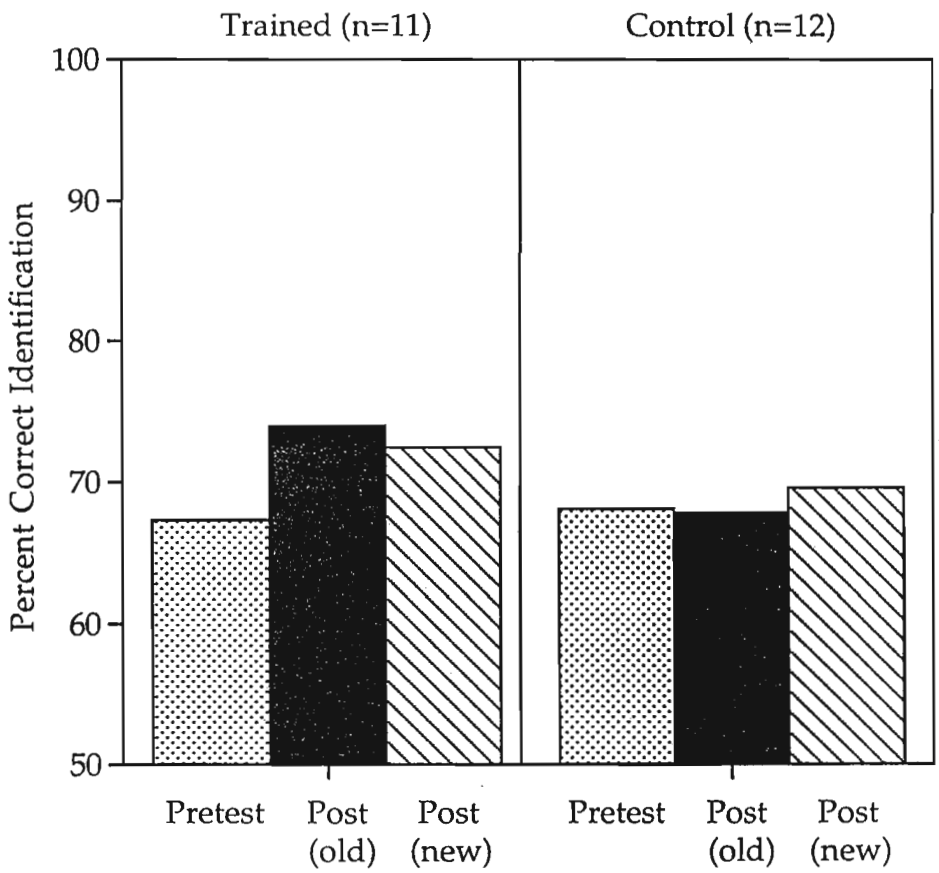


Figure 4. Percent correct performance for trained (left panel) and control (right panel) subjects' productions as judged by American English listeners in the minimal pair identification task. The dotted bar represents the full set of pretest tokens, the solid bar represents the post-test tokens that were included in the perceptual training set, and the slashed bar represents the post-tokens that were not included in the perceptual training set.

the control subjects there was no difference in identification across pretest, "new" post-test or "old" post-test items. These data demonstrate that the overall identifiability of the Japanese trainees' productions improved as a result of the perceptual training program, and that this improvement generalized to both "old" and "new" tokens.

Figure 5 displays the identification accuracy at pretest (left panel) and at post-test (right panel) for the Japanese trainee productions by phonetic environment and by phoneme (/r/ or /l/). This plot includes the two environments that were not included in any of the perception tests (initial triple clusters and medial clusters), as well as the five environments that were included in the training stimulus set. This plot follows the same representation scheme used in Figure 2. The bold portions of each bar represent the proportion of stimuli that were misidentified by the listeners, and the plain portions represent the proportion of /r/ stimuli (bottom part) and /l/ stimuli (upper part) that were correctly identified. Table III gives the accuracy scores represented in this figure.

Insert Table III and Figure 5 about here

A three-factor repeated measures ANOVA was performed with test (pre or post) as the repeated measure, and phoneme and phonetic environment as the within-groups factors. In this analysis, we found a highly significant main effect of the repeated measure factor (i.e., test) ($F(1,154)=27.42$, $p<0.001$) indicating a strong overall improvement in performance from pre- to post-test. This is shown in Figure 5 by the decrease in the bold portions from pretest (left panel) to post-test (right panel). There was also a highly significant main effect of phoneme ($F(1,154)=27.498$, $p<0.001$), due to the generally higher identification accuracy of the /r/ tokens relative to the /l/ tokens at both pre- and post-test. There was no main effect of phonetic environment. There were also no significant interactions between test and either phoneme or environment, indicating that the degree of improvement in speech production was consistent across these factors. Thus, the results show that the Japanese trainees' post-test productions were more accurately identified by native English listeners than their pretest productions. Additionally, these identification data indicate that, at both pretest and post-test, the /r/ tokens were more accurately identified than the /l/ tokens.

The results of the two production evaluation tests (the paired-comparison and the minimal pair identification task) clearly show significant improvements in the Japanese trainees' productions of /r/ and /l/ as a result of perceptual learning. The English listeners consistently judged the post-test utterances to be "better" tokens than the pre-test tokens, and they were more accurate in identifying the post-test tokens in an /r/-/l/ minimal-pair identification task. Furthermore, this improvement in production was robust in that it occurred across all phonetic environments and it even generalized to novel words, i.e., words that the trainees had not been exposed to during perceptual learning. In contrast, the control subjects' productions showed no evidence of any change or improvement across any of these conditions. Having established that the perception training was effective in facilitating improvements in speech production, we now turn to an examination of the relationship between perception and production for individual subjects.

The Relationship Between Learning in Perception and Production

Figure 6 displays the amount of learning in perception and production for each of the 11 Japanese trainees. This figure shows a "perception-production space" where the X-axis represents each trainee's accuracy in perceptual identification of /r/ and /l/ minimal pairs, and the Y-axis represents accuracy in the

Table III.

Pretest and post-test identification accuracies by environment for the Japanese trainee productions as judged by American English listeners.

	/r/		/l/	
Environment	Pretest	Post-test	Pretest	Post-test
initial	68.29	81.29	60.46	68.29
initial cluster	72.83	77.08	49.33	54.92
medial	74.83	77.50	48.00	57.96
final	84.38	82.13	66.54	73.54
final cluster	78.25	81.92	70.29	77.04
initial triple cluster	76.26	84.03	43.47	49.18
medial cluster	68.96	70.63	51.25	59.17
Totals	74.83	79.22	55.62	62.87

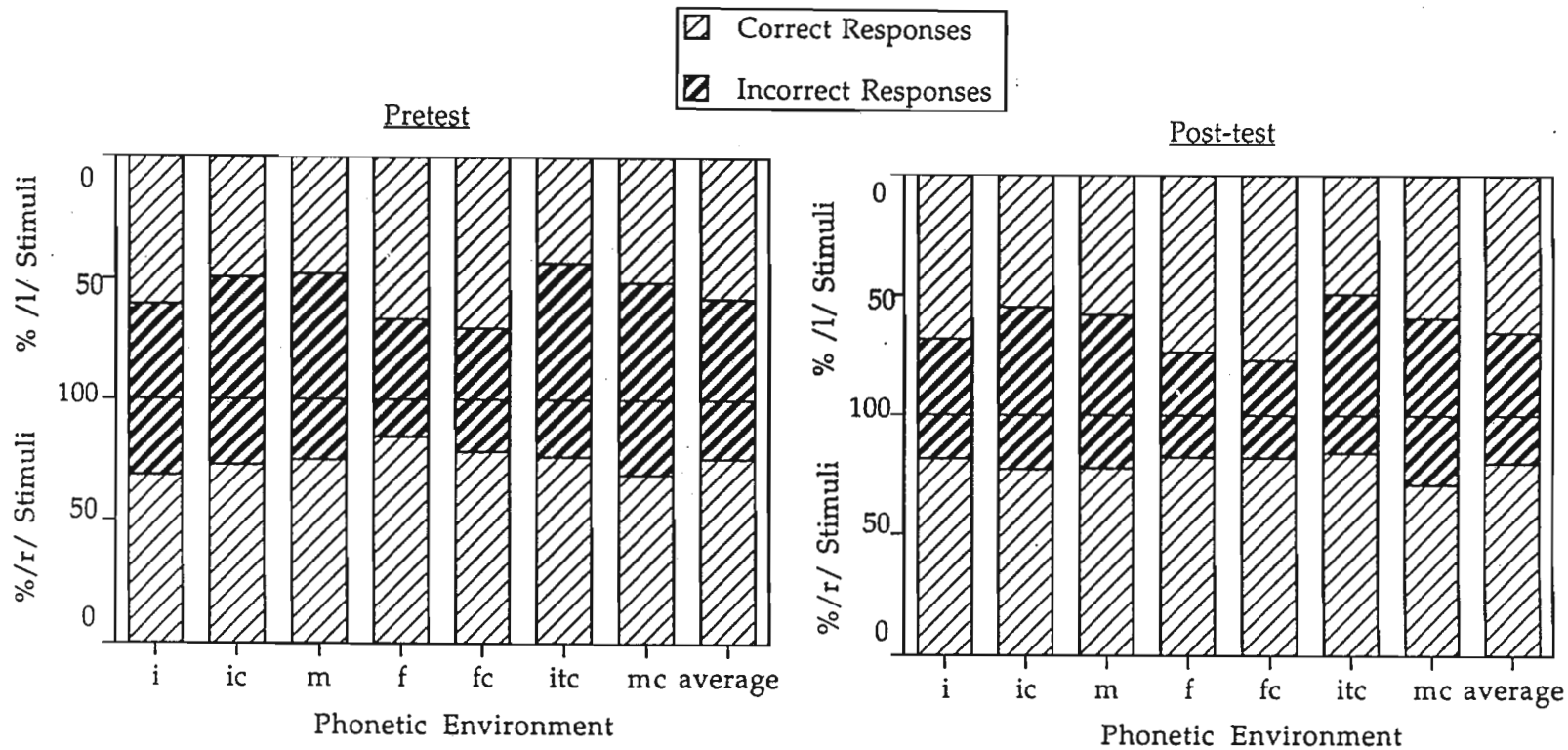


Figure 5. Distribution of the American English listeners' identification accuracy scores for the Japanese trainees' pretest (left panel) and post-test (right panel) productions by phonetic environment. The bottom half of each panel represents the /r/ stimuli, and the top half represents the /l/ stimuli. The unbold portion of each column represents the proportion of correct responses, and the bold portion represents the proportion of incorrect responses. The environments are: i=initial, ic=initial cluster, m=medial, f=final, fc=final cluster, itc=initial triple cluster, mc=medial cluster.

identification of each trainee's productions by American English listeners. Each trainee's performance is represented by a vector whose starting point corresponds to the trainee's pretest performance, and whose end point corresponds to the trainee's post-test performance within this space. The group mean performance is indicated by the bold arrow, and the diagonal is the hypothesized vector that would indicate a perfect correlation between perception and production. Individual subjects' scores are given in Table IV.

Insert Table IV and Figure 6 about here

Considering the perception dimension, it can be seen that each subject's accuracy improved from pretest to post-test, as indicated by the horizontal distance covered by each vector in Figure 6. However, there was considerable variation across subjects in pretest accuracy, in post-test accuracy, as well as in the percentage change from pretest to post-test (see Table IV). For instance, Subjects 2 and 4 both performed at a level above 80% correct at pretest. At post-test, these two subjects approached native-speaker levels of performance. In contrast, Subject 7 was only slightly above chance level at pretest. At post-test, this subject was still performing at a level below the pretest level of the majority of the other subjects. Thus, for the two subjects who performed well in the perception pretest, the training program was effective in enhancing their abilities to identify English /r/ and /l/; whereas, for the poorest performer in the perception pretest, the training program was only moderately effective. It is as if the two high performers used the training sessions to "fine-tune" an already well-defined pre-existing two-way perceptual contrast. Whereas, even after 45 sessions of minimal pair identification training, the poorest performer (Subject 7) barely learned to split a single perceptual category into two new categories, although even this subject showed some improvement after perceptual training. A striking individual difference that emerges from this perceptual identification data can be seen in the comparison between Subjects 9 and 10. These two subjects performed at comparable levels at pretest; however, at post-test a difference of more than 20 percentage points was observed. Subject 10 improved dramatically from 58% to 84% correct identification , whereas Subject 9 showed only a moderate improvement from 56% to 63% correct identification. These individual differences in performance on the perceptual identification tasks are intriguing because every attempt was made to select subjects with similar backgrounds vis à vis exposure to English. Nevertheless, this wide range of individual performance is consistent with previous findings reported in other cross-language studies of /r/-/l/ perception (e.g., Goto, 1971; Mochizuki, 1981; MacKain et al., 1981; Sheldon and Strange, 1982; Yamada et al, 1994).

There are numerous factors that might correlate with individual performance. Here we simply note the strong positive correlation (Pearson $r=+0.664$, $p=0.026$) between pretest level of performance and relative perceptual improvement, where relative perceptual improvement is defined as post-test accuracy minus pretest accuracy divided by 100 minus pretest accuracy. This is a measure of improvement as a proportion of the "room for improvement." This correlation indicates that pretest level of performance is a fairly good predictor of the effectiveness of the perceptual training program for individual subjects; however, as demonstrated by Subjects 9 and 10, there are other important factors at work here too. Furthermore, the present data indicate that there is more variability in the amount of relative perceptual learning for subjects with low pretest performance than for those with high pretest performance. In other words, subjects who performed well at pretest generally showed good relative improvement in perceptual identification after the training program; however, those with low pretest performance showed more variation in relative perceptual improvement.

Table IV.

Individual Japanese trainee perception and production accuracy scores at pretest and at post-test. These data are averaged across /r/ and /l/, as well as across all phonetic environments.

	Perception			Production		
Trainee	Pretest	Post-test	Difference	Pretest	Post-test	Difference
1	67.19	81.25	14.06	57.18	73.00	15.82
2	85.94	95.31	9.37	94.59	95.18	0.59
3	56.25	78.13	21.88	61.18	65.41	4.23
4	82.81	96.88	14.07	99.18	98.95	-0.23
5	65.63	76.56	10.93	60.27	60.91	0.64
6	56.25	76.56	20.31	62.64	72.14	9.5
7	51.56	59.38	7.82	59.64	60.18	0.54
8	68.75	92.19	23.44	74.27	81.32	7.05
9	56.25	62.50	6.25	60.00	76.09	16.09
10	57.81	84.38	26.57	57.00	62.55	5.55
11	67.18	92.18	25.00	56.50	60.55	4.05
Totals	65.06	81.39	16.33	67.50	73.30	5.80

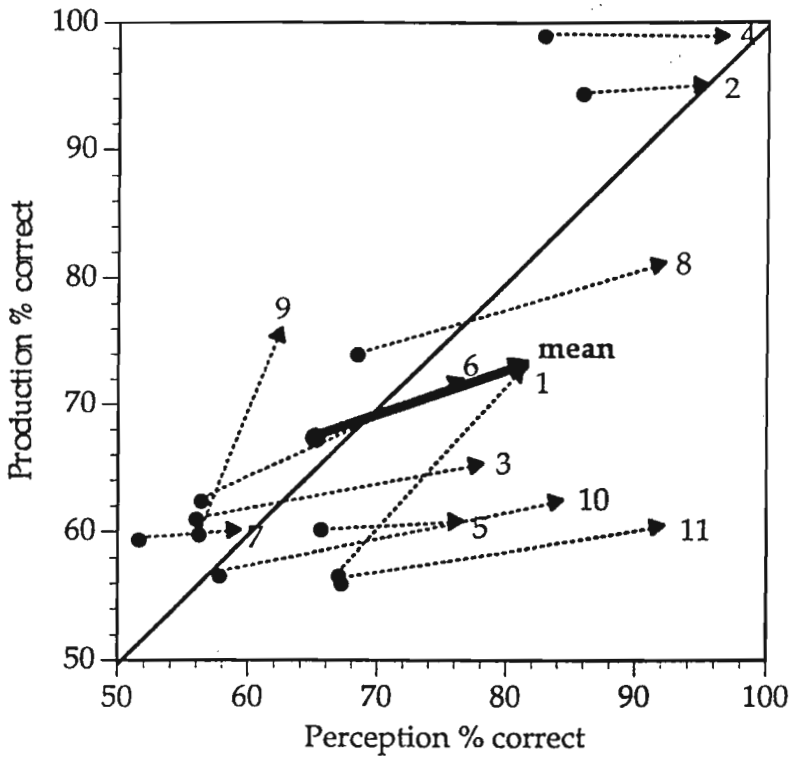


Figure 6. Vector plot of individual Japanese subjects' perceptual identification accuracy (X axis) and production identification accuracy (Y axis) from pretest to post-test. Each individual subject's performance is indicated by a numbered dashed vector. The mean performance is represented by the bold arrow. The diagonal represents the hypothetical vector location and orientation for a perfect correlation between perception and production.

We also observed considerable variation across individual subjects' production performance at pretest and at post-test (Figure 6 and Table IV). Subjects 2 and 4 produced highly intelligible /r/'s and /l/'s at pretest, and therefore had very little room for any improvement to be observed in production. Several of the other subjects' pretest productions were identified at a level around 60% accuracy; however, there were large differences in the degree of production improvement across these subjects. For example, the pretest productions of Subjects 3 and 9 were identified at comparable levels of accuracy; however, Subject 9's post-test productions were identified far more accurately than those of Subject 3.

Given this broad range of individual performance in both perception and production at pretest, we examined some specific characteristics of the relationship between pretest production and perception within individual subjects. Previous studies (Goto, 1971; Sheldon and Strange, 1982) have suggested that /r/-/l/ production can exceed /r/-/l/ perception for Japanese speakers, and that, in general, there is a strong correlation between performance in perception and production (Yamada et al., 1994). An examination of the present perception and production data at the pretest phase indicated that for seven of the eleven subjects, the identifiability of their /r/-/l/ productions exceeded their own ability to identify /r/-/l/ minimal pairs spoken by American English speakers. However, for the remaining 4 subjects, their pretest perception performance exceeded their pretest production performance (see Table IV). Nevertheless, across all subjects, a highly significant correlation was found between performance in perception and production at the pretest phase (Pearson $r=+0.85$, $p=0.001$). In other words, although the relationship between pretest perception and production within individual subjects varied, across all eleven subjects there was a positive correlation between the ability to identify /r/ and /l/ spoken by native English speakers and the ability to produce easily identifiable /r/ and /l/ tokens. Thus, in general, the data from the present study show a similar pattern of results to other earlier findings.

The results from the present investigation allowed us to extend these findings on perception and production by looking at the relationship between changes in one domain (speech perception) and changes in the other domain (speech production). It is clear from the data shown in Figure 6 and Table IV that at the post-test phase, perception performance generally exceeded production performance. This is not surprising since the trainees had extensive training in speech perception, whereas there was no explicit training in speech production. Any improvement observed in speech production was a result of transfer of knowledge gained in perceptual learning to the production domain. By comparing the degrees of improvement in perception and production across the individual trainees, we can obtain some insight into the relationship between learning in the two domains. Specifically, we were interested in investigating whether learning in the two domains proceeds in parallel within individual subjects. This would be the expected if the observed improvement in /r/-/l/ production is a consequence of an improved perceptual representation of the contrast. Alternatively, it is possible that development in perception and production are linked, yet separate processes, in which case learning in the two domains would proceed at different rates within individual subjects.

In order to address this issue we looked at the correlation between the amount of learning in perception and the amount of learning in production across all 11 Japanese trainees. In this analysis we correlated the difference in accuracy scores from pretest to post-test in perception and production across all subjects to test the hypothesis that learning in the two domains proceeds in parallel. In Figure 6, a strong positive correlation would be indicated by a vector that lies parallel to the diagonal shown in the plot, suggesting comparable degrees of improvement in perception and production from pretest to post-test. However, as shown by the bold arrow in the figure (representing the mean improvement in perception and production), there is no such correlation (Pearson $r=+0.049$, $p=0.886$). In other words, it is not the case that improvement in perception and production proceeded in parallel within individual subjects. Rather, it

appears that, although perceptual learning generally transferred to improved production of this non-native contrast, as indicated by the positive slope of the mean vector, the two processes proceed at different rates within individual subjects.

In order to gain a clearer understanding of this lack of correlation between the degrees of learning in perception and production, we investigated specific cases that account for the finding. The first such case is illustrated by Subjects 2 and 4. Both of these subjects had very high pretest performance in perception, and at post-test, both of these subjects performed at a near-native level of /r/-/l/ perceptual identification. In production, both of these subjects performed at an extremely high level at pretest, and therefore had no room for improvement in production. Thus, for these two subjects we observe considerable improvement in perception, but no improvement in production due to a ceiling effect in production.

A second type of situation that leads to the lack of correlation between degrees of learning in perception and production is illustrated by a comparison of Subjects 9 and 10. These two subjects performed at a similar low level of performance in the perception pretest; however, at post-test, Subject 10 performed considerably better than Subject 9. In other words, Subject 10 showed a high degree of perceptual learning; whereas, Subject 9 showed very little improvement in accuracy in the perceptual identification task. However, in production, Subject 9 showed a larger change than Subject 10. A possible explanation for this discrepancy between learning in perception and production is that Subject 9 continued to focus on perceptual cues that are not relevant for /r/-/l/ identification throughout the perceptual training program. However, in the production post-test, this subject was able to implement cues that were effective for improved /r/-/l/ identification in a two-alternative forced-choice identification task with native English listeners. For example, this subject may have focused on durational cues rather than spectral cues, and these cues were sufficient to signal an /r/-/l/ contrast in production but were ineffective for the perceptual identification of /r/ and /l/ by native English speakers.

Finally, a comparison of Subjects 6 and 3 shows that production improvement can vary across individuals, even when initial performance and the degree of learning in perception are comparable. These two subjects performed at similar levels at the pretest phase in both perception and production. They also showed similar degrees of improvement in perceptual accuracy. Nevertheless, Subject 6's post-test productions were identified more accurately by English listeners than Subject 3's post-test productions. In other words, although Subjects 3 and 6 showed comparable gain in perception, the transfer of perceptual learning to production was more effective for Subject 6 than it was for Subject 3.

Our investigation into the relationship between learning in perception and changes in production showed three main results. First, we found a strong positive correlation between initial performance in perception and production within individual subjects. This is seen by the close proximity of the mean vector starting point to the diagonal in Figure 6. Second, we observed a link between perception and production to the extent that perceptual learning generally transferred to improved production. This is seen by the positive slope of the mean vector in Figure 6. And third, we found little correlation between degrees of learning in perception and production after training in perception, due to the wide range of individual variation in learning strategies. This is seen by the deviation of the mean vector from the diagonal in Figure 6. Taken together these findings support the hypothesis that learning in perception and production are closely linked, since perceptual learning generally transferred to improvement in production. However, learning in the perceptual domain is not a necessary or sufficient condition for learning in the production domain: the processes of learning in the two domains appear to be distinct within individual subjects.

General Discussion

The main goal of this study was to investigate the effects of perceptual learning on the production of non-native phonetic contrasts. Our results replicated earlier findings regarding the effectiveness of the high-variability perceptual training program for the acquisition of the English /r/-/l/ perceptual contrast by monolingual Japanese adults. More importantly, the present results extended these earlier findings by showing that the knowledge gained about the novel contrast from perceptual learning transferred to production of English /r/-/l/ words by monolingual Japanese trainees.

Improvement in speech production was assessed in two separate evaluation tests using native English speakers as listeners. The first test was a preference rating task in which the listeners performed a direct comparison of the individual Japanese subjects' pre- and post-test productions. The results of this assessment showed that the English listeners reliably preferred the trainees' post-test productions over their pretest productions. In contrast, the listeners showed no such preference for the control subjects' post-test productions. The second test was a minimal pair identification task in which the listeners identified the test utterances using a two-alternative forced-choice format. The results of this test showed a higher level of identification accuracy for the trainees' post-test productions over their pretest productions. Again, no differences were observed in identification accuracy for the control subjects' pre- and post-test productions. The pattern of results obtained in this study also showed that the learning in both perception and production was highly context-dependent because considerable variation was observed in the Japanese trainees' utterances across different phonetic environments, and across /r/ and /l/. Finally, the data showed that the learning in both perception and production is robust, that is, it transfers to novel items that the trainees had not been exposed to during the training phase of the experiment.

The two sets of data obtained in this study provided us with a means of investigating the relationship between absolute levels of performance in perception and production, as well as the relationship between amounts of learning in the two domains within individual subjects. An examination of the correlations between perception and production showed considerable individual variation across subjects in all aspects of the perception-production relationship. Subjects varied widely in their pretest levels of performance in both perception and production, as well as in the observed improvements in perception and production. Furthermore, the data showed considerable variation across individuals in the transfer of perceptual learning to speech production: subjects who showed a relatively high degree of improvement in perception did not necessarily show a comparable degree of improvement in production. Thus, regarding the relationship between perception and production, the data suggest considerable individual variation in all aspects of this complex relationship. Nevertheless, we observed a high correlation between pretest performance in perception and in production; and in general, the knowledge gained during perceptual training transferred to improved production of the non-native contrast. Thus, when taken together, the results suggest a close link between perception and production.

Having observed transfer of perceptual learning to aspects of speech production, at this point we can speculate as to the mechanisms that facilitate this transfer, and what this tells us about the relationship between speech perception and production at the neural level. In this regard, we considered two theoretical possibilities. The first account supposes that the learning in production involves a mechanism by which articulatory commands are tuned to internal acoustic representations. In this view, the learning in perception leads to more accurate internal acoustic representations of the target speech sounds, and these improved representations function as acoustic templates that play a role in monitoring the articulatory output. Thus, the learning in production occurs during production per se, that is, there is no change in the articulatory commands until they are actually activated during articulation.

An alternative possibility supposes that the articulatory commands are modified during perceptual training. Under this view, the neural changes that result from the perceptual training constitute permanent changes in the internal representation that is common to both perception and production. Thus, according to this view, the learning results in changes in a single, shared representation for perception and production, and as such, accounts for the apparent transfer of learning in perception to improvement in production. The present data do not favor either one of these theoretical possibilities regarding the underlying mechanisms that facilitate the transfer of perceptual learning to speech production. It is possible that both proposed mechanisms operate together to result in the observed transfer from perception to production.

At this point we can also speculate about some of the possible sources of the observed individual differences in non-native perception and production. These might include different general auditory processing strategies, variation in motor control required for speech production, or variation in higher-level cognitive processing strategies. For example, Redmond (1977, cited in Flege, 1988) noted different processing strategies by L2 listeners: some listeners used a "code using" strategy to perceive L2 sounds in terms of L1 categories; whereas, others used a "code forming" strategy to form new categories based on the auditory characteristics of the target L2 categories. Another example of different listener strategies comes from a voice learning study by Nygaard et al. (1994) who found that, after nine days of voice identification training, subjects could be divided into two groups according to their performance level on this task. The "good learners" reached a performance criterion of 70% correct identification of ten voices; whereas the "poor learners" failed to reach this level of performance. This study demonstrates inter-subject variability in a task that requires the development of long-term memory codes for information carried in the speech signal. Although these examples do not reveal the underlying causes or mechanisms of the observed individual differences, they demonstrate the wide range of capabilities relevant for speech perception that are subject to this type of individual difference. Furthermore, as demonstrated by the voice-learning study (Nygaard et al., 1994), these individual differences are not limited to second-language acquisition. In fact, it may be that differences in non-native phonetic contrast perception and production are closely related to differences in specific linguistic or domain-specific cognitive abilities. These speculations regarding the mechanisms that underlie the transfer from perception to production, and regarding individual differences in the acquisition of a non-native phonetic contrast remain to be addressed by both behavioral and neuro-imaging studies.

From an applied point of view, this study provides very encouraging data regarding the acquisition of non-native speech contrasts in laboratory settings. Our findings show very clearly that the high-variability perceptual training procedure is not only effective in training monolingual Japanese adults to perceive the English /r/-/l/ contrast, but that this perceptual training program is also effective in improving the pronunciation of these non-native speech sounds without any explicit training in production. This result is consistent with the recent findings of Rochet (1995) that a change in VOT categorization was accompanied by a change in VOT production for Mandarin speakers exposed to a synthetic French VOT continuum. Additionally, the present results are consistent with the recent findings on phonologically delayed children reported by Jamieson and Rvachew (Jamieson and Rvachew, 1992; Rvachew, 1994; Jamieson and Rvachew, 1994), which showed clear benefits of perception training in conjunction with traditional production therapy for that population. The research by Jamieson and Rvachew as well as the present study focused on cases where the observed pre-training phoneme inventory is reduced relative to the non-native target inventory, and where difficulties in both perception and production are known to occur. Similarly, the research by Rochet focused on a case where the trainees' native categories differed phonetically from the target categories in both perception and production. This type of situation, where

there is a clear match between perception and production characteristics, is likely to be the case where transfer of perceptual learning to speech production will be observed.

Recent models, such as Best's "Perceptual Assimilation Model" (PAM) (Best et al., 1988; Best, 1994; Best, 1995) and Flege's "Speech Learning Model" (SLM) (Flege, 1987, 1992, 1995), provide theoretical frameworks within which we can consider further which non-native contrasts are likely to show transfer of perceptual learning to production, as well as the nature of the mechanisms that underlie this transfer. Both of these models propose that non-native phoneme perception abilities can be explained, at least in part, by reference to the native phonetic space. For example, in the case of the perception of English /r/ and /l/ by Japanese speakers, the observed difficulties can be explained in these models by the fact that Japanese has no such contrast in its native inventory. The most similar native Japanese phoneme is /R/, which is described as an alveolar flap or stop depending on the phonetic context. Thus, with respect to the native Japanese phoneme inventory, English /r/ and /l/ are equally categorizable as this Japanese phoneme, and the contrast is therefore not supported by the native system in either perception or production.

Whereas Best's PAM model makes predictions about the initial difficulty that a given non-native contrast proposes to listeners from a given native language background, Flege's SLM also makes predictions about the persistence of foreign-accented production of a non-native contrast. SLM predicts that as long as native categories subsume non-native categories, accurate perception and production of the target categories will be blocked. Thus, in SLM, it is assumed that improvement in speech production as a consequence of perceptual learning is due to a reorganization of the auditory-acoustic phonetic space which is the underlying system used for both speech perception and production. Thus, SLM would predict that, with respect to adult second-language learning, changes in perception will transfer to changes in production, and that the changes will proceed in parallel. Although PAM is not a model of speech learning, it would make similar predictions regarding the transfer of perceptual learning to speech production. PAM would predict that as the listener becomes more "attuned" to the gestural constellation that characterizes a non-native phoneme, he/she should learn to produce the required gestures for the target phonetic segment.

Both models can account for the main findings of the present study showing transfer of perceptual learning to speech production. However, neither model provides an adequate account of the detailed findings that learning in perception and production do not necessarily proceed in parallel. When individual subject data were examined, we found that subjects who showed the most improvement in perceptual identification of English /r/ and /l/ were not necessarily the same subjects who showed the most improvement in their own productions of English /r/ and /l/. In other words, there were domain-specific aspects to this transfer process that neither SLM nor PAM address directly. It is possible that the domain-specific characteristics and individual subject differences that were found in the present study are related to variability in the sensitivity that the subjects develop to specific /r/ and /l/ characteristics. For instance, as was noted above, some subjects may have developed an increased sensitivity to cues that are not relevant for the identification of /r/ and /l/ words spoken by native English speakers, but these cues may be sufficient to enhance the identifiability of /r/-/l/ productions by native English listeners. The present data thus provide additional insights into the acquisition of non-native phonetic categories in both speech perception and production, and in so doing present a challenge for the future development of these models of cross-language phonetics which rely almost exclusively on the relationship between the pre-training and the target phonetic categories.

In conclusion, we would like to emphasize that our primary goal has been to develop new techniques for the modification of the structure of the trainee's phonetic system. Rather than explicitly focusing the trainee's attention on the detailed physical attributes of the perception and production of the

target contrast, our approach has been to present the trainee with real-world exemplars of the target categories so that he or she can learn to integrate the exemplars into a linguistically meaningful phonetic space. The present study replicated earlier studies that showed the effectiveness of this high-variability approach to the acquisition of a non-native perceptual contrast. More importantly, this study has now extended these results by showing that the changes produced by this approach to non-native phoneme acquisition occur at a level beyond the perceptual domain, that is, the modification of phonetic perception transferred to promote changes in speech production.

Appendix

1. Pretest and post-test word list (taken from Strange and Dittmann, 1984):

Initial singleton		Initial cluster		Intervocalic		Final singleton	
read	lead	breed	bleed	mirror	miller	dear	deal
room	loom	broom	bloom	berry	belly	core	coal
road	load	grow	glow	correct	collect	war	wall
right	light	grass	glass	arrive	alive	tire	tile

Filler words:

deep	keep	swimming	swinging
hope	soap	defend	descend
boat	boot	him	hip
get	got	mad	man

2. Training word list:

initial singleton		initial cluster		intervocalic		final singleton		final cluster	
lash	rash	blue	brew	pallet	parrot	bail	bare	bold	board
late	rate	glue	grew	pilot	pirate	bowl	bore	build	bird
ied	red	play	pray	allay	array	dale	dare	called	card
leer	rear	ply	pry	elect	erect	dial	dire	cold	cord
lewd	rude	blade	braid	oleo	oreo	fail	fair	gold	gourd
lice	rice	blain	brain			feel	fear	halt	heart
limb	rim	blaze	braze			hail	hare	malt	mart
lime	rhyme	bleach	breach			male	mare	mild	mired
lip	rip	blues	bruise			mile	mire	shield	sheared
lit	writ	blush	brush			pail	pare	tiled	tired
lock	rock	clam	cram			roll	roar	wield	weird
long	wrong	clash	crash			soul	sore		
lot	rot	click	crick			tail	tear		
		climb	crime			while	wire		
		clock	crock			misfile	misfire		
		cloud	crowd						
		clout	kraut						
		clown	crown						
		clutch	crutch						
		gloat	groat						
		plop	prop						
		plod	prod						
		blacken	bracken						
		blackfish	brackfish						

3. Test of generalization 1 (new words, new speaker):

initial singleton	initial cluster	intervocalic	final singleton	final cluster
loose	plate	aisle	hole	mars
lush	grows	arrow	wear	malls
rung	groom	pairing	whale	hold
lung	gloom	hearing	veal	health
rug	green	healing	shore	ford
lug	graze		shoal	fold
rough	grade		rare	farce
route	glade		pyre	false
root	fruit		pile	
loot	flute		liar	
loon	fright		leal	
robe	fresh		near	
lobe	flesh		kneel	
rob	fled		hear	
roam	flap		fire	
loam	frame		file	
live	frail		dare	
ring	crack			
ling	clack			
life	bright			
rid	break			
lid	plow			
lick	fly			
leave	flee			
leap	fray			
reek	crew			
leaf	clue			
leech	clay			
raise	claw			
lamb				
rake				
lake				
rare				
rain				
rag				
lag				
lace				

4. Test of generalization 2 (new words, old speaker):

initial singleton	initial cluster	intervocalic	final singleton	final cluster
race	craw	railing	kale	hauled
lack	Cray	raring	heal	hard
rack	flay	aloe	leer	hearth
laid	free	ire	meal	hoard
raid	flow		near	kneeled
lane	fro		peal	neared
lair	fry		peer	pelt
ram	prow		rail	pert
rap	Blake		real	pulse
lathe	bled		rear	purse
wraith	bread		veer	Burt
lave	blight		whore	
rave	class		poor	
laze	crass		pole	
reach	clique			
leak	creek			
leal	cloak			
Reeve	croak			
Lev	clone			
Rick	crone			
ride	flail			
lisle	flame			
ride	frap			
look	Fred			
rook	flight			
lout	flock			
luff	frock			
loan	flog			
roan	frog			
lob	glaze			
rush	glean			
life	prate			
luge	flap			
rouge				
ruse				
lath				
wrath				

5. Production word list:

(i) "Old" words (from training set):

initial singleton		initial cluster		intervocalic		final singleton		final cluster	
late	rate	breach	bleach	allay	array	feel	fear	shield	sheared
lock	rock	blade	braid	pilot	pirate	tail	tear	halt	heart
lewd	rude	clock	crock	oleo	oreo	wire	while	cold	cord
lice	rice	gloat	groat	pallet	parrot	soul	sore	build	bird
lip	rip	blues	bruise	elect	erect	dial	dire	tiled	tired

(ii) "New" words:

initial singleton		initial cluster		intervocalic		final singleton		final cluster	
link	rink	blink	brink	Cilia	Syria	seal	sear	pealed	peered
lope	rope	clop	crop	alight	aright	ball	bar	bald	bard
list	wrist	bloke	broke	telly	Terry	foal	four	colt	cart
lent	rent	bland	brand	Helen	heron	tall	tar	scowled	scoured
lamp	ramp	gland	grand	jelly	Jerry	call	car	wild	wired

medial clusters		initial triple clusters	
eaglet	egret	splay	spray
gently	gentry	splint	sprint
		split	sprit

References

Atkinson, R.C. (1972). Ingredients for a theory of instruction. *American Psychologist* **27**, 921-931.

Best, C.T., McRoberts, G.W., and Sithole, N.M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance* **14**, 345-360.

Best, C.T. (1994). Learning to perceive the sound pattern of English. In *Advances in Infancy Research,* edited by C. Rovee-Collier and L. Lipsitt (Norwood, NJ: Ablex).

Best, C.T. (1995). A direct-realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language speech research,* edited by W. Strange (Timonium, MD: York Press), Pp. 171-206.

Browman, C.P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica,* **49**, 155-180.

Flege, J.E. (1987). The production of new and similar phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, **15**, 47-65.

Flege, J.E. (1988). The production and perception of foreign language speech sounds. In *Human communication and its disorders: A review*, edited by H. Winitz (Norwood, NJ: Ablex), Pp. 224-401.

Flege, J.E. (1992). Speech learning in a second language. In *Phonological Development: Models, research, and application,* edited by C. Ferguson, L. Menn, and C. Stoel-Gammon. Timonium, MD: York Press.

Flege, J.E. (1995). Second language speech learning: Theory, findings and problems. In *Speech perception and linguistic experience: Issues in cross-language research* edited by W. Strange. Timonium, MD: York Press, Pp. 233-272.

Flege, J.E. and Hillenbrand, J. (1987). Limits on phonetic accuracy in foreign language speech production. In *Interlanguage phonology: The acquisition of a second language sound system*, edited by G. Ioup and S. Weinberger Newbury House: Cambridge.

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minn. Papers Ling. Philos. Lang.*, **6**,59-72.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds 'l' and 'r'. *Neuropsychologia*, **9**, 317-323.

Jamieson, D.G. and Morosan, D.E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception and Psychophysics,* **40**, 205-215.

Jamieson, D.G. and Rvachew, S. (1992). Remediating speech production errors with sound identification training. *Journal of speech-language pathology and audiology*, **16**, 201-210.

Jamieson, D.G. and Rvachew, S. (1994). Perception, production and training of new consonant contrasts in children with articulation disorders. Proceedings of the International Conference on Spoken Language Understanding. Yokohama: Acoustical Society of Japan, 199-1202.

Lively, S.E., Logan, J.D., and Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ ad /l/: II. the role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America,* **94**, 1242-1255.

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., and 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.

Logan, J.D., Lively, S.E., and Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874-886.

MacKain, K.S., Best, C.T., and Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, **2**, 369-390.

Magnuson, J.S., Yamada, R.A., Tohkura, Y., Pisoni, D.B., Lively, S.E., and Bradlow, A.R. (1995). The role of talker variability in non-native phoneme training. Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan, 393-394.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A.M., Jenkins, J.J., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception and Psychophysics*, **18**, 331-340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics*, **9**, 283-303.

Morosan, D., and Jamieson, D. (1989). Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research,* **32**, 501-511.

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

Redmond, L.D. (1977). *Learning to recognize foreign speech sounds: Strategies of auditory processing.* Unpublished Ph.D. thesis, University of California, San Diego, California.

Rochet, B.L. (1995). Perception and production of second-language speech sounds by adults. In *Speech perception and linguistic experience: Issues in cross-language research* edited by W. Strange. Timonium, MD: York Press, 379-410.

Rvachew, S. (1994), Speech perception training can facilitate sound production learning. *Journal of Speech and Hearing Research*, **37**, 347-357.

Sheldon, A., and Strange, W. (1982), The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, **3**, 243-261.

Strange, W., and Dittmann, S. (1984), Effects of discrimination training on the perception of /r/-/l/ by Japanese adults earning English. *Perception and Psychophysics,* **36**, 131-145.

Tahta, S., Wood, M., and Loewenthal, K. (1981). Foreign accents: Factors relating to transfer of accent from the first language to a second language. *Language and Speech,* **24**, 265-272.

Yamada, R.A. (1993). Effects of extended training on /r/ and /l/ identification by native speakers of Japanese. *Journal of the Acoustical Society of America*, **93**(2), 2391.

Yamada, R.A., and Tohkura, Y. (1991). Perception of American English /r/ and /l/ by native speakers of Japanese. In *Speech Perception, Production, and Linguistic Structure,* edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Tokyo: Ohmsha, Ltd.).

Yamada, R.A., Strange, W., Magnuson, J.S., Pruitt, J.S., and Clarke, W.D. III (1994). The intelligibility of Japanese speakers' productions of American English /r/, /l/, and /w/, as evaluated by native speakers of American English. Proceedings of the International Conference of Spoken Language Processing, Yokohama, 1994.

Yamada, R.A. and Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception and Psychophysics,* **52**, 376-392.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Intelligibility of Normal Speech I: Global and Fine-Grained Acoustic-Phonetic Talker Characteristics[1]

Ann R. Bradlow, Gina M. Torretta and David B. Pisoni

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

1 This research was supported by NIDCD Training Grant DC-00012 and by NIDCD Research Grant DC-00111 to Indiana University. We are grateful to Luis Hernandez for technical support, and to John Karl for compiling the Indiana Multi-talker Sentence Database. Earlier versions of some of this work were presented at the 127th meeting of the Acoustical Society of America, at the XIIIth International Congress of Phonetic Sciences, and at the 130th meeting of the Acoustical Society of America. Copies of the Indiana Multi-Talker Sentence Database can be obtained in CD-ROM form for a nominal cost for media and postage.

Abstract

This study used a multi-talker database containing intelligibility scores for 2000 sentences (20 talkers, 100 sentences), to identify talker-related correlates of speech intelligibility. We first investigated "global" talker characteristics (e.g., gender, F0 and speaking rate). Findings showed female talkers to be more intelligible as a group than male talkers. Additionally, we found a tendency for F0 range to correlate positively with higher speech intelligibility scores. However, F0 mean and speaking rate did not correlate with intelligibility. We then examined several fine-grained acoustic-phonetic talker-characteristics as correlates of overall intelligibility. We found that talkers with larger vowel spaces were generally more intelligible than talkers with reduced spaces. In investigating two cases of consistent listener errors (segment deletion and syllable affiliation), we found that these perceptual errors could be traced directly to detailed timing characteristics in the speech signal. Results suggest that a substantial portion of variability in normal speech intelligibility is traceable to specific acoustic-phonetic characteristics of the talker. Knowledge about these factors may be valuable for improving speech synthesis and recognition strategies, and for special populations (e.g., the hearing-impaired and second-language learners) who are particularly sensitive to intelligibility differences among talkers.

Intelligibility of Normal Speech I: Global and Fine-Grained Acoustic-Phonetic Talker Characteristics

It is well known that even under "ideal" speaking and listening conditions, there is a wide range of individual differences in overall speech intelligibility across normal talkers (e.g. Black, 1957; Hood and Poole, 1980; Bond and Moore, 1994). Additionally, recent studies on the role of talker variability in speech perception and spoken word recognition have shown that listeners are sensitive to talker variability to the extent that speech intelligibility scores decrease with increased talker variability in the test materials (e.g. Mullennix et al. 1989; Pisoni, 1993; Sommers et al., 1994; Nygaard et al., 1994). Moreover, listeners show evidence of encoding talker-specific voice attributes in memory along with information about the specific test words (Palmeri et al., 1993). Nygaard et al. (1994) also reported that familiarity with a talker's voice leads to an advantage in intelligibility of speech produced by that talker, suggesting a direct link between listener sensitivity to paralinguistic, talker-specific attributes and overall speech intelligibility. Thus, there is a growing body of research showing that the linguistic content of an utterance and the indexical, paralinguistic information, such as talker- and instance-specific characteristics, are not only simultaneously conveyed by the acoustic signal, but also are not dissociated, or normalized away, during speech perception (Ladefoged and Broadbent, 1957; Laver and Trudgill, 1979).

Similarly, studies of within-talker variability in speech production have shown that talkers systematically alter their speech patterns in response to particular communicative requirements in ways that have substantial effects on the overall intelligibility of an utterance. For example, in a series of studies on speech directed towards the hard of hearing, Picheny et al. (1985, 1986, 1989) and Uchanski et al. (in press) found systematic, acoustic-phonetic differences between "clear" and "conversational" speech within individual talkers. Clear speech had consistently higher intelligibility scores, and was found to be slower and to exhibit fewer phonological reduction phenomena than conversational speech. Lindblom (1990) and Moon and Lindblom (1994) showed that talkers adapt their speech patterns to both production-oriented and listener-oriented factors as demanded by the specific communicative situation. For example, formant frequencies of vowels embedded in words spoken in "clear speech" exhibited less contextually conditioned undershoot than those embedded in words spoken in "citation form."

Recently, Bond and Moore (1994) investigated whether the acoustic-phonetic characteristics that apparently distinguish "clear" versus "conversational" speaking styles within a talker also distinguish the speech across talkers who differ in overall intelligibility. Indeed, in a comparison of the acoustic-phonetic characteristics of the speech of a relatively high intelligibility talker and two talkers with relatively low intelligibility, Bond and Moore found that "inadvertently" clear speech shared many of the acoustic-phonetic characteristics of intentionally clear speech. Finally, Keating et al. (1994) and Byrd (1994) investigated inter-talker variability in pronunciation of American English from tokens in the TIMIT database of American English dialects (Lamel et al., 1986; Pallett, 1990; Zue et al., 1990). Both of these studies revealed the broad range of pronunciation characteristics in American English, and pointed out how paralinguistic factors, such as the talker's gender, dialect and age, in addition to linguistic factors, such as phonetic context, contribute to the observed pronunciation variability. However, since the TIMIT database does not include perceptual data, neither of these studies made inferences regarding the effects of these inter-talker differences on overall speech intelligibility.

The goal of the present study was to extend our understanding of the talker-specific characteristics that lead to variability in speech intelligibility by investigating the acoustic correlates of different talkers' productions in a large database that includes both sentence productions from multiple talkers and intelligibility data from multiple listeners per talker (Karl and Pisoni, 1994). The basic question we asked

was, "What acoustic characteristics make some talkers more intelligible than others?" By directly assessing talker-specific correlates of speech intelligibility at the acoustic-phonetic level this investigation aimed to extend our understanding of the relationship between the indexical and linguistic aspects of speech communication: we hoped to identify some of the aspects of talker variability that might on the one hand, be expected to help identify a particular talker, and on the other hand have a direct effect on overall speech intelligibility.

We acknowledge that it is misleading to ascribe all of the variability in sentence intelligibility to acoustic-phonetic characteristics of the talker. Such an approach incorrectly disregards any listener-talker-sentence interactions that affect the resultant intelligibility score. Nevertheless, while keeping in mind the contribution of listener- and sentence-related factors to overall intelligibility, we were interested in investigating what talker-related characteristics, independently of the listener- and sentence-related characteristics, might correlate with overall intelligibility, and therefore might account for some portion of the observed variability in overall intelligibility. We hoped that the results of this investigation combining both acoustic-phonetic measurements with perceptual data might lead to a better understanding of the salient acoustic-phonetic characteristics that listeners respond to during speech perception, and would therefore help to differentiate highly intelligible speech from less intelligible speech.

We adopted an approach that focused on two aspects of talker-specific characteristics. First, we focused on "global" talker characteristics, such as gender, fundamental frequency and rate of speech. These characteristics are "global" because they extend over the entire set of utterances from a given talker, rather than being confined to local aspects of the speech signal that are related to the articulation of individual segments. Second, we focused on specific pronunciation characteristics, such as vowel category realization and segmental timing relations that are fine-grained, acoustic-phonetic indicators of instance-specific variability. Whereas the global characteristics provide information about some of the invariant speech attributes of the individual talkers, the fine-grained acoustic-phonetic details at the local, segmental level provide information about the instance-specific pronunciation characteristics of particular utterances. We expected that a wide range of these talker-related characteristics would contribute to variability in overall intelligibility, and hoped that this approach would provide a better understanding of some of the talker- and instance-specific factors that are associated with highly intelligible normal speech.

The Indiana Multi-Talker Sentence Database

The materials for this study came from the Indiana Multi-Talker Sentence Database (Karl and Pisoni, 1994). This database consists of 100 Harvard sentences (IEEE, 1969) produced by 20 talkers (10 males and 10 females) of General American English. The sentences are all mono-clausal and contain 5 key words plus any number of additional function words. None of the talkers had any known speech or hearing impairments at the time of recording, and all recordings were live-monitored for gross misarticulations, hesitations, and other disfluencies. (See Table II for examples of the sentences.) The sentences were presented to the subjects on a CRT screen in a sound-attenuated booth (IAC 401A). The stimuli were transduced with a Shure (SM98) microphone, and digitized on-line (16-bit analog-to-digital converter (DSC Model 240) at a 20 kHz sampling rate). The average root mean square amplitude of each of the digital speech files was then equated with a signal processing software package (Luce and Carrell, 1981), and the files were converted to 12-bit resolution for later presentation to listeners in a transcription task using a PDP-11/34 computer.

Along with the audio recordings, this database also includes speech intelligibility data in the form of sentence transcriptions by 10 listeners per talker, for a total of 200 listeners. In collecting these

transcriptions, each group of 10 listeners heard the full set of 100 sentences produced by a single talker. The sentence stimuli were low-pass filtered at 10kHz, and presented binaurally over matched and calibrated TDH-39 headphones using a 12-bit digital-to-analog converter. The listeners heard each sentence in the clear (no noise was added) at a comfortable listening level (75 db SPL), and then typed what they heard at a computer keyboard. A PDP-11/34 computer was used to control the entire experimental procedure in real-time. The listeners were all native speakers of American English, and were students at Indiana University with no speech or hearing impairments at the time of testing.

The sentence transcriptions were scored by a key word criterion that counted a sentence as correctly transcribed if, and only if, all 5 keywords were correctly transcribed. Any error on a keyword resulted in the sentence being counted as mistranscribed. With this scoring method, each sentence for each talker received an intelligibility score out of a possible 10. Each talker's overall intelligibility score was then calculated as the average score across all 100 sentences.

Figure 1 (see also Table I) shows a summary of the speech intelligibility scores that we collected from the 20 talkers: the overall sentence intelligibility scores ranged from 81.1% to 93.4% correct transcription, with a mean and standard deviation of 87.8% and 3.1%, respectively. Thus, the materials in this large multi-talker sentence database showed considerable variation and covered a range of talker intelligibility that could be used as the basis for an investigation of the effects of global and fine-grained acoustic-phonetic talker characteristics on overall speech intelligibility.

Insert Figure 1 and Table I about here

It is important to note here that intelligibility scores must be interpreted in a relative sense. For example, Hirsh et al. (1954) observed that authors on this subject almost always caution readers "... to regard such scores as specific to a given crew of talkers and a given crew of listeners." In the present study, we were specifically interested in exploring the individual characteristics of our "crew of talkers," however, our database was constructed in such a way that it did not provide the means of investigating the contribution of the "crew of talker" independently of the "crew of listeners." This is because for each talker, a different group of 10 listeners, drawn from the same population, transcribed the recordings of the full set of 100 sentence. Therefore, the intelligibility scores for the 20 talkers shown in Figure 1, as well as the talker-related correlates of intelligibility that we discuss below, should, strictly speaking, be regarded as reflecting characteristics of the particular talker-listener situation, rather than of the talker independently of the listener.

Global Talker Characteristics

Gender

We began by investigating global talker characteristics that could provide an indication of the relationship between source-related acoustic characteristics and overall speech intelligibility scores. Although all of the talkers in our database were judged to have normal voice qualities, we investigated whether some voice qualities would be associated with higher speech intelligibility scores than others. In particular, we wondered whether talker gender would be a correlate of variability in intelligibility.

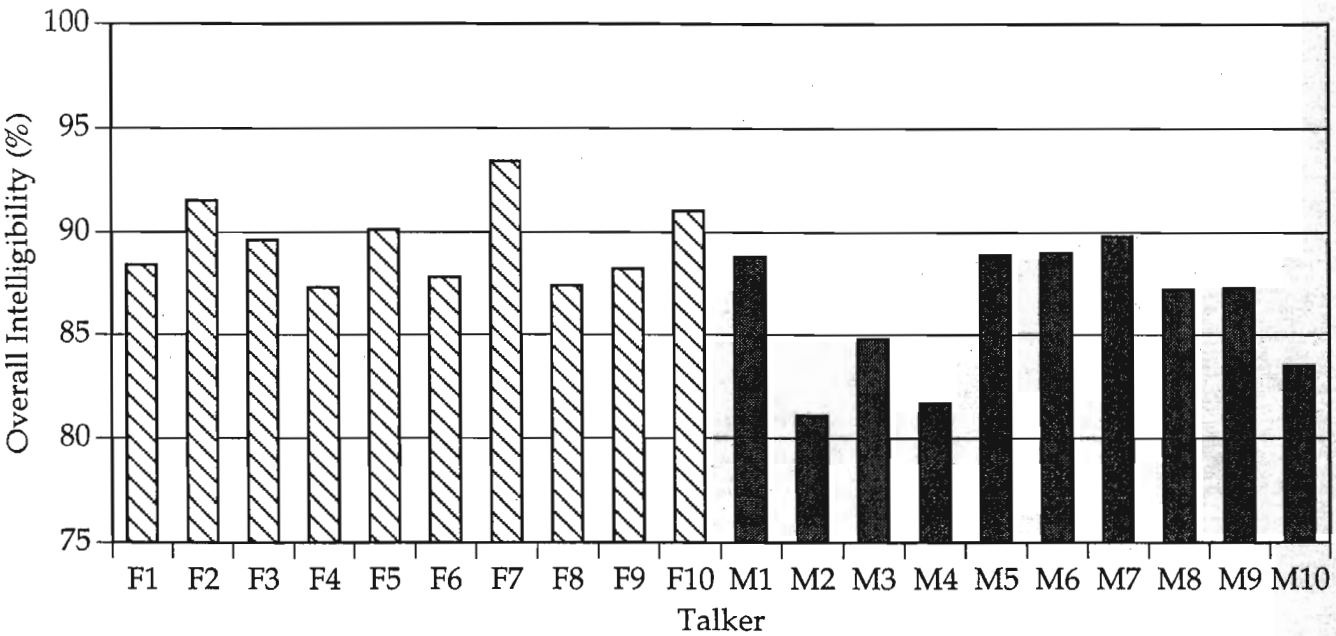


Figure 1. Overall intelligibility scores for each of the 10 female talkers (columns filled with slanted lines) and 10 male talkers (shaded columns).

Table I

Intelligibility scores across all 100 sentences, mean sentence duration, fundamental frequency mean and range for each individual talker.

Talker	Intelligibility (100 sentences)	F0 Mean (Hz)	F0 Range (Hz)	Mean Sentence Duration (sec.)
F1	88.4	208.4	218.9	2.8
F2	91.5	168.4	134.0	1.9
F3	89.6	178.5	150.2	1.9
F4	87.3	210.7	164.5	1.8
F5	90.1	220.7	226.8	1.9
F6	87.8	203.3	231.0	2.2
F7	93.4	162.8	139.7	2.2
F8	87.4	206.9	124.5	1.9
F9	88.2	206.9	158.3	2.2
F10	91.0	237.3	205.6	2.0
M1	88.8	119.6	114.4	2.3
M2	81.1	141.8	131.8	2.1
M3	84.8	130.3	96.0	2.8
M4	81.7	102.5	76.6	1.9
M5	88.9	110.2	100.1	2.1
M6	89.0	140.3	140.1	2.0
M7	89.8	100.0	73.4	2.3
M8	87.2	118.7	103.4	2.0
M9	87.3	118.7	115.8	2.1
M10	83.5	104.0	79.5	1.9

Male and female glottal characteristics differ considerably (Klatt and Klatt, 1990; Hanson, 1995), and listeners are generally able to distinguish male from female voices quite easily (Nygaard et al., 1994; Tielen, 1992). Furthermore, Byrd (1994) found that male speech in the TIMIT database of American English was characterized by a greater prevalence of phonological reduction phenomena, such as vowel centralization, alveolar flapping, and reduced frequency of stop releases, relative to female speech in this database. Thus, there is evidence that gender is a salient characteristic that we might expect to affect overall intelligibility. Specifically, we hypothesized that more "reduced" speech would lead to lower overall intelligibility, and therefore that the group of female talkers in our database might have a higher mean overall intelligibility score than the group of male talkers. Indeed, we found that the group of 10 female talkers did have a significantly higher overall intelligibility score than the group of 10 male talkers (89.5% versus 86.2% correct transcription with standard deviations of 2.0% and 3.2%, respectively; $t(18)=2.72$, $p=0.01$ by a 2-tail unpaired t-test). Furthermore, the four talkers with the highest overall intelligibility scores were female and the four talkers with the lowest overall intelligibility scores were male (see Figure 1). Thus, for the 20 talkers in our database, there was a significant gender-based difference in overall intelligibility with the female talkers being generally more intelligible than the male talkers. This result raised the question of what specific acoustic-phonetic characteristics lead to this gender-based intelligibility difference. Byrd's (1994) analyses suggest that this intelligibility difference might be due to an increased prevalence of specific reduction phenomena for male speech relative to female speech, rather than due to the source-related (voice quality) differences between males and females. However, before turning to a discussion of fine-grained pronunciation differences, we examined several other global talker characteristics that might provide information about the relationship between talker-specific factors and overall speech intelligibility.

Fundamental Frequency

Fundamental frequency is a global talker characteristic that typically differs markedly across male and female talkers. However, it is not clear that it is an acoustic attribute that directly affects overall intelligibility. Bond and Moore (1994) found no reliable difference in mean fundamental frequency between their higher and lower intelligibility talkers. Similarly, Picheny, Durlach and Braida (1986) found that for all three talkers in their study, clear speech was characterized by a somewhat wider range in fundamental frequency with a slight bias towards higher fundamental frequencies than conversational speech, however, these differences were not dramatic. In the present study, we investigated both fundamental frequency mean and range as possible correlates of overall intelligibility; however, based on these previous studies, we had no strong predictions regarding the relationship between fundamental frequency characteristics and intelligibility.

All fundamental frequency measurements were made using the Entropics WAVES+ software (version 5.0) on a SUN workstation. For each sentence produced by each talker, the mean, minimum and maximum fundamental frequency was extracted from the voiced portions of the digital speech file using the pitch extraction program included in the Entropics WAVES+ software package. Each talker's overall mean, minimum and maximum fundamental frequency was then calculated across all 100 sentences. These values are given in Table I and Figure 2.

Insert Figure2 about here

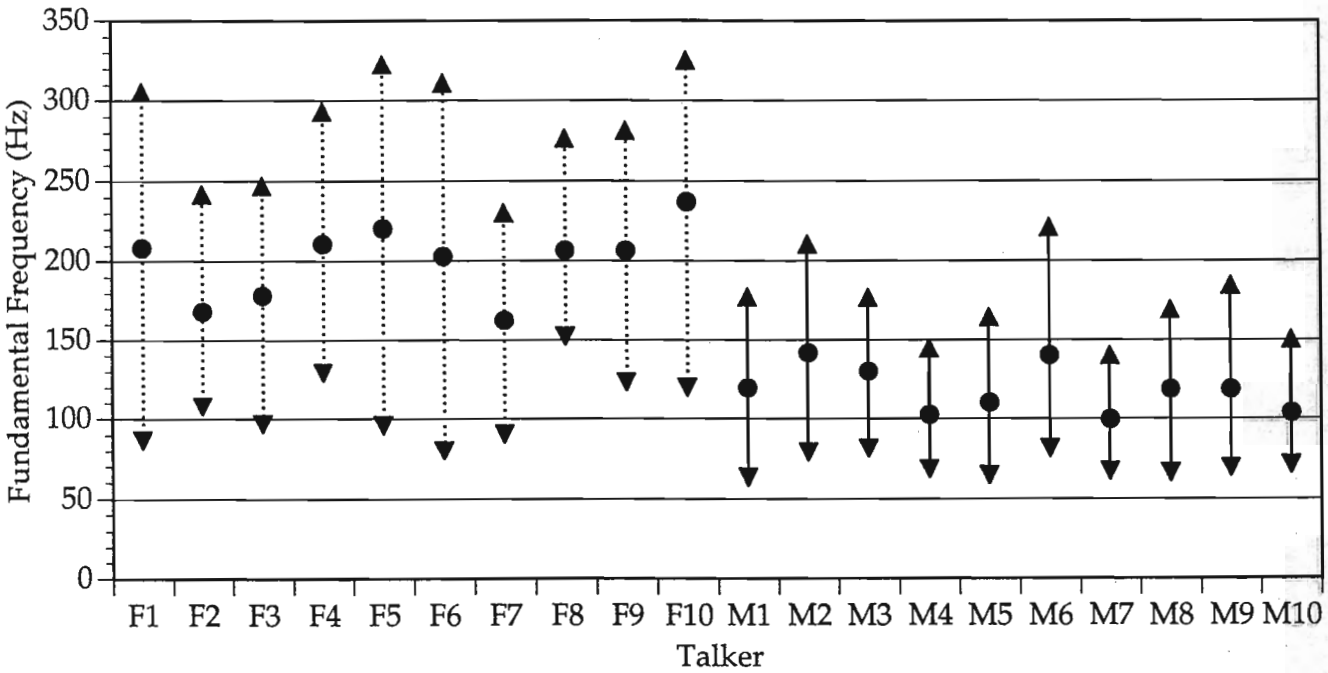


Figure 2. Minimum, maximum and mean fundamental frequency in Hertz for each of the 10 female talkers (dotted arrows) and 10 male talkers (solid arrows).

Because the female talkers taken as a group had a higher mean intelligibility score, and as expected the female talkers had a higher mean fundamental frequency, across all talkers we found a slight tendency for a higher mean fundamental frequency to correlate with a higher mean intelligibility score (Spearman rho = +0.341, p = 0.14). However, when we looked at the males and females separately, we found no such correlation between mean fundamental frequency and intelligibility. Thus, in our database, overall intelligibility is not correlated with mean fundamental frequency independently of the gender-based difference in overall intelligibility. With respect to fundamental frequency range, across all 20 talkers we found a tendency for a wider range in fundamental frequency to correlate with a higher overall intelligibility score (Spearman rho = +0.384, p= 0.095). We also found a significantly greater fundamental frequency range for the group of female talkers than for the group of male talkers ($t(18)=4.87$, $p<0.001$ by a 2-tailed unpaired t-test.) The fundamental frequency range, mean and standard deviation were 175 Hz and 41 Hz for the female group, and 103 Hz and 23 Hz for the male group, respectively. Since this finding is correlational, we cannot be certain whether the wider fundamental frequency range leads to higher intelligibility or whether both the higher intelligibility and wider fundamental frequency range are simply consequences of some other voice quality attribute of our female talkers. One piece of evidence that bears on this issue comes from a recent study by Tielen (1992) who found that, although female speakers of Dutch typically had higher mean fundamental frequencies than their male counterparts, they did not have significantly wider fundamental frequency ranges. For the purposes of the present study, this finding indicates that a wider fundamental frequency range is not a necessary consequence of a higher fundamental frequency mean. It is therefore possible that the wider female fundamental frequency range is one of the female speech characteristics that contributes to the generally higher intelligibility of female speech relative to male speech in our database.

Speaking Rate

The final global talker characteristic that we investigated was overall speaking rate. Although speaking rate is not a source-related, voice-quality characteristic, it is one of the most salient global talker-specific characteristics, and one that is known to distinguish "clear" versus "conversational" speech within individuals (Picheny et al., 1989; Krause and Braida, 1995; Uchanski et al., in press). Additionally, many phonological reduction phenomena are directly related to increased speaking rate. In Byrd's analyses of the TIMIT database, which included sentences from 630 talkers, she found that across all dialects, the males had significantly faster speaking rates than the females on the two calibration sentences that were read by all talkers. However, Byrd's study also found an interaction of gender and dialect region such that the slowest speaking region for the male speakers (the South Midland) was only the fourth slowest for the female speakers. Bond and Moore (1994) found no word duration differences in their analyses of two talkers that differed in overall intelligibility when the words were embedded in sentences, although for isolated words the less intelligible talker had shorter durations than the more intelligible talker. Furthermore, in a recent study of the effects of speaking rate on the intelligibility of clear and conversational speaking modes, Krause and Braida (1995) reported that trained talkers were able to achieve an intelligibility advantage for the clear speech mode even at faster speaking rates. In other words, it is possible to produce fast clear speech. Thus, although there is some evidence that overall speaking rate varies with paralinguistic (indexical) characteristics such as speaker's gender and dialect, and that speaking rate can be associated with a "reduced," conversational speaking style, a direct link between speaking rate and intelligibility remains unclear.

In our database, we measured overall speaking rate for each of the 20 talkers, as the mean sentence duration across all 100 sentences. All duration measurements were made using the Entropics WAVES+ software on a SUN workstation. The questions we asked here were: (1) Does overall speaking rate correlate

with overall speech intelligibility across all 20 talkers? And, (2) Can the gender-based intelligibility difference be traced to a gender-based difference in overall speaking rate? As shown in Figure 3 and Table I, we observed considerable variability across all 20 talkers in mean sentence duration (mean sentence duration = 2.115 seconds, with a standard deviation of 0.276 seconds). However, we failed to find a clear relationship between mean sentence duration and overall speech intelligibility scores: there was no correlation between speaking rate and speech intelligibility across all 20 talkers, and there was no significant difference in the means between the male and female speaking rates.

Insert Figure3 about here

Thus, in our multi-talker sentence database, overall speaking rate did not appear to be a talker-related correlate of variability in speech intelligibility. This result is consistent with the recent finding of Krause and Braida (1995) that fast speech can also be "clear" speech, and with Bond and Moore (1994) who found no difference in duration for words in sentences spoken by a high and a low intelligibility talker. Furthermore, even though the present data do not show a difference between male and female speaking rates as reported by Byrd (1994), it is likely that these data reflect the interaction between speaker gender and dialect region that she found in the TIMIT database: most of our speakers were from the South Midland region, which Byrd found to have the slowest speaking rate for males but an average speaking rate for females.

Taken together, the present data indicate that the global talker characteristics that we examined were not strong correlates of variability in overall speech intelligibility. We found a significant difference in overall intelligibility across the male and female talkers in our database, which suggests that a higher fundamental frequency is associated with higher overall intelligibility; however, this was not the case for the separate groups of male and female talkers. We also found a tendency for the range in fundamental frequency to correlate positively with speech intelligibility. However, there was no clear relationship between overall speaking rate and speech intelligibility. Other global, source-related characteristics might affect overall intelligibility of normal speech, such as spectral tilt and other details of the glottal waveform. However, in view of our generally weak findings regarding the simplest global talker characteristics, we turned our attention to more fine-grained talker characteristics that are indicators of specific pronunciation characteristics at the segmental level.

Fine-Grained Acoustic-Phonetic Talker Characteristics

Vowel Space Characteristics

We began our investigation of fine-grained acoustic-phonetic talker characteristics with an examination of vowel spaces. Vowel centralization is a typical feature of casual, or reduced speech (Picheny et al., 1986; Lindblom, 1990; Moon and Lindblom, 1994; Byrd, 1994). Additionally, vowel space expansion has been shown to correlate with speech intelligibility. For example, Bond and Moore (1994) found more peripheral vowel category locations in an F1 by F2 space for a higher-intelligibility talker relative to a lower-intelligibility talker. In a study of vowel production by deaf adolescents, Monsen (1976) found a significant positive correlation between range in F2 and intelligibility. Both of these studies lead us to hypothesize that in our multi-talker sentence database we would find a positive correlation between overall intelligibility and measures of vowel space expansion. Specifically, we predicted that relatively expanded vowel spaces would be associated with enhanced speech intelligibility scores.

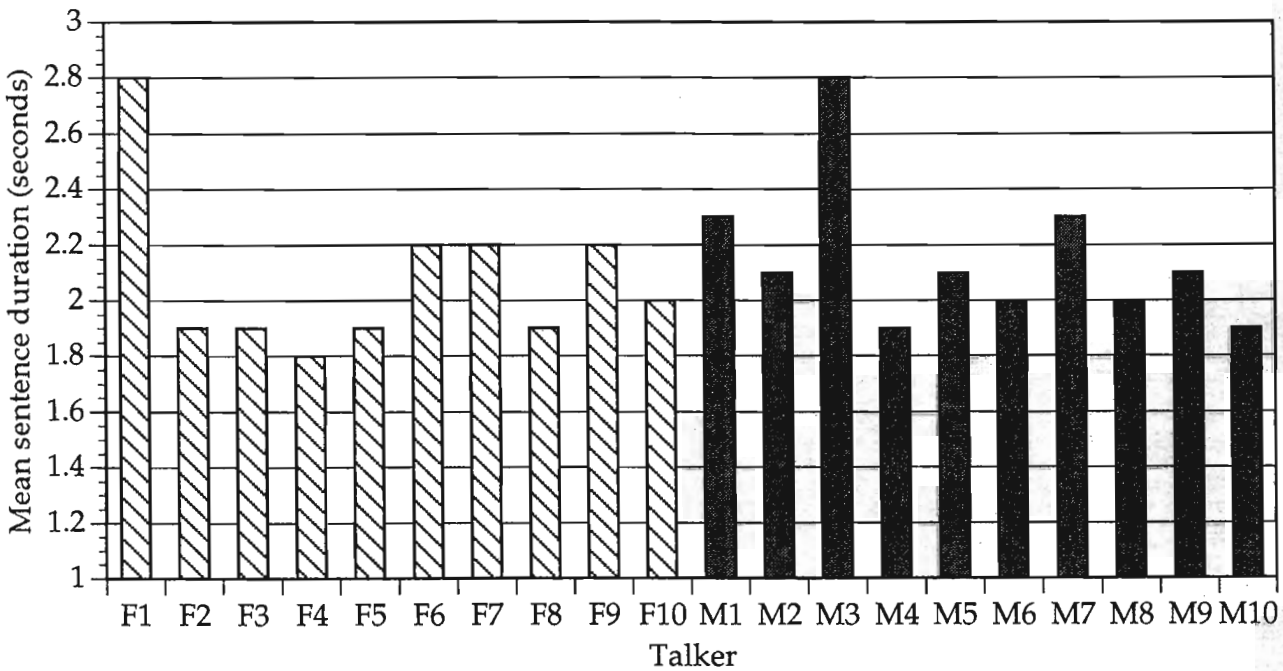


Figure 3. Mean sentence duration in seconds for each of the 10 female talkers (columns filled with slanted lines) and 10 male talkers (shaded columns).

In order to measure each talker's vowel space, we selected six occurrences of the three peripheral vowels, /i, a, o/, from the sentence materials in the database. (The point vowel /u/ was avoided due to excessive allophonic variation for this vowel in General American English). All of the words containing the target vowels were content words, and none was the final keyword in the sentence. Table II lists the subset of 18 sentences containing the words with the target vowels from which the vowel space measurements were taken.

Insert Table II about here

The first and second formants were measured from each of the 18 target vowels as produced by each of the 20 talkers. All formant measurements were made using the Entropics WAVES+ software package on a SUN workstation. Both LPC spectra (calculated from a 25 ms Hanning window) and spectrograms were used to determine the location of the first two formant frequencies at the vowel steady-state. These F1 and F2 measurements were then converted to the perceptually motivated mel scale (Fant, 1973). (The exact equation for converting frequencies from Hertz to mels is $M = (1000/\log 2)\log((F/1000)+1)$, where M and F are the frequencies in mels and Hertz, respectively.) Each talker's vowel space was then represented by the locations of the 18 individual vowel tokens in an F1 by F2 space. In all of the following analyses of the relations between vowel space characteristics and speech intelligibility, we used each talker's average intelligibility score across the 18 sentences, given in Table II, that formed the subset of sentences with the words that contained the target vowels (see Table III below). Across all 20 talkers, the overall intelligibility scores for the total set of 100 sentences and for the subset of 18 sentence were significantly correlated (Spearman rho = +0.629, $p = 0.006$), thus this subset of 18 sentences was a good indicator of the talkers' overall intelligibility scores.

Insert Figure 4 and Table III about here

The first measure that we used to assess the relationship between vowel space and overall speech intelligibility was the Euclidian area covered by the triangle defined by the mean of each vowel category. Here we hypothesized that the greater the triangular area, the higher the overall intelligibility. Figure 4a shows the vowel triangles for the highest intelligibility talker (Talker F7) and the lowest intelligibility talker (Talker M2). It is clear from Figure 4a that the vowel triangle for Talker F7 covers a greater area within this space than the vowel triangle for Talker M2. However, across all 20 talkers we failed to find a positive correlation between triangular vowel space area and speech intelligibility scores (see Table III for each individual talker's vowel space area). One problem with triangular vowel space area as a measure of vowel category differentiation is that the points used to calculate this measure are the category averages, and these may not be representative of the individual vowel tokens actually produced by the talker. For this reason, we devised a different measure of vowel space expansion that took into account the location of each individual vowel token, and then reanalyzed the data.

Figure 4b shows each vowel token's distance from a central point in the talker's vowel space for the highest intelligibility talker (Talker F7) and the lowest intelligibility talker (Talker M2). A measure of each talker's "vowel space dispersion" was calculated as the mean of these distances for each talker. This measure thus provided an indication of the overall expansion, or compactness, of the set of individual vowel tokens from each talker (see Table III for each individual talker's vowel space dispersion measure).

Table II

Subset of 18 sentences containing the words with the target vowels from which the vowel space measurements were taken, with the IPA phonemic transcription for the target word. All 5 keywords are underlined, with the word with the target vowel in italics.

/i/:	
1. It's *easy* to tell the depth of a well.	/izi/
2. The fruit *peel* was cut in thick slices.	/pil/
3. Adding fast *leads* to wrong sums.	/lidz/
4. This is a grand *season* for hikes on the road.	/sizIn/
5. The walled town was *seized* without a fight.	/sizd/
6. The *meal* was cooked before the bell rang.	/mil/
/a/:	
7. A *pot* of tea helps to pass the evening.	/pat/
8. A *rod* is used to catch pink salmon.	/rad/
9. The wide road shimmered in the *hot* sun.	/hat/
10. The show was a *flop* from the very start.	/flap/
11. The hogs were fed *chopped* corn and garbage.	/tʃapt/
12. A large size in *stockings* is hard to sell.	/stakINz/
/o/:	
13. The horn of the car *woke* the sleeping cop.	/wok/
14. Bail the *boat* to stop it from sinking.	/bot/
15. Mend the *coat* before you go out.	/kot/
16. Hoist the *load* to your left shoulder.	/lod/
17. The dune *rose* from the edge of the water.	/roz/
18. The young girl gave *no* clear response.	/no/

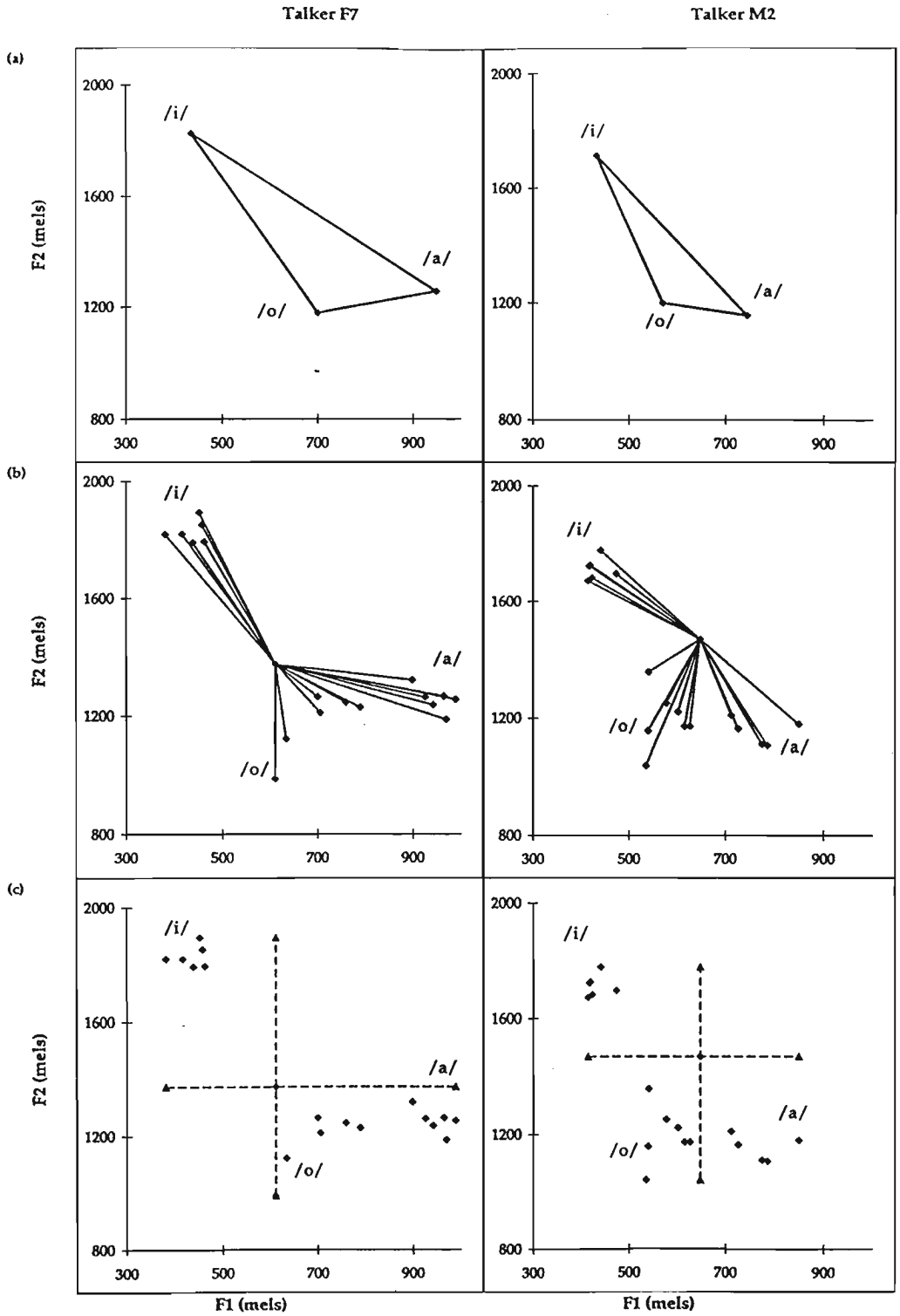


Figure 4. Vowel space characteristics for a high-intelligibility talker (Talker F7) and a low-intelligibility talker (Talker M2): (a) vowel space area, (b) vowel space dispersion, (c) range in F1 and F2.

Table III

Intelligibility scores across the 18 sentences used for the vowel space meaurements, vowel space area, dispersion, F1 range, F2 range, withing category clustering, F2-F1 distance for /i/ and /a/ for each individual talker.

Talker	Intelligibility (18 sentences)	Vowel Space Area ($mels^2$)	Vowel Space Dispersion (mels)	F1 Range (mels)	F2 Range (mels)	Category Clustering (mels)	F2-F1 /i/ (mels)	F2-F1 /a/ (mels)
F1	93.3	82747.76	349.23	649.72	802.95	66.899	1426.44	311.19
F2	92.8	40844.95	301.91	569.19	746.18	75.077	1396.64	339.11
F3	91.1	81688.80	327.33	518.35	1168.76	90.939	1314.26	412.39
F4	85.0	49686.44	268.21	545.66	932.08	110.827	1219.59	422.63
F5	91.1	85160.25	311.06	604.91	1172.23	107.85	1357.83	496.94
F6	91.7	69203.79	321.54	542.88	852.91	67.482	1370.43	333.86
F7	92.8	98726.79	346.85	607.76	904.66	62.253	1394.58	307.04
F8	92.8	55770.36	291.36	572.61	757.56	77.757	1351.63	405.06
F9	88.3	28993.41	251.30	564.30	651.23	73.608	1243.95	339.65
F10	90.6	61950.01	252.13	472.95	672.00	66.179	1253.27	450.47
M1	91.1	61092.87	285.57	470.40	844.43	61.688	1230.42	355.51
M2	78.3	41005.79	272.45	435.03	737.49	65.893	1282.32	413.46
M3	82.2	114352.73	360.09	498.13	1053.08	94.095	1238.78	393.10
M4	81.5	73394.01	278.47	456.37	745.809	65.757	1211.41	529.38
M5	85.0	13531.00	250.13	476.60	636.17	55.574	1268.43	375.96
M6	86.7	72398.30	280.60	475.77	663.73	47.559	1250.89	443.88
M7	88.3	35205.43	273.67	408.63	811.51	45.863	1204.81	482.01
M8	90.6	49982.49	262.61	430.11	751.19	99.5	1080.55	382.81
M9	86.7	63413.41	263.44	453.04	756.13	66.161	1177.70	387.12
M10	85.0	79670.34	309.08	575.71	878.97	72.69	1343.15	399.74

The measures of vowel space area and vowel space dispersion were highly correlated (Spearman rho = +0.782, $p < 0.001$), however, the correlation was not perfect indicating that each measure captures a slightly different aspect of the talkers' vowel production characteristics. With respect to the correlation between vowel space dispersion and intelligibility, we found a moderate, positive rank order correlation (Spearman rho = +0.431, p=0.060) across all 20 talkers, and this correlation increased when only the 10 highest intelligibility talkers were included in the analysis (Spearman rho = +0.698, p=0.036). Thus, using a measure of vowel space dispersion, the data showed that higher overall speech intelligibility is associated with a more expanded vowel space, particularly for the talkers in the top half of the distribution of intelligibility scores.

Based on the finding that overall vowel space dispersion and speech intelligibility were correlated, we then investigated which of the two dimensions, F1 or F2, in the vowel space representations was more responsible for this correlation. In his study of the vowel productions of deaf adolescents, Monsen (1976) found a stronger positive correlation between range in F2 and intelligibility (r=+0.74) than he did for range in F1 and intelligibility (r=+0.45). As Monsen notes, these correlations do not suggest that range in F2 is more important for normal speech intelligibility than range in F1, rather these correlations arise from the fact that the vowels of these deaf subjects occupy a more normal range in F1 than in F2. For the purposes of our investigation of variability in normal speech, Monsen's finding simply indicated the usefulness of investigating range in F1 and F2 as separate dimensions that might correlate with overall intelligibility.

Accordingly, we measured each talker's range in F1 and F2 as the difference between the maximum and minimum values on each of these dimensions. Figure 4c shows the F1 and F2 range measurements for the highest intelligibility talker (Talker F7) and the lowest intelligibility talker (Talker M2). (See Table III for each individual talker's range in F1 and F2). Across all 20 talkers, we found a significant positive rank order correlation between range in F1 and intelligibility (Spearman rho = +0.531, p=0.020), but we failed to find a significant rank order correlation between range in F2 and intelligibility (Spearman rho = +0.239, p=0.300). This correlation of F1 range and intelligibility was strengthened when only the top 10 talkers were included in the analysis (Spearman rho = +0.817, p=0.014). Thus, it appears that the area covered in F1 was a better correlate of overall intelligibility than the area covered in F2. This finding is not surprising in view of the fact that the English vowel system has several vowel height distinctions (of which F1 frequency is an important acoustic correlate), whereas there are many fewer distinctions along the front-back dimension (of which F2 frequency is the primary acoustic correlate). It may be that in order for the numerous English vowels to be well distinguished, a wide range in F1 (vowel height) is advantageous, whereas less precision can be more easily tolerated in the F2 (front-back) dimension.

The vowel space measures that we have reported so far have established relations between relative vowel space expansion and overall speech intelligibility, particularly for talkers in the top half of the intelligibility score distribution. An additional measure of vowel articulation that might be expected to correlate with intelligibility is the relative compactness of individual vowel categories. We might expect that the more tightly clustered categories enhance intelligibility since they are less likely to lead to inter-category confusion. As a measure of tightness of within-category clustering, we first calculated the mean of the distances of each individual token from the category mean, as we did for our measure of overall vowel space dispersion. Then a single measure for each talker was calculated as the mean within-category dispersion across all three vowel categories (see Table III for these values for each talker). However, analysis of the results showed that across all 20 talkers, as well as for only the 10 highest intelligibility talkers, there was no correlation between within-category dispersion and intelligibility. Thus, tightness of within-category clustering per se was not a good correlate of overall intelligibility, suggesting that overall

vowel space expansion, rather than within-category compactness, is associated with overall speech intelligibility.

The final measure of vowel space that we examined as a possible correlate of speech intelligibility was the acoustic-phonetic implementation of the point vowels /i/ and /a/. Each of these two vowel categories defines an extreme point in the American English general vowel space. In the acoustic domain, they each display extreme F2-F1 distances: /i/ is characterized by a wide separation between the first two formant frequencies, whereas /a/ is characterized by very close F1 and F2 frequencies. Thus, the F2-F1 distance for these point vowels provided an indication of the extreme locations in the F1 by F2 space for these vowels (Gerstman, 1968). Accordingly, we hypothesized that the F2-F1 distance for /i/ would be positively correlated with overall intelligibility, and that the F2-F1 distance for /a/ would be negatively correlated with overall intelligibility. Indeed, across all 20 talkers, we found a positive rank order correlation between F2-F1 distance for /i/ and overall intelligibility (Spearman rho = +0.601, p=0.009), and a negative rank order correlation between F2-F1 distance for /a/ and overall intelligibility (Spearman rho = -0.509, p=0.027). (See Table III for these values for each talker). When only the 10 highest intelligibility talkers were included in the analysis, these correlations were strengthened further (Spearman rho = +0.866, p=0.009 and Spearman rho = -0.673, p=0.043, for /i/ and /a/ respectively.) Thus, relatively high overall speech intelligibility is associated with more extreme vowels as measured by the precision of individual vowel category realization, as well as by overall vowel space expansion for a given talker.

In summary, the general pattern that emerged from these measures of the acoustic-phonetic vowel characteristics as correlates of overall intelligibility was that talkers with more reduced vowel spaces tended to have lower overall speech intelligibility scores. The measures of vowel space reduction that were shown to correlate with overall speech intelligibility were overall vowel space dispersion, particularly range covered in the F1 dimension, and the extreme locations in the F1 by F2 space of the point vowels /i/ and /a/ as measured by F2-F1 distance. The analyses also showed that the correlations between vowel space reduction and overall intelligibility were stronger for talkers in the top half of the distribution of intelligibility scores, suggesting a greater degree of variability for talkers with lower intelligibility scores that is not accounted for by these measures of a talker's vowel space.

Acoustic-Phonetic Correlates of Consistent Listener Errors

Another strategy we used for investigating the correlation between fine-grained acoustic-phonetic characteristics of a talker's speech and overall intelligibility involved analyses of the specific portions of sentences that showed consistent listener transcription errors. With this approach we hoped to identify specific pronunciation patterns that resulted in the observed listener errors. These analyses differed from the methods used in the analysis of vowel spaces because here we focused on specific cases where there were known listener errors, rather than on more general statistical indicators of overall phonetic reduction. In particular, in our database we found two specific cases of consistent listener error that revealed the importance of highly precise inter-segmental timing for speech intelligibility (see also Neel, 1995).

Segment Deletion

The first case of consistent listener error occurred in the sentence, "The walled town was seized without a fight." The overall intelligibility of this sentence across all 20 talkers was 60% correct, with 94% of the listener transcription errors occurring for the phrase, "walled town." Of the listener errors on this portion of the sentence, 82% involved omitting the word final /d/ in "walled." In order to determine what specific talker-related acoustic characteristics might lead to this common listener error, we measured the

durations of various portions of the acoustic waveform from this phrase, and then correlated these measurements with the rate of /d/ detection for each talker.

We began by measuring the total vowel-to-vowel duration, that is, the portion of the waveform that corresponds to the talker's /dt/ articulation between the /al/ of "wall" and the /aU/ of "town." This portion of the acoustic signal was measured from the point at which there was a marked decrease in amplitude and change in waveform shape as the preceding vowel-sonorant sequence (the /al/ from "wall") ended, until the onset of periodicity for the following vowel (the /aU/ from "town"). In almost all cases, this portion consisted of a single /d/-like closure portion and a single /t/-like release portion: most talkers (18/20) did not release the /d/ and then form a second closure for the /t/. Figure 5 shows waveforms of this portion of the sentence for two talkers, with vertical cursors demarcating the salient acoustic boundaries.

Insert Figure 5 about here

Across the group of 20 talkers, we found a significant positive rank order correlation between the vowel-to-vowel duration and rate of /d/ detection (Spearman rho = +0.713, p=0.002). Based on this finding, we then looked at the rate of /d/ detection in relation to the separate durations of the /d/ closure portion and of the /t/ release portion, which together comprised the vowel-to-vowel portion. Here we found a significant positive correlation with /d/ closure duration (Spearman rho = +.641, p=.005), but no correlation with /t/ release duration. The /d/ closure portion generally consisted of a period with very low amplitude, low frequency vibration, followed by a silent portion. Accordingly, we then examined the correlation between rate of /d/ detection and the separate durations of each of these portions of the total closure duration. A highly significant positive correlation was found between rate of /d/ detection and the duration of voicing during the /d/ closure (Spearman rho = +0.755, p<0.001), whereas no correlation was found between the duration of the silent portion of the /d/ closure and rate of /d/ detection. This correlation suggests that the duration of voicing during closure, in an absolute sense, is a reliable acoustic cue to the presence of a voiced consonant in this phonetic environment. However, an extremely strong (and highly significant) rank order correlation was found between the rate of /d/ detection and the duration of the voicing during the /d/ closure relative to the duration of the preceding vowel-sonorant sequence, /wal/ (Spearman rho =+.810, p=.0004). In other words, listeners appeared to rely heavily on relative timing between the duration of voicing during the /d/ closure and the overall rate of speech, as determined by the duration of the preceding syllable portion, in detecting the presence or absence of a segment. This finding is consistent with studies on rate-dependent processing in phonetic perception that have shown that listeners adjust to overall rate of speech in the identification of phonetic segments (e.g. Miller 1981), and that relative timing between segments can play a crucial role in segment identification (Port, 1981; Port and Dalby, 1982).

Figure 5 contrasts two talkers with varying amounts of this low frequency voicing during the /d/ closure relative to the preceding /wal/ portion of the waveform. Talker M1 had a considerably longer relative duration of voicing during the /d/ closure than talker M9, and consequently all of the listeners for Talker M1 detected the presence of the /d/, whereas the only 1 of the 10 listeners for Talker M9 detected the /d/. In other words, talkers who did not produce sufficient voicing during the /d/ closure as determined by the overall rate of speech (e.g. Talker M9 in Figure 5) were likely to be mis-heard, even though the syntactic structure of the phrase should have lead listeners to expect " *walled* town" rather than "*wall* town." For the purposes of the present investigation, this particular case of consistent listener-error

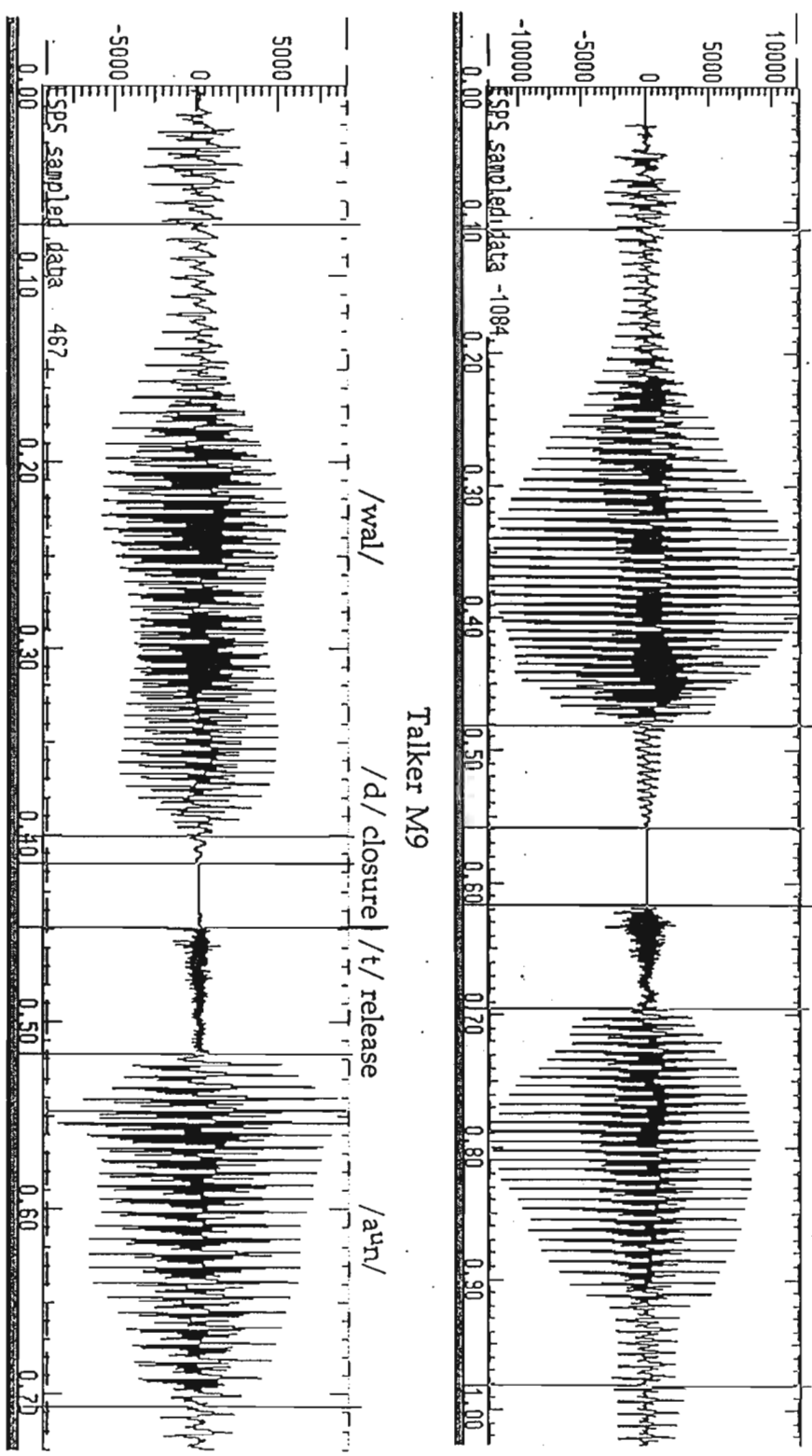


Figure 5. Waveforms of the sentence portion, "walled town," as produced by Talker M1, who had a relatively long duration of voicing during of the /d/ closure, and Talker M9, who had a very short duration of voicing during the /d/ closure.

revealed just how some of the observed variability in speech intelligibility scores for these sentences can be traced directly to specific pronunciation characteristics of the talker. Furthermore, this case indicates the importance of articulatory precision in the realization of gestural timing relations for speech intelligibility.

Syllable Affiliation

The second case of a consistent listener error occurred in the sentence, "The play seems dull and quite stupid." The overall intelligibility of this sentence across all 20 talkers was 75% correct, with 70% of the listener transcription errors occurring for the phrase, "play seems." Of the perceptual errors on this portion of the sentence, 95% involved mis-syllabification of the word initial /s/, resulting in "place seems". In this case, we measured the duration of the /s/ (marked by the high frequency, high amplitude turbulent waveform) and of the preceding and following syllables (/plej/ and /simz/ respectively). Figure 6 shows waveforms of this portion of the sentence for Talker F6 and Talker F1 with vertical cursors marking these three segments. We then examined the correlation of these durations and the rate of correct transcription of "play." We expected to find a correlation between /s/ duration relative to the durations of surrounding syllables and rate of correct transcription. Indeed, results showed a significant negative correlation between rate of "play seems" transcription and /s/ duration as a proportion of the preceding syllable, /plej/, duration (Spearman rho =-0.631, p=0.006). We also found a tendency for the correct transcription rate to correlate with /s/ duration as a proportion of the following syllable, /imz/, duration (Spearman rho =-0.432, p=0.060). In other words, the shorter the /s/ relative to the surrounding syllables, the more likely it was to be correctly syllabified by the listener as onset of the following syllable, rather than as both coda of the preceding word and the onset of the following word. In Figure 6, this may be seen by the shorter relative durations of the /s/ for Talker F6, whose /s/ was correctly syllabified by all 10 listeners, as opposed to the relatively longer /s/ for Talker F1, whose /s/ was correctly syllabified by only 3 of the 10 listeners. Thus, in this case, as in the case of segment deletion discussed above, the listeners drew on global information about the speaking rate of the talker in perceiving the placement of the word boundary. The talker's precision in inter-segmental timing had a direct effect on the listener's interpretation of the speech signal.

Insert Figure 6 about here

Furthermore, in this case, there was a gender-related factor in the timing relationship between the medial /s/ and the surrounding syllables. In general, the duration of the /s/ relative to the preceding and following syllables was shorter for the female talkers than for the male talkers. Consequently, the female talkers' renditions of this phrase were more often correctly transcribed: 7 of the 10 female talkers had no errors of this type, whereas 6 of the 10 males had this error for at least 30% of the listeners. Thus, in this case, the female talkers as a group were more precise with respect to controlling this timing relationship than the group of male talkers. Although this case is not a matter of phonological reduction (in fact, the correct form is shorter in duration), this example does demonstrate that the gender-based difference in overall speech intelligibility that we observed in our database may be due to the use of more precise articulations by our female talkers. Moreover, both this case of syllable affiliation and the previous case of segment deletion indicate why global talker-related characteristics, such as overall speech rate, may not be good candidates for the primary determiners of talker intelligibility: finer acoustic-phonetic details of speech timing and the precision of specific articulatory events "propagate up" to higher levels of processing during speech perception to modulate and control overall speech intelligibility in sentences.

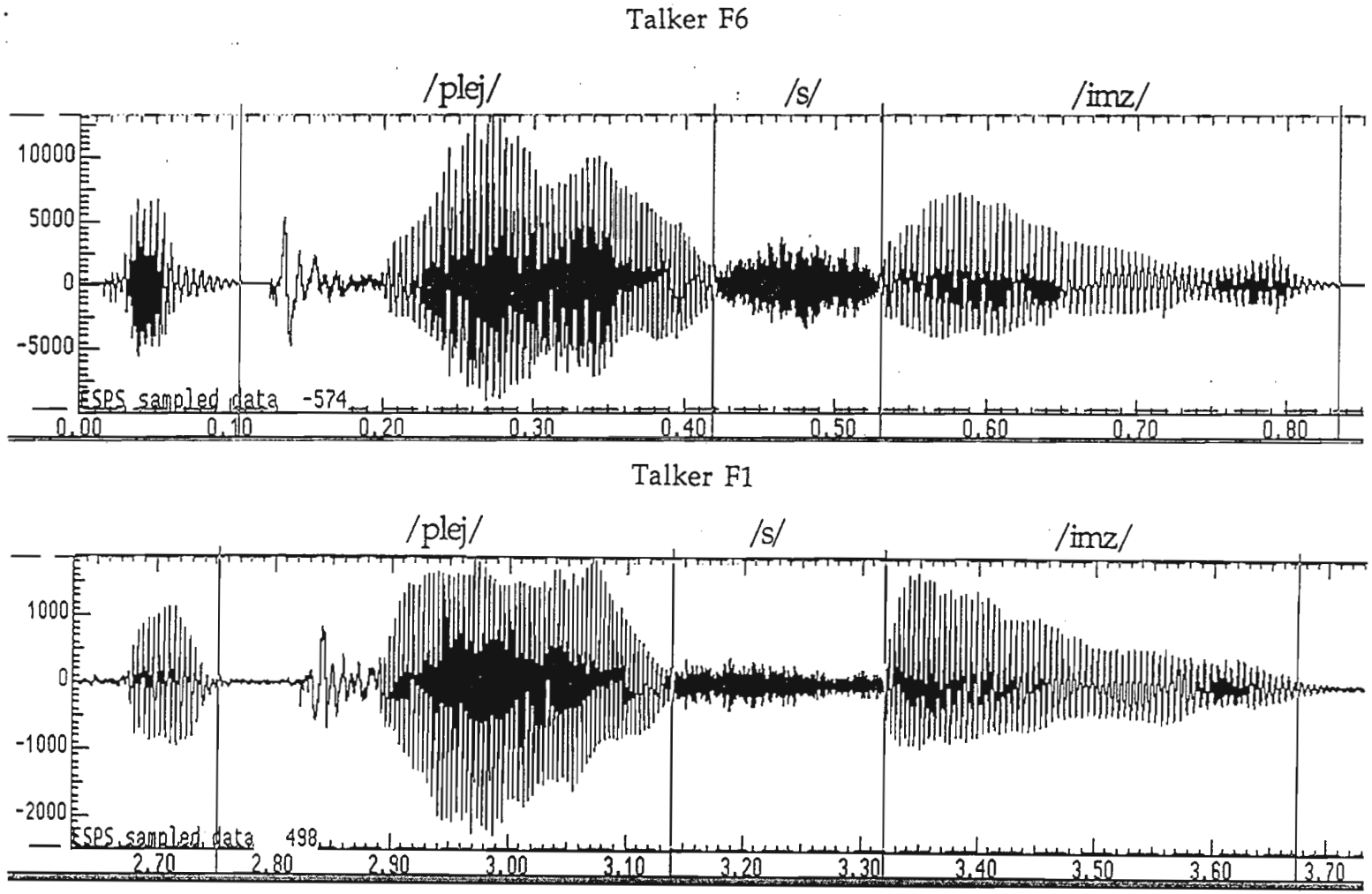


Figure 6. Waveforms of the sentence portion, "play seems," as produced by Talker F6, who had a relatively short /s/ duration relative to the durations of the /plej/ and /imz/ portions, and Talker F1, who had a relatively long /s/ duration relative to the durations of the /plej/ and /imz/ portions.

General Summary and Discussion

The overall goal of this investigation was to identify some of the talker-related acoustic-phonetic correlates of speech intelligibility. Specifically, we asked, "What makes one talker more intelligible than another?" The results of this study showed that global talker characteristics such as overall speaking rate and mean fundamental frequency did not correlate strongly with speech intelligibility scores. In contrast, we observed a tendency for a wider fundamental frequency range to be associated with higher overall intelligibility, and we found a significant gender-related difference in intelligibility such that the female talkers in our database were generally more intelligible than the male talkers. We also found strong evidence for a negative correlation between degree of vowel space reduction and overall intelligibility, suggesting that a talker's vowel space is a good indicator of overall speech intelligibility. Specifically, we found that talkers who produce vowels that are widely dispersed in the phonetic vowel space, particularly along the F1 dimension, have relatively high overall intelligibility scores. Finally, by examining two cases of common listener perceptual errors, segment deletion and mis-syllabification, we found that talkers with highly precise articulations at the fine-grained acoustic-phonetic level were less likely to be mis-heard than talkers who showed less articulatory precision.

These findings suggest that for normal talkers, although global characteristics do appear to have some bearing on overall intelligibility, there are substantially stronger correlations between fine-grained changes in articulation and speech intelligibility. In response to the question we posed at the start of this investigation, the present results suggest that highly intelligible talkers are those with a high degree of articulatory precision in producing segmental phonetic contrasts and a low degree of phonetic reduction in their speech. Based on these findings we can construct a profile of a highly intelligible talker: such a talker would be a female who produces sentences with a relatively wide range in fundamental frequency, employs a relatively expanded vowel space that covers a broad range in F1, precisely articulates her point vowels, and has a high precision of inter-segmental timing.

This characterization of a highly intelligible talker has several broader implications for our understanding of acoustic-phonetic variability and its effects on overall speech intelligibility. First, these findings suggest that a substantial portion of the observed variability in overall speech intelligibility can be traced directly to talker-specific characteristics. As we noted in the introduction to this paper, the speech intelligibility scores in our database reflect both listener- and sentence-related characteristics as well as talker-related characteristics. Nevertheless, by focusing exclusively on talker-related characteristics, we were able to identify several correlates of variability in speech intelligibility.

A second implication of our findings deals with the role that talker-related characteristics play in situations that require "clear" speech. In a review of speech intelligibility tests used with disordered speakers Weismer and Martin (1992) noted that indices of intelligibility deficits are considerably more useful when they include an explanatory component, in terms of the acoustic-phonetic bases of these deficits, that can serve as a guide for the remediation of such deficits. Knowledge of how speech varies across normal talkers, and how these variations affect speech intelligibility, might help direct attention to the crucial aspects of speech production for special populations, such as the hearing-impaired, and second language learners. Similarly, such fundamental knowledge about the production of normal speech may be very useful for the development of the next generation of speech output devices. For example, speech synthesizers and speech synthesis-by-rule systems could be designed to focus and emphasize the talker-characteristics that result in highly intelligible natural speech. Additionally, by knowing how natural speech varies across talkers and how these specific variations affect overall intelligibility, speech recognition

systems might be able to achieve higher levels of performance over a much wider range of individual talkers and operational environments.

Finally, by establishing a direct link between overall speech intelligibility and some of the fine-grained acoustic-phonetic variations that exist across talkers, the results of this investigation add to the growing body of research demonstrating the important role that talker-specific attributes play in speech perception and spoken language processing. We believe it is now possible to provide a principled explanation for why some talkers are more intelligible than others and to specify the attributes of their speech with greater precision than has been possible in the past. Part of the success of this approach lies in having a large digital database of spoken sentences along with speech intelligibility scores for each sentence. Thus, detailed acoustic-phonetic measures of the speech signal can be related directly to listeners' perceptual responses.

References

Black, J.W. (1957). Multiple-choice intelligibility tests. *Journal of Speech and Hearing Disorders*, **22**, 213-235.

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

Byrd, D. (1994). Relations of sex and dialect to reduction, *Speech Communication*, **15**, 39-54.

Gerstman, L.J. (1968). Classification of self-normalized vowels, *IEEE Transactions on Audio and Electroacoustics,* **AU-16**, 78-80.

Hanson, H.M. (1995). Glottal characteristics of female speakers- Acoustic, physiological, and perceptual correlates, *Journal of the Acoustical Society of America,* **97**(2), 3422.

Hirsh, I.J., Reynolds, E.G. and Joseph, M. (1954). Intelligibility of different speech materials, *Journal of the Acoustical Society of America,* **26**, 530-538.

Hood, J.D and Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility, *Audiology* , **19**, 434-455.

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

Karl, J. and Pisoni, D.B. (1994). The role of talker-specific information in memory for spoken sentences, *Journal of the Acoustical Society of America* , **95**(2), 2873.

Keating, P.A., Byrd, D., Flemming, E. and Todaka, Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English, *Speech Communication*, **14**, 131-142.

Klatt, D. and Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America*, **87**, 820-857.

Krause, J.C. and Braida, L.D. (1995). The effects of speaking rate on the intelligibility of speech for various speaking modes, *Journal of the Acoustical Society of America*, **98**(2), 2982.

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

Lamel, L., Kassel, R. and Seneff, S. (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus, *Proceedings DARPA Speech Recognition Workshop,* February 1986, 100-109.

Laver, J. and Trudgill, P. (1979). Phonetic and linguistic markers in speech, in *Social markers in speech* ed. by K. R. Scherer and H. Giles (Cambridge University Press, Cambridge)., 1-32.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H & H theory, in *Speech Production and Speech Modeling*, ed. by W. J. Hardcastle and A. Marchal. Kluwer Academic Publishers, Dordrecht. 403-439.

Luce, P.A. and Carrell, T.D. (1981). Creating and editing waveforms using WAVES, *Research in Speech Perception Progress Report No. 7*. Indiana University Speech Research Laboratory, Bloomington.

Miller, J.L. (1981). Effects of speaking rate on segmental distinctions, in *Perspectives on the study of speech,* ed. by P. D Eimas and J. L. Miller. Lawrence Erlbaum, Hillsdale. 39-74.

Monsen, R.B. (1976). Normal and reduced phonological space: the productions of English vowels by deaf adolescents, *Journal of Phonetics*, **4**, 189-198.

Moon, .S.J. and Lindblom, B. (1994). Interaction between duration, context and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, **96**, 40-55.

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

Neel, A.T. (1995). Intelligibility of normal speakers: Error analysis, *Journal of the Acoustical Society of America,* **98**(2), 2982.

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

Nygaard, L.C., Sommers, M.S. and Pisoni, D.B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory, *Perception and Psychophysics,* **57**, 989-1001.

Pallett, D. (1990). Speech corpora and performance assessment in the DARPA SLS program, *ICSLP 90 Proceedings.*, 24.3.1-24.3.4.

Palmeri, T.J., Goldinger, S.D. and Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words, *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**, 1-20.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28**, 96-103.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, **29**, 434-446.

Picheny, M.A., Durlach, N.I. and Braida, L.D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to difference in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, **32**, 600-603.

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.

Port, R.F. (1981). Linguistic timing factors in combination, *Journal of the Acoustical Society of America,* **69**, 262-274.

Port, R.F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English, *Perception and Psychophysics,* **32**, 141-152.

Runyon, R.P. and Haber, A. (1991). *Fundamentals of Behavioral Statistics* (McGraw-Hill)., 201-205.

Sommers, M.S., Nygaard, L.C. and Pisoni, D.B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude, *Journal of the Acoustical Society of America,* **96**, 1314-1324.

Tielen, M.T.J. (1992). Male and female speech: An experimental study of sex-related voice and pronunciation characteristics, Doctoral dissertation, University of Amsterdam.

Uchanski, R.M., Choi, S., Braida, L.D., Reed, C.M. and Durlach, N.I. (in press). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research* .

Weismer, G. and Martin, R.E. (1992). Acoustic and perceptual approaches to the study of intelligibility, in *Intelligibility in speech disorders: Theory, measurement and management,* ed. by R. D. Kent. John Benjamins, Amsterdam/Philadelphia. 67-118.

Zue, V., Seneff, S. and Glass, J. (1990). Speech database development at MIT: TIMIT and beyond, *Speech Communication,* **9**, 351-356.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Talker-Specific Perceptual Learning in Spoken Word Recognition: Preliminary Findings and Theoretical Implications[1]

Lynne C. Nygaard[2] and David B. Pisoni[3]

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

[1] This research was supported by NIH NIDCD Research Grant DC-00111 to Indiana University, Bloomington, IN.
[2] Department of Psychology, Emory University, Atlanta, GA.
[3] Also, DeVault Otologic Research Laboratory, Departtment of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Abstract

Three experiments investigated the effects of perceptual learning of talker identity on the recognition of spoken words and sentences. Listeners were trained to learn a set of ten talkers' voices and were then given an intelligibility test to assess the influence of learning the voices on the processing of the linguistic content of speech. In the first experiment, listeners learned novel voices from isolated words and then were tested with isolated words mixed in noise at several signal-to-noise ratios. The results showed transfer of training from the talker identification task to the word intelligibility task. Listeners who were given words produced by familiar talkers at test showed better identification performance than listeners who were given unfamiliar talkers at test. In the second experiment, listeners learned novel voices from sentence-length utterances and then were presented with isolated words. The results showed that learning a talker's voice from sentences did not generalize very well to isolated words suggesting that listeners were attending to different talker-specific information when learning voices from sentences than when learning voices from isolated words. In the third experiment, listeners learned voices from sentence-length utterances and then were given sentence-length utterances produced by familiar and unfamiliar talkers to transcribe at test. The results showed that perceptual learning of novel voices from sentence-length utterances improved speech intelligibility for words in sentences. Generalization and transfer from voice learning to linguistic processing was found to be sensitive to the talker-specific information available during learning and test. These findings demonstrate that increased sensitivity to talker-specific information in the speech signal through perceptual learning affects the perception of spoken language. In learning to recognize a novel voice, listeners encode perceptual details and processing operations that transfer to other tasks which require analysis of the phonetic content of the same acoustic signal. The results of these experiments suggest that current models of speech perception and spoken language processing need to include a role for talker-specific perceptual learning in the perception of speech and language in order to account for the observed interactions between these two sources of information that are carried by the same acoustic signal.

Talker-Specific Perceptual Learning in Spoken Word Recognition: Preliminary Findings and Theoretical Implications

During everyday conversation, listeners effortlessly understand talkers with a wide variety of individual vocal characteristics and styles. It is only when we encounter an unfamiliar talker with an unusual dialect or accent that we become consciously aware that we have to adjust to the idiosyncratic vocal attributes of a novel talker. Presumably, this adjustment involves a period of perceptual adaptation in which listeners learn to differentiate the unique properties of each talker's speech patterns from the underlying intended linguistic message. Listening to speech produced by talkers of different dialects and accents is an extreme example of what occurs routinely as we encounter unfamiliar talkers. The purpose of the present investigation was to study this process of perceptual learning and adaptation to individual talkers and to determine how sensitivity to talker identity affects the intelligibility of the linguistic aspects of speech, specifically the recognition of spoken words in isolation and in sentence contexts.

The Abstractionist Approach

Traditionally, the perception of linguistic content of speech -- the words, phrases, and sentences, of an utterance has been studied separately from perception of talker identity. Research on the perception of the linguistic aspects of spoken language has considered variation in the acoustic realization of linguistic components due to differences in individual talkers as a source of noise that serves to obscure the underlying abstract symbolic linguistic message. Variability is considered a perceptual problem that listeners must solve if they are to recover the linguistic constituents that carry meaning (Shankweiler, Strange, & Verbrugge, 1977). The proposed solution to this problem of talker variability is a perceptual normalization process in which talker-specific acoustic-phonetic properties are evaluated relative to prototypical mental representations (Joos, 1948; Ladefoged & Broadbent, 1957, Summerfield & Haggard, 1973). Variation is assumed to be stripped away to arrive at canonical representations that underlie further linguistic analysis. Implicit in this view of normalization is the assumption that the end product of perception is a series of abstract, symbolic, idealized, linguistic units (Halle, 1985; Joos, 1948; Kuhl; 1991, 1992).

This abstractionist approach to the perception of spoken language with its emphasis on context-free processing units falls short of providing a satisfactory explanation for the relationship between the processing of linguistic content and the analysis of a talker's voice. Although the speech signal carries a considerable amount of personal information about the talker along with the linguistic content into the communicative setting (Ladefoged & Broadbent, 1957; Laver, 1989; Laver & Trudgill, 1979), little, if any, role for talker information has been assumed in current theoretical accounts of the perception of speech or spoken language processing. A separate body of research has addressed the perception and identification of talker identity viewing the speech signal as simply a carrier of talker information (e.g., Legge, Grossmann, & Pieper, 1984; Van Lancker, Kreiman, & Emmorey, 1985). This explicit dissociation of research involving linguistic processing, on the one hand, and voice perception, on the other hand, reflects an implicit theoretical separation. Talker identification and perception are assumed to involve a distinct set of perceptual mechanisms which operate on attributes of the acoustic speech signal that are separate and autonomous from the attributes that underlie spoken word recognition and comprehension of the linguistic message.

An alternative to the abstractionist approach to speech perception and spoken language recognition suggests that the traditional view of perceptual normalization and its long-standing emphasis on the search

for abstract, canonical linguistic units as the endpoint of perception may need to be reconsidered or abandoned entirely. Over the last few years, a number of researchers have demonstrated that stimulus variability is a rich source of information that is encoded and stored in memory along with the linguistic content of a talker's utterance (e.g., Pisoni, 1993; Palmeri, Goldinger, & Pisoni, 1993). These findings suggest that speech perception does not involve a mapping of invariant attributes or features in the signal onto idealized symbolic representations in memory, but rather employs highly detailed and specific encodings of speech which preserve many attributes of the acoustic signal.

The Role of Indexical Information

The human voice conveys a considerable amount of information about a speaker's physical, social, and psychological characteristics and these aspects, referred to as 'indexical' information (Abercrombie, 1967), complement the processing of linguistic content during spoken communication. Individuals differ in the size and shape of their vocal tracts (Fant, 1973; Joos, 1948; Peterson & Barney, 1952), in their idiosyncratic methods of articulation (Ladefoged, 1980) as well as in their individual glottal characteristics. These properties provide information about a speaker's identity (Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985) in addition to more general information about a speaker's origin and background (Labov, 1972). The speech signal also provides important information about more short-term aspects of a speaker's voice such as emotional or psychological states. These psychological factors are readily perceived when anger, depression, or happiness is recognized in a speaker's voice (Costanzo, Markel, & Costanzo, 1989; Markel, Bein, & Phillis, 1973; Murray & Arnott, 1993).

In everyday conversation, the indexical properties of the speech signal become quite important as perceivers use this information to govern their own speaking styles and responses. From more permanent characteristics of a speaker's voice that provide information about identity to the short-term vocal changes related to emotion or 'tone of voice,' indexical information contributes to the overall interpretation of a speaker's utterance. How then is the perception and encoding of the indexical properties of the speech signal related to the analysis of the more abstract linguistic content of an utterance? On the one hand, according to the traditional abstractionist accounts of speech perception, information conveyed by a talker's voice introduces variability or noise into the signal, presumably obscuring the linguistic content of an utterance. On the other hand, recent studies have shown that listeners are able to exploit this variation in the signal to apprehend talker-specific information. The essence of the problem is that both types of information are conveyed simultaneously along the same acoustic dimensions within the speech signal (Remez, Fellowes, & Rubin, in press). As the acoustic waveform constituting a talker's utterance reaches the listener's ear, information about the talker must be disentangled from information about the linguistic content of the utterance. Consequently, any explanation of 'perceptual normalization' for talker variability will necessarily need to include an account of the processing and representation of both the linguistic and indexical information that is carried in parallel in the speech signal.

A number of recent experiments have been reported explicitly addressing the relationship between linguistic analysis and talker variability. Several studies have shown that talker variability has a significant impact both on the perceptual processing of spoken utterances and on the memory representations constructed during the perception of spoken language. For example, talker variability has been shown to affect both vowel perception (Assmann, Nearey, & Hogan, 1982; Summerfield, 1975; Summerfield & Haggard, 1973; Verbrugge, Strange, Shankweiler, & Edman, 1976; Weenink, 1986) and spoken word recognition (Cole, Coltheart, & Allard, 1974; Creelman, 1957; Mullennix, Pisoni, & Martin, 1989). Mullennix et al. (1989) found that perceptual identification of words presented in noise was significantly poorer when the words were produced by multiple talkers than when they were produced by a single talker

(see also Sommers, Nygaard, & Pisoni, 1994). In addition, using a speeded classification task (Garner, 1974), Mullennix and Pisoni (1990) found that listeners had difficulty ignoring irrelevant variation in a talker's voice when asked to classify syllables by initial phoneme. When asked to classify the same stimuli according to the sex of the speaker, listeners also had difficulty ignoring irrelevant variation in initial phoneme. Taken together, these results suggest that variability due to changes in a talker's voice affects the recovery of the linguistic aspects of the speech signal. Aspects of the speech signal related to classifying talker identity seem to be integrally linked to those attributes related to the processing of the linguistic content of the signal.

In a more recent study, Remez, Fellowes, & Rubin (in press) found that information encoded in sine-wave replicas of spoken utterances also supports talker identification. These nonspeech signals are assumed to preserve only the time-varying phonetic information essential for linguistic interpretation and none of the acoustic attributes traditionally proposed to underlie the identification of talker's voice (Bricker & Pruzansky, 1976). Remez et al. found that listeners were able to discriminate and match to sample a set of sinewave replicas of utterances produced by unfamiliar talkers as well as identify sinewave replicas of utterances produced by a set of familiar talkers. These results show that time-varying phonetic information preserves at least some of the unique acoustic information that characterizes individual talkers' voices. Thus, it appears that information for talker identity and linguistic analysis are carried along the same acoustic-phonetic dimensions within the speech signal.

In addition to this recent evidence linking talker variability to linguistic analysis in perception, there is now considerable evidence that talker information affects memory as well. Martin, Mullennix, Pisoni, and Summers (1989) found that serial recall of spoken word lists produced by multiple talkers was poorer than recall of lists produced by a single talker; but the result was found only in the primacy portion of the serial recall curve. Martin et al. interpreted these findings to suggest that variation in a talker's voice from word to word in a list competes for processing resources in the recall task. Analysis of talker information during a memory task appears to be time- and resource-demanding leaving fewer resources for the rehearsal and transfer of words into long-term memory. In addition, Martin et al. found that recall of a series of visually presented preload digits was poorer when followed by a multiple-talker list than when followed by a single-talker list, again suggesting that talker variability increases the capacity demands of the working memory system.

In a subsequent series of experiments, Goldinger, Logan, and Pisoni (1991) investigated the serial recall of multiple-talker and single-talker lists using presentation rate as an additional experimental variable. Goldinger et al. found that at relatively fast presentation rates, serial recall in initial list positions was poorer for multiple-talker lists than for single-talker lists, replicating Martin et al. At longer presentations rates, however, recall performance was poorer in initial list positions for the *single-talker* rather than for the multiple-talker lists (see also Nygaard, Sommers, & Pisoni, 1995). This interaction between presentation rate and serial recall for the multiple- and single-talker word lists suggests that at fast presentation rates, when processing is constrained by time, talker variability affects both the perceptual encoding and rehearsal of items in the serial recall task. At slower presentation rates, when listeners have more time and resources to encode talker information, they are able to use that information to aid in the encoding of item and order information. These findings suggest that talker information may not be discarded in the process of spoken word recognition but rather retained in memory along with the more abstract, symbolic linguistic content of the utterance.

A stronger demonstration that detailed talker-specific information is retained in long-term memory comes from a series of recent experiments conduced by Palmeri, Goldinger, and Pisoni, (1993). Using a

continuous recognition memory task (Shepard & Teghtsoonian, 1961), Palmeri et al. found that talker-specific information is retained in memory along with lexical information, and this information can aid listeners' recognition memory. In the continuous recognition memory task, listeners were asked to listen to a list of spoken words and identify each word as 'old' or 'new.' Words repeated in the same voice were recognized better than words repeated in a different voice. This advantage for same voice repetitions suggests that listeners are simultaneously processing attributes of the linguistic content and attributes of the talker's voice and both sets of stimulus attributes are encoded and preserved in memory. Thus, variations in a talker's voice appear to be incorporated in memory into a highly-detailed, rich representation of a talker's utterance (see also Craik & Kirsner, 1974; Geiselman, 1979; Geiselman & Bellezza, 1976, 1977; Geiselman & Crawley, 1983).

Church and Schacter (1994) have also reported similar findings in a series of experiments aimed at assessing implicit savings for surface characteristics of spoken language. Using an implicit memory paradigm to study priming, Church and Schacter found that repetition of surface characteristics such as talker's voice, affective tone (happy or sad), and fundamental frequency from the study to test phase of their task resulted in better implicit word identification than when prime and target were dissimilar from study to test along each of these dimensions. Explicit recognition memory was not affected by these manipulations. The basis for this implicit savings was hypothesized to be a general-purpose perceptual representation system (PRS) which operates in a modality-specific manner to preserve detailed instance-specific perceptual information (Schacter, 1990). This detailed perceptual information can then be used, in this case in addition to lexical content, to implicitly access prior events. These findings taken together with those of Palmeri et al. (1993; see also Pollack, Pickett, & Sumby, 1954) suggest that the effects of talker variability on perception and memory are a consequence of the additional processing time and resources that are devoted to encoding talker-specific information when voice changes from item to item in these tasks. It appears that variability in the acoustic speech signal is encoded and retained in memory together with linguistic information. That both types of information, lexical and voice, are retained in memory suggests that a simple perceptual normalization process which simply discards surface characteristics of the speech signal is not adequate to account for these results and must be reconsidered (Garvin & Ladefoged, 1963; Johnson, 1990; Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989).

The research reviewed above makes a convincing case for the notion that talker-specific information is retained in memory and can be used as a cue, in addition to linguistic content, to retrieve specific linguistic events. The question still remains, however, as to the relationship between the processing of talker information and the processing of linguistic content. Are perception of talker identity and perception of linguistic content independent processes such that each contributes information separately about a to-be-remembered event? Or, are the perceptual analyses that extract both types of information integrally linked? The present series of experiments seeks to address these questions by focusing on perceptual learning of novel voices. If perception of talker identity and perception of linguistic content are independent mechanisms, then perceptual learning of voice information should be unrelated to the recovery of linguistic aspects of the speech signal. Conversely, if learning to become familiar with a talker's voice affects the intelligibility of their speech, then a direct link in processing between language perception and voice perception can be established. To that end, we sought to assess the effects of talker familiarity on spoken language processing. The goal was to evaluate traditional theoretical accounts of speech perception and spoken language processing as well as to address somewhat more general issues of perceptual learning in the high dimensional domain of spoken language.

Perceptual Learning

Relatively few studies have been conducted on the role of perceptual learning in the perception of speech and language in adults (but see, Lively, Logan, & Pisoni, 1993; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Logan, Lively, & Pisoni, 1991; Samuel, 1977; Strange & Dittmann, 1984). Although the role of categorization on perceptual sensitivity has long intrigued psychologists (Gibson, 1969; Gibson, 1991; Goldstone, 1994; Wohlwill, 1958), increased perceptual sensitivity to aspects of the speech signal has traditionally been considered an interesting empirical demonstration rather than a routine aspect of our everyday perceptual experience. Yet, in our use of language, we are often aware that through exposure and learning of a novel talker's voice for example, we become increasingly able to recover the linguistic aspects of an utterance that had seemed difficult to understand only moments earlier. If perceptual learning of novel voices can be shown to influence spoken language recognition, then the processing of linguistic content cannot be characterized as impervious to general mechanisms of perceptual learning and adaptation.

In a more general sense, it is possible to use the relationship between learning of talker identity and linguistic processing as a test case for the study of perceptual learning in a highly complex natural stimulus domain like speech. According to Gibson (1969), perceptual learning involves, 'an increase in the ability to extract information from the environment, as a result of experience and practice with stimulation coming from it' (page 3). Gibson identified two types of perceptual learning. The first type of perceptual learning suggests that perceptual sensitivity can be enhanced by preexposure or 'predifferentiation' to a set of stimuli (Hall, 1991). Mere experience with the stimulus domain increases perceivers' sensitivity. In the second type of perceptual learning, explicit experience categorizing or identifying stimuli allows perceivers to become attuned to specific diagnostic physical features (Gibson & Gibson, 1955). For this type of learning, the organization of stimuli into categories has been shown to have an important influence on subsequent perceptual sensitivity (Goldstone, 1994). Within this domain of perceptual learning, Lawrence (1949) developed a theory of acquired distinctiveness of cues, such that cues or features that are relevant to a task become generally distinctive. In the case of talker learning, categorizing or identifying talker's voices may lead to increases distinctiveness of the perceptual dimension of talker identity. If a benefit of perceptual learning of voice can be demonstrated for linguistic processing as well, then it would suggest that the same underlying dimensions subserve both perceptual abilities. In other words, if perceptual learning, or some type of acquired distinctiveness, along a dimension such as talker identity can be demonstrated to increase sensitivity to another higher-order dimension such as linguistic content, this finding would suggest that properties associated with talker identity are integral to the properties which underlie linguistic processing and are not independent of each other.

Clues to the issues just raised come from experiments examining talker identification and discrimination and from a handful of studies investigating perceptual learning of category structure in spoken language. At the outset, several studies have shown that listeners can learn to identify a set of talkers from their voices alone (e.g., Doddington, 1985; Williams, 1964) and are quite good at discriminating among unfamiliar talkers (e.g., Van Lancker & Kreiman, 1987). These studies have shown that a number of factors, such as a priori distinctiveness of the set of voices to be learned, the number of talkers to be identified or discriminated, and the length or duration of the utterances used during training (i.e., syllables, words, phrases, passages), can mediate learning of voices. Not surprisingly, listeners learn to recognize talkers' voices most readily when utterances of long duration from a few highly distinct talkers are used. These results suggest that a period of perceptual learning is required for listeners to become sensitive to talker-specific information in the speech signal. Listeners do not appear to effortlessly acquire expertise in talker recognition, but rather learn over time to attend to the unique acoustically distinct properties of each talker's voice.

The crucial research question then becomes whether perceptual learning of talker identity can influence perception of the linguistic properties of spoken language. That is, given experience with the particular aspects of the speech signal relating to talker identity, does it follow that listeners also become sensitive to talker-specific linguistic properties as well? Outside the domain of adaptation to voice (but see, Lively et al., 1994), selective training on particular acoustic dimensions has been shown to modify even highly stable low-level phonetic categories. For example, Logan, Lively, & Pisoni (1991; see also, Lively et al., 1993; Lively et al., 1994) have demonstrated perceptual learning of the /r/-/l/ contrast by adult native speakers of Japanese. This contrast is not phonemic for native speakers of Japanese and adult speakers have difficulty reliably categorizing instances from these categories. However, Logan et al. (as well as Lively et al., 1993; Lively et al., 1994) found, using a high-variability training program with explicit feedback, that native Japanese speakers can learn to discriminate the relevant acoustic dimensions and reliably classify /r/ and /l/. The authors conclude that perceptual learning of nonnative contrasts is possible and suggest that a certain amount of perceptual plasticity exists in adult speech perception. Thus, perceptual mechanisms that subserve phonetic categorization are susceptible to general processes of learning and adaptation.

In addition to perceptual learning of phonetic category structure, perceptual adaptation to continuous synthesized speech has also been demonstrated in several studies. Greenspan, Nusbaum, & Pisoni (1988) showed that repeated practice transcribing synthetic speech resulted in better comprehension performance. That is, exposure to the unique properties of synthetic speech resulted in better comprehension for novel instances of speech synthesized in the same manner. The learning, however, was somewhat specific to the training and testing materials used (see also Schwab, Nusbaum, & Pisoni, 1985). Practice with synthesized sentences improved transcription performance for synthesized sentences and isolated words. Practice with isolated synthetic words improved word but not sentence transcription suggesting that exposure during learning must be specific to the stimulus dimensions that will be relevant at test. Similarly, Dupoux and Green (in press) have found evidence for rapid perceptual adaptation to compressed speech. A group of listeners received exposure to digitally compressed speech showed better subsequent transcription performance for compressed speech than a group of listeners who were not previously exposed. Taken together, these practice effects with synthetic and compressed speech materials suggest that the speech processing system is capable of adjusting to a variety of distortions, both synthetic and natural, that occur in the acoustic signal. Further, listeners do not appear to just become familiar with the sound of synthetic or compressed speech in these experiments but rather they appear to learn the specific acoustic-phonetic mapping rules that describe the relationship between the rule-governed synthetic manipulations and each listeners' underlying linguistic knowledge (Greenspan et al., 1988).

The variation in spoken language that is introduced by individual talkers' speaking style and vocal tract anatomy is analogous to the distortions imposed when speech is synthesized by rule. Each talker's vocal style shapes the acoustic realization of linguistic constituents in different but systematic and predictable ways. Nevertheless, perceptual adaptation to individual talkers' voices, as mentioned previously, has traditionally been cast as a problem of subtracting variation due to individual differences in speakers' voices from underlying linguistic constants, rather than as a perceptual learning process in which listeners become sensitive to higher-order relational information in the speech signal which then subserves both talker identification and linguistic processing. Consequently, perceptual normalization procedures have been proposed that decode the talker-specific alterations to the speech signal to uncover the abstract, invariant linguistic categories (Garvin & Ladefoged, 1963; Johnson, 1990; Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989). In this sense, perceptual adaptation is assumed to be a mandatory process that works very quickly and automatically to strip away talker-specific information. According to this view,

perceptual learning of talker identity should be irrelevant with respect to linguistic analysis and speech intelligibility (Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989).

The Present Experiments

Demonstrating an influence of perceptual learning of novel voices on the recovery of the linguistic content of an utterance would provide evidence for interdependence of the mechanisms subserving each function. The goal of the present investigation was to determine if such a link exists and to assess the role of general cognitive processes such as perceptual learning, attention, and memory in the comprehension of spoken language. It should be noted that an influence of perceptual learning of talker identity on linguistic processing is a type of associative perceptual learning that has not been demonstrated previously for speech. Rather than simply providing practice recovering linguistic content in a variety of contexts (e.g., Greenspan et al., 1988), listeners in our experiments were trained to direct their attention to a seemingly unrelated stimulus dimension, the name of the talker. Categorization of voices focuses attention on talker-specific attributes that have been deemed distinct from linguistic attributes. To the extent that perceptual learning of one stimulus dimension transfers and affects perception of another stimulus dimension, one is justified in concluding that the two dimensions are 'cooperatively' or 'integrally' processed (Goldstone, 1994).

In the first experiment, we sought to initially establish the effects of learning novel voices on the perception of lexical content. Listeners learned to identify a set of ten voices (five male and five female) over a nine-day period and were then asked to recognize novel words produced either by talkers they had or had not heard during training. The purpose of the experiment was twofold. First, we wanted to investigate the perceptual learning of voices in its own right. Would listeners be able to learn to identify talkers' voices from lists of short isolated words? Of interest were issues concerning the identifiability or distinctiveness of individual talkers as well as individual differences in listeners' ability to learn the set of talkers. Second, given that listeners could successfully learn to identify a set of talkers, we sought to assess the effects of voice learning on their ability to recognize words mixed in noise. If words produced by familiar talkers are more easily recognized or more intelligible than words produced by unfamiliar talkers, this result would suggest that the perceptual learning in the talker identification task transferred to the word recognition task. This transfer of learning has several theoretical implications. One is that the perceptual learning of talker identity draws attention to the same perceptual attributes of the acoustic speech signal that are also important for word recognition. Therefore, the underlying representational code must somehow be integrated or linked in processing. Another implication is that mutual dependence of the perception of talker identity and linguistic identity would argue against traditional accounts of spoken language processing emphasizing abstract, context-free linguistic units. Instead, highly detailed information about the entire speech event is retained in long-term memory along with linguistic content.

In two additional experiments, the processes of perceptual learning and generalization were explored in greater detail. Again, listeners were asked to learn a set of ten novel voices. Then, they attempted to identify the linguistic content of speech produced by familiar or unfamiliar talkers. However, in both these experiments, listeners learned to identify talkers from sentence-length utterances rather than from isolated words as in the first experiment. In one experiment, listeners were asked to transcribe isolated words mixed in noise produced both by talkers learned during training and by a set of unfamiliar talkers. In the other experiment, listeners were asked to transcribe sentence-length utterances mixed in noise produced by familiar and unfamiliar talkers. Our goal was to assess what type of talker-specific information is learned when listeners are trained with sentence-length versus word-length utterances and whether this learning would be task-specific, generalizing only to similar test stimuli (Greenspan, Nusbaum, & Pisoni,

1988). We hypothesized that learning voices from isolated words would be a difficult task, focusing listeners' attention on detailed talker-specific acoustic-phonetic variation. Thus, we reasoned that any benefit or transfer from the dimension of talker identity to spoken words would be greatest when fine acoustic-phonetic distinctions were required, as in the isolated word recognition task. Learning voices from sentence-length utterances conversely was hypothesized to be a much easier learning task, focusing listeners' attention on more coarse-grained, global attributes of talker identity such as prosody, intonation, and rhythm. Thus, we expected that listeners would show considerable transfer of learning to a sentence transcription task. In these experiments, the role of attention in perceptual learning is explicitly addressed. Attention to different inventories of talker-specific information was directly manipulated to determine the ultimate pattern of task-specific generalization and its benefit in spoken word recognition.

Experiment 1

Recently, Nygaard, Sommers, and Pisoni (1994) found that learning a talker's voice facilitates subsequent phonetic analyses. In their study, listeners were trained over a nine-day period to identify a set of five female and five male voices from isolated words and were then give a speech intelligibility task. During training, listeners were required to associate one of ten common names with each voice and were given explicit feedback regarding their performance The results showed that listeners who heard familiar talkers at test were better able to extract the linguistic content of isolated words than those who heard unfamiliar talkers at test. Their initial findings suggested that perceptual learning of voice can modify the linguistic processing of isolated words.

The present investigation was designed to replicate and extend Nygaard et al.'s (1994) analyses of instance-specific factors in perceptual learning and to report additional data from this initial experiment. In Nygaard et al. (1994), large individual differences in listeners' ability to learn the set of talkers were observed. Some listeners improved dramatically over the nine-day training period while others showed little to modest improvement. Nygaard et al. only reported and analyzed data from listeners who were able to learn the set of talkers' voices and reached a minimum performance criterion. Of interest as well, however, is the performance of listeners who were not able to learn the talkers' voices to criterion. These subjects provide an ideal comparison group because while they received the same training and exposure to the set of talkers' voices, they were not able to attend specifically or successfully to the unique talker-specific aspects of the speech signal. Thus, in the present report, we assessed the nature of perceptual learning in these experiments by comparing listeners who learned the voices to listeners who did not learn the set of voices well enough to identify them reliably to criterion. The question of interest here is whether the perceptual learning of voices that benefits spoken word recognition is due to the mere exposure of listeners to the set of talkers used or whether successful identification and categorization of voices is necessary to increase perceptual sensitivity to the linguistic content of familiar talkers' utterances (Gibson, 1969).

In addition to evaluating the nature of perceptual learning and transfer of training in this paradigm, we report more detailed analyses of voice learning for all subjects. In these analyses, we address issues concerning individual differences in listeners' ability (listener-specific factors) and differences in talker identifiability and intelligibility (talker-specific variables). Do individual listeners differ in their talker-learning strategies? Are all talkers equally identifiable? Are male and female voices different in terms of how easily they are learned? Our aim was to answer these questions by investigating in depth what factors mediate voice learning and how perceptual sensitivity to voice is acquired under these learning conditions.

Two groups of listeners were asked to explicitly learn to recognize a set of ten talkers' voices over a nine-day period. At the end of training, one group of listeners (the experimental group) received a list of

novel words produced by the familiar talkers they had learned to categorize and the other group of listeners (the control group) received novel words from unfamiliar talkers. Here we report the data from all listeners, those that were successful and those that were unsuccessful at learning the set of talkers' voices. In addition, we reportmore detailed analyses of voice learning for all listeners.

Method

Subjects.

Sixty-six undergraduate and graduate students at Indiana University participated in this study. Subjects were assigned to one of four conditions. Nineteen listeners served in the trained experimental (Group I) and in the trained control conditions (Group II). Fourteen listeners served in the untrained experimental (Group III) and in the untrained control conditions (Group IV). All listeners were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. The listeners were paid for their participation.

Stimulus Materials.

Three sets of stimuli were used in this experiment. All items were selected from a database of 360 monosyllabic words produced by ten male and ten female speakers taken from the vocabulary of the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965) and from phonetically balanced (PB) word lists (Egan, 1948). Each word was recorded on audiotape using a high-quality professional microphone and digitized at 10 kHz by a 12-bit analog-to-digital converter. The root mean squared (RMS) amplitude levels for all words were digitally equated. Word identification tests in quiet showed greater than 90% intelligibility for all words. In addition, all words were rated to be highly familiar on a seven-point rating scale (Nusbaum, Pisoni, & Davis, 1984).

Procedure.

Training. Two groups of nineteen listeners completed nine training sessions over a period of two weeks. Listeners were asked to learn to recognize each talker's voice and to associate each voice with one of ten common names (see Lightfoot, 1989). Digitized stimuli were presented using a 12-bit digital-to-analog converter and were low-pass filtered at 4.8 kHz. Stimuli were presented to listeners over matched and calibrated TDH-39 headphones at approximately 80 dB SPL.

During each of the nine hour-long training sessions, both groups of trained listeners completed three difference phases designed to acquaint them with the ten different voices to be learned. The first phase was a *familiarization task* in which five words from each of the ten talkers were presented in succession to the listeners. Then, a ten-word list composed of one word from each talker was presented. As each item was presented to the listener, the name of the corresponding talker was displayed on a computer screen. The experimenter instructed the subjects to listen carefully to each word presented and to attempt to learn the name associated with each talker's voice. This familiarization procedure was intended to give listeners some direct experience with the range of variability within each talkers' voice.

The second phase of training consisted of a *recognition task* in which ten words from each of the ten talkers were presented in random order to listeners. The hundred words used in this phase did not overlap with those used in the first phase. Listeners were asked to identify the name of the talker that produced each token and were given immediate feedback after each trial as to the correct name. Subjects

responded by pressing the appropriate key on a keyboard. Keys 1 through 5 were labeled with male names (Bill, Joe, Mike, etc.) and keys 6-10 were labeled with female names (Sue, Mary, Carol, etc.).

During each training session, listeners completed two repetitions of the first two phases of training and were then administered a test phase. As in the second phase of training, ten words from each of the ten talkers were presented in random order. Listeners were asked to identify each speakers' voice by choosing the appropriate name on each trial. Feedback was not given during the test phase each day to measure learning.

The same one hundred words were used as stimuli for each of the training phases. However, listeners never heard the same item produced by the same talker in both the test and training phases on a given day. Further, stimuli were reselected from the data base on each day of training so that listeners never heard the same item produced by the same talker in training. So, for example, the ten words that were produced by one talker on the first day of training would be produced by another on the second day of training. This procedure was intended to maximize the number of different tokens listeners heard from each talker.

Generalization. During the tenth session of the experiment, both groups of listeners completed a generalization test. One hundred novel words produced by each of the ten talkers were used. As in the test phase, ten words from each of the ten talkers were presented to listeners in random order. Listeners were asked to identify the talker on each trial and no feedback was given. This generalization test was identical to the test phase used during training each day except that a set of novel words produced by the same ten talkers was used.

Transfer Word Intelligibility. In addition to the generalization test, listeners were given a speech intelligibility test in which they were asked to identify the lexical content of isolated words presented in noise. One hundred novel words were presented at either 80, 75, 70, or 65 dB (SPL) mixed in continuous white noise that was low-pass filtered at 4.8 kHz and presented at 70 dB (SPL). This procedure resulted in four signal-to-noise ratios: +10, +5, 0, and -5. Twenty-five words were presented at each signal-to-noise ratio. In this task, subjects were asked to transcribe each word (rather than to identify the talker's voice) on each trial. Listeners in the trained experimental group were presented with words produced by the ten talkers they had previously learned during training. Listeners in the trained control group were presented with the same words produced by ten new talkers (five male and five female) that they had not heard during training.

In addition to giving the trained listeners (experimental and control) the word intelligibility test, two additional groups of fourteen untrained listeners were also run in these tasks to control for inherent intelligibility differences between the two sets of talkers' voices. Thus, one group of untrained controls (untrained experimental group) was given the word intelligibility test with items produced by the talkers used in the training procedures. This was the same test given to the trained experimental group. The other group of untrained controls (untrained control group) was given the word intelligibility test with items produced by a set of talkers not used in training. This was the same test given to the trained control group.

Results

Training.

Data from the two trained groups (Group I and Group II) revealed large individual differences in listeners' voice identification performance. Figure 1 shows scatterplots of each individual listeners' performance from Day 9 of training plotted against their performance on the generalization test given in the tenth session. Results in this figure and in subsequent training figures were always drawn from the final test with no feedback that was given on each day of training. The top graph shows data from listeners in the experimental group and the bottom graph shows data from listeners in the control group averaged across talkers. Recall that both groups of listeners received identical training with the same group of talkers. It is the stimulus set used for the subsequent word intelligibility test that distinguishes these groups. This figure illustrates two aspects of our results. First, performance on the ninth day of training is well correlated with performance in the generalization test ($r = +.83$, $p < .01$, for the experimental group; $r = +.88$, $p < .01$, for the control group). Second, for both groups of listeners, individual subjects differed greatly in their performance on this task. The range on Day 9 was 69 percentage points from the poorest to the best learner.

Insert Figure 1 about here.

Due to the large individual differences, listeners were divided into two groups based on their voice identification scores. A criterion of 70% correct voice identification on the ninth day of training was selected to group listeners into 'good' and 'poor' learners. Our rationale for dividing our listeners into two groups was that to assess the effects of voice learning on word intelligibility, we needed to have a group of listeners who had indeed learned the set of voices used in the experiment. This partitioning of the data based on this performance criterion also allowed us to compare listeners who learned the voices with listeners who were simply exposed to the voices over the nine days of training. Using this criterion, nine subjects from both the experimental and control conditions were classified as 'good' learners[4] and ten subjects from both experimental and control groups were classified as 'poor' learners.

Figure 2 shows listeners' voice identification performance, averaged across talker's voice, for the test phase of Days 1-9 of training and for the generalization test on Day 10. Percent correct voice identification is plotted as a function of day of training for 'good' and 'poor' learners in both the experimental and control conditions. Again, recall that the experimental and control groups were given training on the same set of voices. All subjects identified talkers consistently above chance even on the first day of training and all listeners improved over the nine days of training. A three-way repeated measures analysis of variance (ANOVA) with training group (experimental vs. control), day of training (Days 1-9 and generalization), and listener learning group ('good' vs. 'poor') as factors was conducted on the percent of correct responses. A significant main effect of days of training was found [$F(9,306) = 69.58$, $p < .001$] indicating that overall, listeners' voice identification performance improved over days of training. A significant main effect of learning group was also found [$F(1, 34) = 78.31$, $p < .001$] indicating that 'good' learners identified talkers' voices more accurately than 'poor' learners. In addition, a significant interaction between days of training and learner group was found [$F(9, 306) = 9.55$, $p < .001$]. 'Good' learners

[4] One listener from the control group fell just short of 70% correct on the ninth day of training. However, their performance rose to 75% correct on the generalization test and consequently, was included with the 'good' learners from the control group.

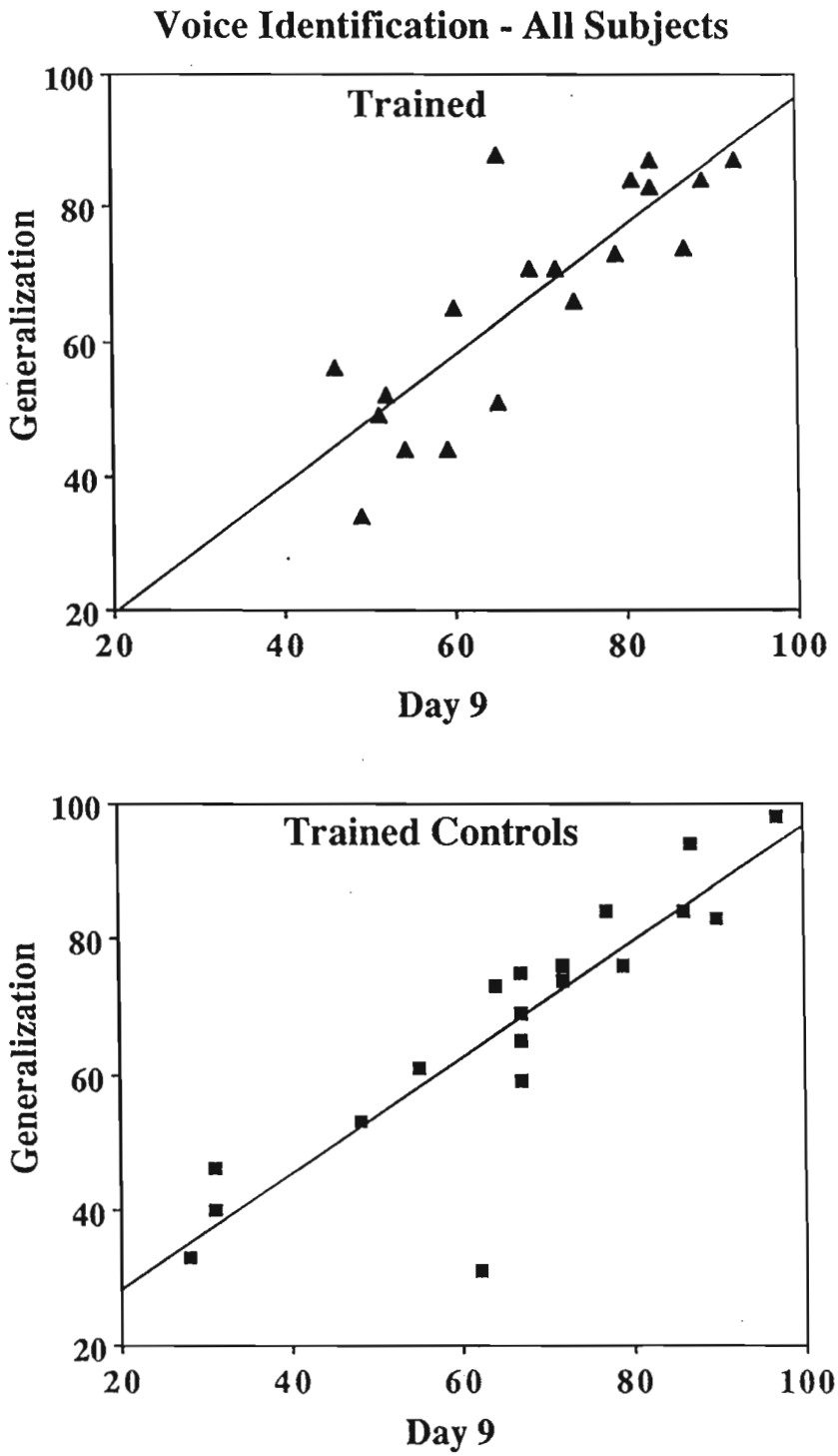


Figure 1. Scatterplots of Day 9 voice identification performance for each listener plotted as a function of generalization on Day 10. The top panel shows the results for the trained experimental group and the bottom panel shows the results for the trained control group.

improved to a greater extent over days of training than did 'poor' learners. From Day 1 to Day 9, 'good' learners performance rose from 46.06% to 81.72% correct while 'poor' learners performance only rose from 38.9% to 54.5% correct.

Insert Figure 2 about here.

In order to examine the perceptual spaces for these voices and how they changed with training, multidimensional scaling was performed on the confusion matrices generated during the first and last day of training for 'good' and 'poor' listeners. The matrices were constructed using the number of times listeners confused each voice with each of the other nine voices during the test phase administered to listeners at the end of the first and ninth session of training. Four separate three-dimensional scaling solutions were calculated (see Nygaard & Kalish, 1994) for each of the four day versus group combinations. Figure 3 shows two-dimensional (dimensions 2 and 3) representations of each of the four solutions. Dimension 1 is not represented in this figure because across all solutions, it uniformly corresponded with sex of the speaker. Although Dimensions 2 and 3 do not map onto obvious acoustic dimensions of the speech signal, the differences in perceptual distances among talkers for 'good' and 'poor' listeners is diagnostic. For both 'poor' and 'good' learners, there is a considerable amount of perceptual confusion on the first day of training. However, 'good' and 'poor' learners differ after the last day of training. Male and female talkers are well separated in perceptual space for the 'good' learners while there is no such separation for the 'poor' learners. For 'good' learners, male speakers are well represented along Dimension 3 and female speakers are well represented along Dimension 2. For 'poor' learners, male and female speakers are not separated along either Dimension 2 or Dimension 3.

Insert Figure 3 about here.

The results of the scaling solutions also illustrate the differences in identifiability of the voices used in training. Individual talkers' voices were quite different in how easily they could be learned by listeners. Figure 4 shows scatterplots of Day 9 talker identification performance plotted against generalization test scores for individual talkers, averaged across listeners. The top graph shows data from listeners in the experimental group and the bottom graph shows data from listeners in the control group. This figure illustrates two aspects of the learning data. First, individual talker identification on Day 9 of testing is significantly correlated with performance on the generalization test for both groups ($r = +.89$, $p < .01$, for the experimental group; $r = +.91$; $p < .01$, for the control group). Second, for both groups of listeners, identification performance varied greatly depending on the individual voice. For example, identification scores for male voices were superior to identification scores for female voices, at least for the set of voices used in this experiment.

Insert Figure 4 about here.

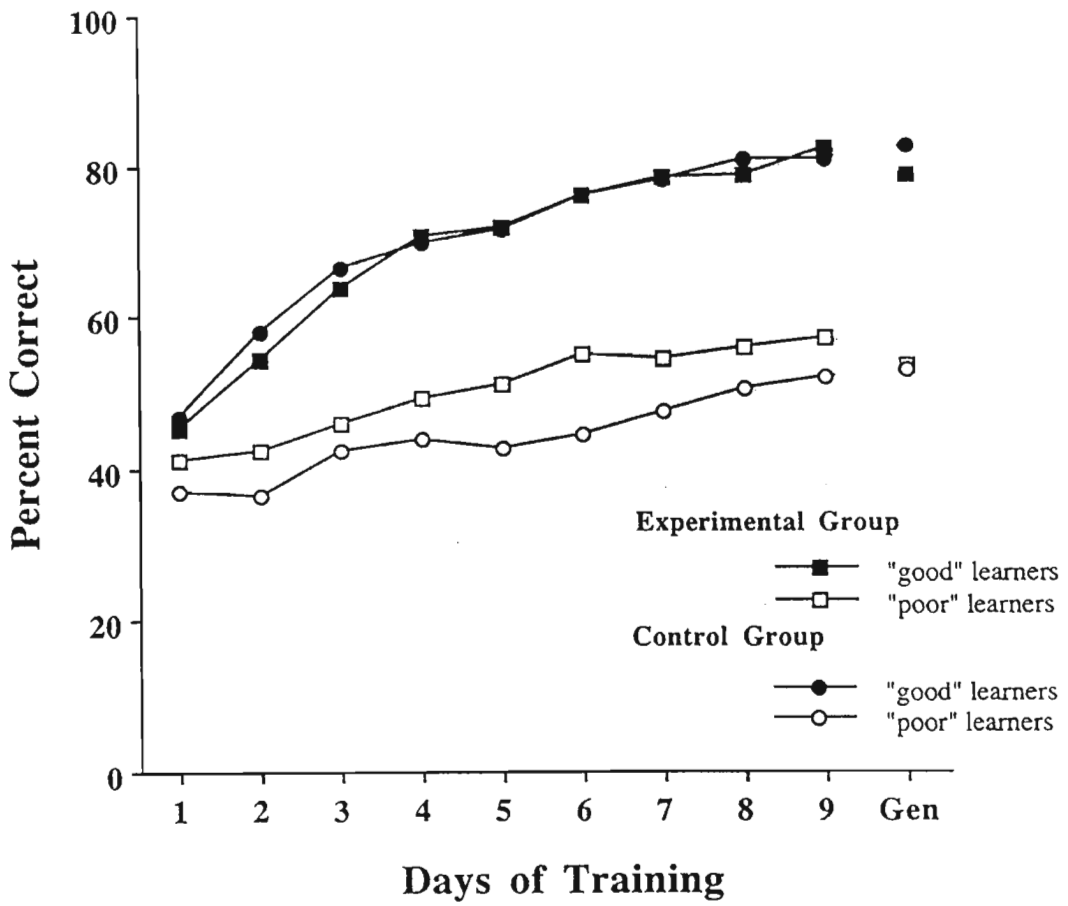


Figure 2. Percent correct voice identification from isolated words is plotted for each day of training and for the generalization test for both 'good' and 'poor' learners in the trained experimental and control groups.

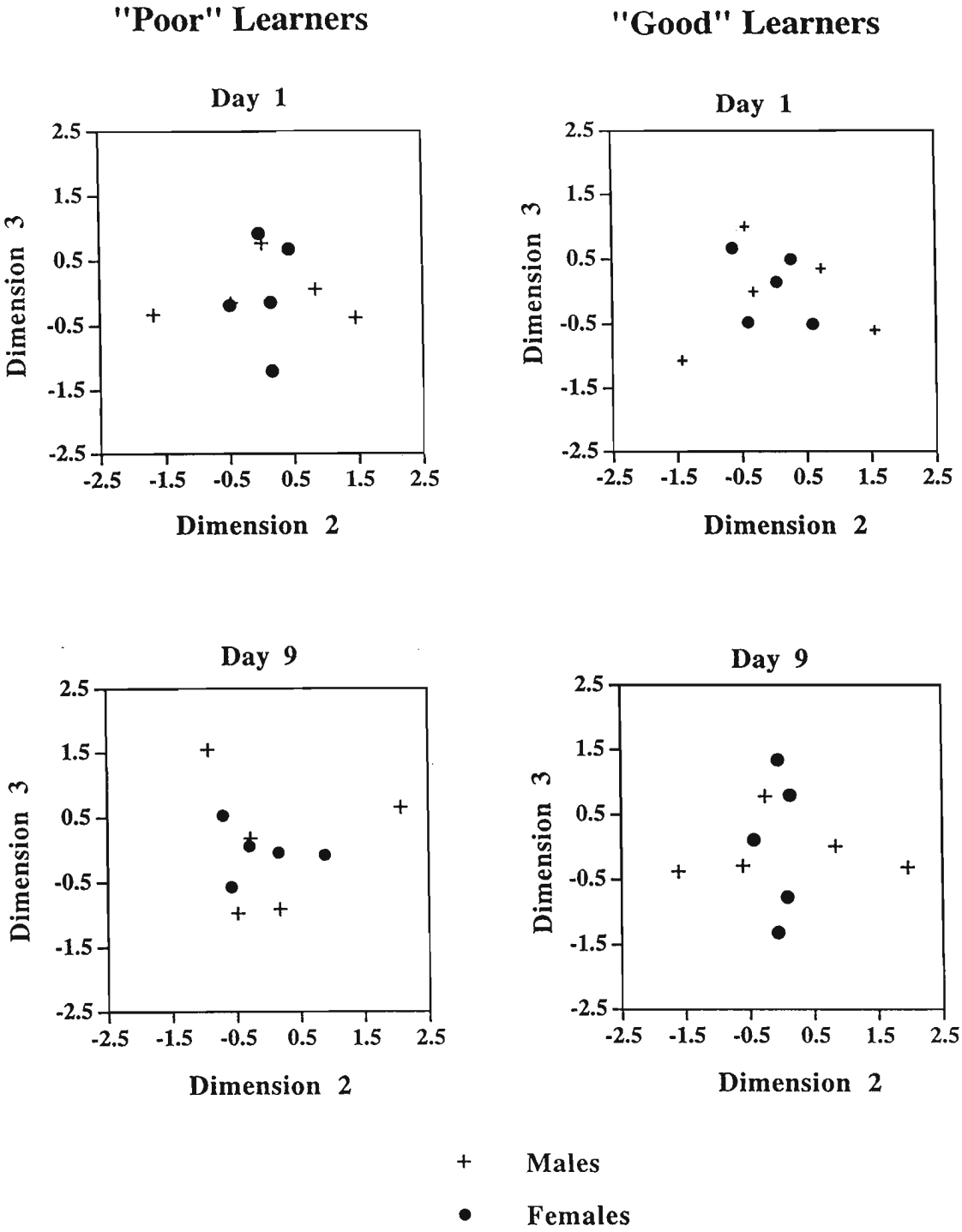


Figure 3. Dimensions 2 and 3 of multidimensional scaling solutions are plotted for Day 1 of training (top panels) and Day 9 of training (bottom panels). Scaling solutions for the 'poor' learners are on the right and solutions for the 'good' learners are on the left.

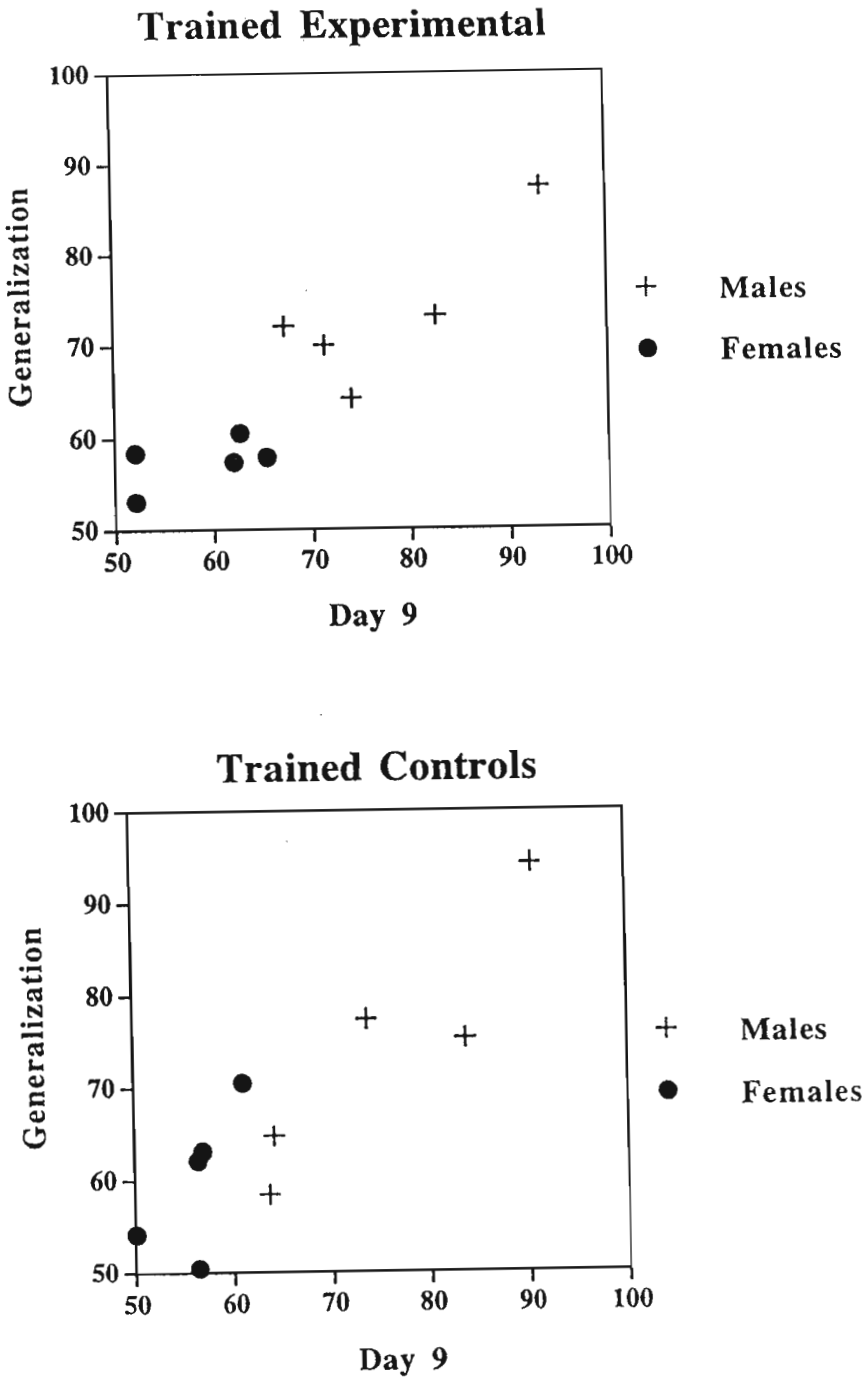


Figure 4. Scatterplots of Day 9 voice identification performance for each individual talker (males and females) plotted as a function of generalization performance. The top panel shows the results for the trained experimental group and the bottom panel shows the results for the trained control group.

Generalization.

The generalization test showed almost identical recognition of voices from the novel words presented on Day 10 as on the final day of training. These results are also shown in Figure 2. In the experimental condition, percent differences between the generalization test and Day 9 of training were 3.55 and 3.66 for 'good' and 'poor' learners respectively. In the control condition, percent differences between the generalization test and Day 9 of training were 1.89 and 1.00 for 'good' and 'poor' learners respectively. T-tests revealed no significant differences in listeners' performance across conditions between the two tests.

Word Intelligibility.

Figure 5 shows the percentage of correct word identification as a function of signal-to-noise ratio for both groups of trained listeners and for both groups of untrained listeners. The top graph shows data from 'good' learners and the bottom graph shows data from the 'poor' learners. Two separate repeated measures ANOVAs were conducted for the 'good' and 'poor' learners using training condition (trained experimental, trained control, untrained experimental, and untrained control) and signal-to-noise ratio (+10, +5, 0, -5) as factors. Data from all listeners in both untrained control groups were used as comparison for both 'good' and 'poor' learners and the same data are included in both analyses and are included in both graphs of Figure 5.

'Good' learners. The analysis for the 'good' learners revealed a significant main effect of signal-to-noise ratio [$F(3, 126) = 351.55$, $p < .001$]. As expected, identification performance decreased from the +10 to the -5 signal-to-noise ratio for all four groups. The analysis also revealed a significant main effect of training condition [$F(3, 42) = 7.43$, $p < .001$] indicating that identification performance differed across the four training conditions. A significant interaction was found between training condition and signal-to-noise ratio [$F(9, 126) = 3.03$, $p < .001$] suggesting that identification performance among the groups was larger for some signal-to-noise ratios than others.

Insert Figure 5 about here.

A follow-up ANOVA excluding the data from the trained experimental group was conducted to determine if there were any inherent differences among the three control groups for whom the talkers' voices were unfamiliar. Using training group (trained control, untrained experimental, and untrained control) and signal-to-noise ratio as factors, we again found a significant main effect of signal-to-noise ratio [$F(3, 102) = 221.38$, $p < .001$], but no significant main effect was observed across control conditions [$F(2, 34) = 0.16$, $p > .9$]. This analysis confirms that the significant main effect for training condition found in the original analysis was due to better word identification performance in the trained experimental group who received words produced by familiar voices at test than in the three control groups who received words produced by unfamiliar voices at test.

'Poor' learners. The analysis for the 'poor' learners also revealed a significant main effect of signal-to-noise ratio [$F(3, 132) = 290.85$, $p < .001$] indicating that identification performance decreased from the +10 to the -5 signal-to-noise ratio for all four groups. No main effect of training group ($p > .73$) or interaction between training group and signal-to-noise ratio ($p > .14$) was found. Word identification performance across training groups did not differ significantly.

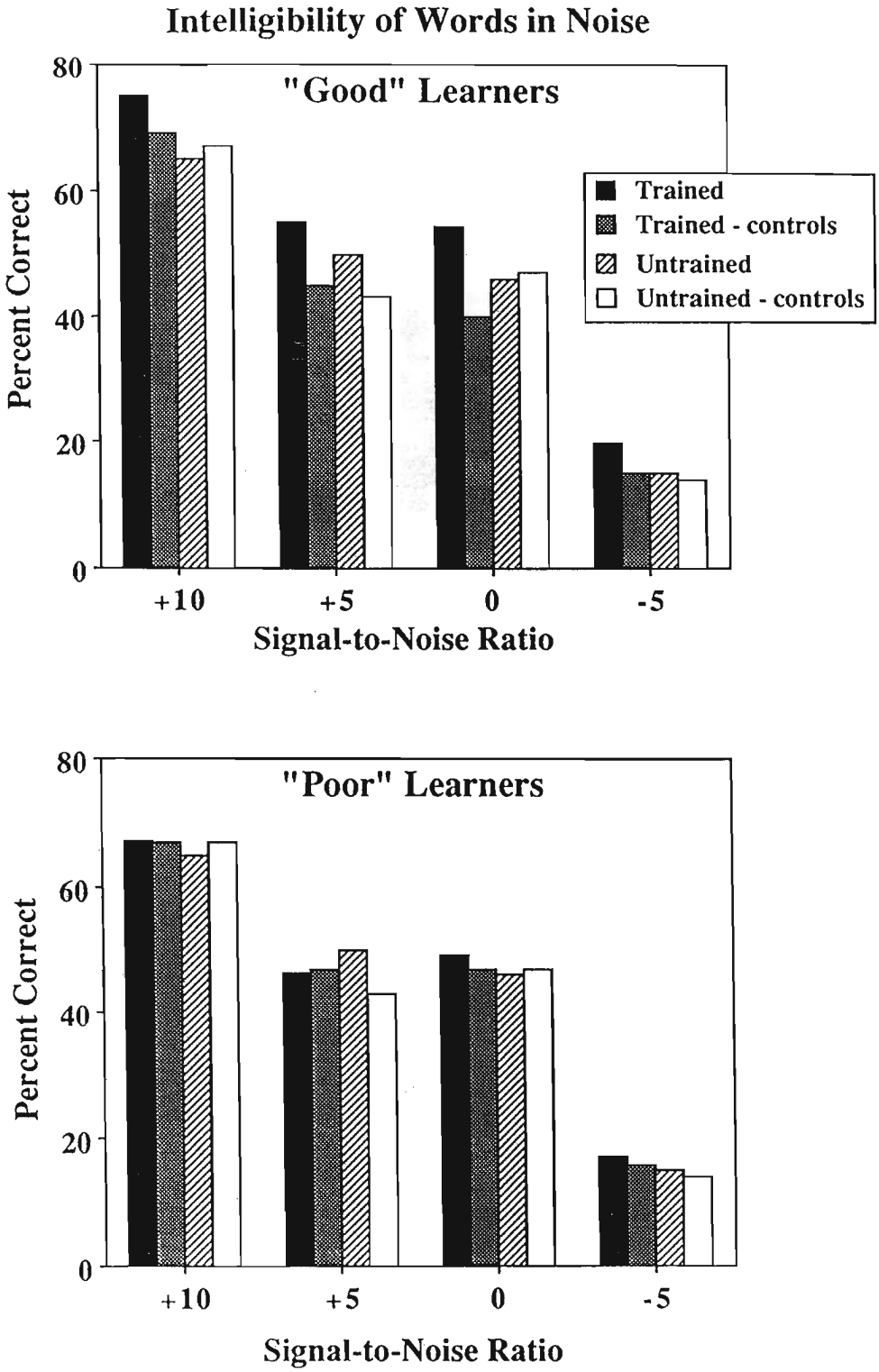


Figure 5. Percent correct word recognition for both training groups (experimental and control) and both untrained groups is plotted for each signal-to-noise ratio. The top panel shows results for the 'good' learners and the bottom panel show results for the 'poor' learners.

Discussion

Several important findings were obtained in Experiment 1. First, listeners display large individual differences in their ability to learn to identify a set of ten voices. Individual listener performance across training groups ranged from 28% correct for the poorest learner to 97% correct for the best learner after nine days of training. This finding suggests that simple exposure to the set of voices over the nine day period is not sufficient for perceptual learning of talkers' voices to occur. Rather, listeners must attend to and process fully the talker-specific aspects of the speech signal to show improvement in voice learning.

Second, given that listeners differ in their ability to learn the voices, it is possible to characterize some listeners as 'good' voice learners and others as 'poor' voice learners. Although voice learning performance represents a continuum of ability (see Figure 1), grouping listeners in this manner provides a useful heuristic for analyses of the consequences of different learning abilities. For example, we found that 'good' and 'poor' learners not only differed in their absolute ability to identify voices by Day 9, but they also differed in the amount of learning that took place over time. That is, 'good' and 'poor' learners start out roughly similar at the end of the first day of training, but their identification performance quickly diverges over the next nine days of training. 'Good' learners improve to a much greater extent that 'poor' learners. This divergence suggests that through practice categorizing and explicitly identifying voices, 'good' learners become attuned to the acoustic/linguistic details that distinguish each talkers' voice. 'Poor' learners do not seem to acquire the same kind of perceptual sensitivity or distinctiveness using these voice dimensions during this type of laboratory training, at least under the conditins used here.

A clue to the differences between the two groups comes from our multi-dimensional scaling analyses. 'Good' and 'poor' learners appear to develop qualitatively different strategies to identify different talkers which may account for their ultimate success in the voice recognition task. Although both groups were able to distinguish male from female voices, they differed in the other dimensions they used to distinguish individual female and individual male talkers. By the end of the ninth day of training, 'good' learners appeared to use Dimension 3 to distinguish female talkers and Dimension 2 to distinguish male talkers. 'Poor' learners appeared to be using both dimensions to distinguish all the talkers and it is this difference in strategy that may account for differences in identification performance. Of course, a note of caution should be mentioned here. These scaling solutions only provide a suggestion of strategic differences in voice learning and no acoustic dimensions have been identified that relate to the dimensions that result from scaling. In addition, it should be noted that the 'poor' learners do improve somewhat over the nine days of training and presumably, could eventually learn to identify our set of talkers to criterion. Given the limitations of the study, it was not possible to determine whether 'poor' learners, or 'good' learners for that matter, would continue to improve if given additional training or whether 'poor' learners would eventually switch to a more useful learning strategy. This is obviously a question for future research.

The individual differences found among listeners is complemented by individual differences in the identifiability of talkers' voices. Both the scaling analyses and performance differences in identifying individual talkers' voices suggest that talkers vary a great deal in their perceptual distinctiveness. In particular, it appears that male voices, at least those used in this experiment, were significantly easier to identify than the female voices. In addition, whatever makes each voice more or less perceptually distinctive in terms of ease of identification is at least somewhat abstract with respect to the items used during training. Listeners were quite good at generalizing what they had learned to a new set of stimulus materials in the generalization task. Listeners were able to use talker-specific knowledge to identify voices from linguistic tokens that they were never exposed to before, suggesting that listeners are learning something more general about each talkers' voice and style of speaking. Thus, it appears that both talker-specific and

listener-specific variables contribute to the eventual identification of a talker's voice (Bradlow, Nygaard, & Pisoni, 1995). Learning is general in the sense that while individual talkers' voices are learned and retained, this knowledge is not necessarily tied to the specific stimulus tokens used in training.

Finally, the most important finding is that familiarity with a talkers' voice influences linguistic processing; specifically, the intelligibility of isolated words mixed in noise. Perceptual learning of a set of talkers' voices caused listeners to be better able to recover the linguistic content of the signal. This finding marks the first experimental demonstrations that mechanisms responsible for analyzing talker identity are not independent from the mechanisms responsible for analyzing the lexical content of an utterance. Perceptual learning and long-term retention of talkers' voices selectively modified the ability of listeners to process the phonetic content of these speech signals.

The present results also demonstrate that through learning to associate a name with each talkers' voice, listeners began to attend to talker-specific aspects of the speech signal that were also relevant for perceiving the linguistic content of the same signal. The perceptual dimension related to talker identity appeared to become much more distinctive during categorization training and this perceptual sensitivity transferred to processing of linguistic information. This transfer of learning or sensitivity from talker identity to linguistic identity is crucial because it implies that these two sources of information, and the perceptual processing of these two sets of dimensions, are inexorably linked in perception. In a more general sense, talker identity and linguistic information appear to be integral dimensions analogous to color dimensions such as brightness and saturation (Goldstone, 1994). Although lexical and indexical information are arguably higher-order aspects of spoken language, they may nevertheless behave like lower-level perceptual dimensions (see Mullennix & Pisoni, 1990).

The differences in performance between the 'good' and 'poor' learners in this experiment indicate that associative learning was a necessary but not sufficient condition for listeners to learn each talker's voice and consequently for listeners to show a benefit or transfer of training from talker identity to word recognition. Although all listeners received the same amount of training, only listeners that could successfully identify the talkers' voices explicitly showed a benefit in the word recognition test. This indirect test of the type of perceptual learning necessary to impact word intelligibility provides evidence for the assertion that mere exposure or mere repetition of the voices over a period of time does not result in sufficient perceptual differentiation along the voice dimension to modify processes of spoken word recognition (see Gibson, 1969). One explanation of these results is that the 'poor' learners did not receive sufficient training to 'fine tune' or adjust their attentional mechanisms to the relevant talker-specific information in the signal. For whatever reason, the 'good' learners were able to attend to the specific acoustic-phonetic details that not only reliably distinguished one talkers' voice from another but also reliably help in processing the phonetic aspects of the speech signal. It should be noted that 'poor' learners do not necessarily have difficulty processing speech from a variety of talkers, but rather when the perceptual system is taxed, as when words are presented in noise, they were unable to utilize their prior knowledge of each talkers' idiosyncratic style of speech to help recover the phonetic content and lexical information in the signal.

Although we proposed that attentional differences between the 'good' and 'poor' groups of learners account for both the differences in perceptual learning of voice and their ability to identify linguistic aspects of the signal produced by 'pre-exposed' talkers, we have no direct evidence that it is attention during learning to talker-specific details that results in perceptual sensitivity for linguistic information as well. In the next two experiments, we address this issue more directly by experimentally focusing listeners' attention on specific aspects of talker voice information and then evaluating how well listeners generalize

this specific learning to a linguistic task. By specifically manipulating the type of talker information available during training and then evaluating linguistic processing for matched or mismatched material, we can evaluate how perceptual learning of talker identity relates to perceptual sensitivity for linguistic information in the signal.

Experiment 2

Experiment 2 was designed to assess the nature and extent of the perceptual learning demonstrated in the first experiment. To that end, listeners were trained to recognize a set of ten talkers from sentence-length rather than from word-length utterances. After training was completed, intelligibility was assessed using isolated words produced by familiar and unfamiliar talkers. The aim was to determine if the information learned about a talker's voice from sentences generalizes to the perception of isolated spoken words. The assumption was that training with sentence-length utterances would focus listeners' attention at a different level of analysis than training with isolated words. It was hypothesized that because sentences contain highly distinctive prosodic and rhythmic information in addition to the specific acoustic-phonetic implementation strategies unique to individual talkers, perceptual learning of voices from sentences would require attentional and encoding demands specific to those test materials.

Two groups of listeners learned to identify voices from sentence-length utterances over a three-day training period. The experimental group was then tested with isolated words mixed in noise to assess intelligibility of talkers they had been exposed to in training. The control group was tested with isolated words produced by a set of unfamiliar talkers. The isolated words used at test also differed in their lexical characteristics. Half were 'easy words' -- high frequency words from sparse lexical neighborhoods and half were 'hard words' -- low frequency words from dense lexical neighborhoods (Luce, Pisoni, & Goldinger, 1990). Because lexically hard words require attention to fine acoustic/phonetic detail for successful lexical access, it was hypothesized that perceptual learning of talkers' voices might improve the identification of lexically hard words in noise to a greater extent than lexically easy words.

Method

Subjects.

Participants were 46 undergraduate and graduate students at Indiana University. Twenty-seven subjects served in the experimental condition and nineteen subjects served in the control condition. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. Subjects were paid for their participation.

Stimulus Materials.

Two sets of stimuli were used in this experiment. The sentence training stimuli consisted of 100 Harvard sentences (IEEE, 1969; Egan, 1948) produced by 10 male and 10 female talkers. Harvard sentences are all meaningful English, mono-clausal sentences containing five key words plus a variable number of function words. The key words all contained one or two syllables. The isolated word stimuli consisted of 100 monosyllabic words produced by 10 of the same talkers (5 male and 5 female) that produced the sentence materials. None of the isolated words were used in the sentence stimuli. The isolated words varied in their neighborhood characteristics (Luce, Pisoni, & Goldinger, 1990). Fifty 'easy' and fifty 'hard' words were selected. 'Easy' words are high frequency items that come from a sparse lexical neighborhoods. 'Hard' words are low frequency words that come from dense lexical neighborhoods. A

lexical neighborhood consists of the set of words which differ by one phoneme from the target word. In addition, all of the isolated words were rated as highly familiar (Nusbaum, Pisoni, & Davis, 1984). All stimuli were digitized on-line at a sampling rate of 20 kHz using 16-bit resolution. The root mean squared (RMS) amplitude levels for all stimuli were digitally equated.

Procedure.

Two groups of listeners completed three training sessions with sentence-length materials. Digitized stimuli were presented using a 16-bit digital-to-analog converter and were low-pass filtered at 10 kHz. Stimuli were presented to listeners over matched and calibrated TDH-39 headphones at approximately 80 dB SPL. A pretest-posttest design was used in which both groups of listeners received identical pre- and posttests with isolated words produced by the same set of talkers. Each group was then trained using different sets of talkers. For the experimental group, the same talkers were used for pre- and posttests and for training. For the control group, different talkers were used during training than in the pre- and posttests. Thus, in this experiment, we were able to directly compare word intelligibility performance for the same set of words.

Pretest Word Intelligibility. A pretest-posttest design was used to directly evaluate the effects of talker familiarity on word intelligibility. In both the pretest and posttest, 100 isolated words produced by ten talkers (5 male and 5 female) were presented at either 80, 75, 70, or 65 dB (SPL) mixed in continuous white noise that was low-pass filtered at 10 kHz and presented at 70 dB (SPL) over matched and calibrated TDH-39 headphones. This manipulation yielded four signal-to-noise ratios: +10, +5, 0, -5. Equal numbers of words were presented at each of the four signal-to-noise ratios. Listeners were asked to identify each word by typing their response on a keyboard. Responses were recorded on-line by a PDP 11/34 computer. For listeners in the experimental condition, the words were produced by the ten talkers they would subsequently hear in training. For listeners in the control condition, the words were produced by ten talkers they would not subsequently hear in training.

Training. The two groups of listeners also completed three days of training to familiarize themselves with the voices of ten talkers. The experimental group of 27 subjects learned the voices of the same ten talkers that were used for the pre- and posttests. The control group of 19 subjects learned the voices of ten different talkers. As in Experment 1, both groups were required to identify each talker's voice and associate that voice with one of 10 common names. On each day of training, both groups of listeners completed three different phases. The first was a *familiarization task* in which one sentence from each talker was presented in succession. Each time a sentence was presented, the name of the talker appeared on a CRT screen in front of the listener. Subjects were asked to listen carefully to the sentences and to attend specifically to the talker's voice. As in the first experiment, no response was required by the listeners during the familiarization phase.

The second phase of training consisted of a *recognition task* in which subjects were asked to identify the talker who had produced each sentence. Ten new sentences from each of ten talkers were presented in random order to listeners who were asked to identify each voice by pressing the appropriate button on a keyboard. On each trial, after all subjects had responded, the correct name appeared on a CRT screen. Stimuli were selected so that each sentence was produced by a different talker on each day of training. Stimuli were rotated in this manner to maximize listeners' exposure the range of variability within each talker's voice.

The third phase, testing, was identical to the second phase except that no feedback was given. Again, in this test phase, individual tokens did not overlap with those used in training so that listeners never heard the same word produced by the same talker twice in training or at test.

Posttest Word Intelligibility. After three days of sentence training, listeners received a posttest word recognition test which was identical to the pre-test. Subjects were asked to identify the linguistic content of isolated words produced by familiar or unfamiliar talkers at four signal-to-noise ratios.

Generalization. After the posttest, the experimental group received a generalization test in which the set of words used in the pre- and posttests were presented to listeners for voice identification. Recall that the experimental and control groups learned different sets of talkers' voices during training but received the same voices at test. For the control group, the voices used at test were unfamiliar. However, for the experimental group, the voices used at test were familiar because they were used to produce the sentences. Thus, the experimental group was given an additional generalization test in which the had to identify the talker (rather than the word) on each trial from the same isolated words used in the posttest. This test allowed us to get a direct measure of how well the perceptual learning of voices from sentences generalized to identification of voices from isolated words.

Results

Training.

As in the first experiment, we found large individual differences in listeners' voice identification performance. However, far fewer listeners failed to reach the criterion performance of 70% correct on the third day of training when learning voices from sentences. Because there were too few 'poor' listeners, particularly in the control condition, eleven subjects from the experimental condition and two subjects from the control condition were eliminated from the analyses. That left sixteen subjects in the experimental conditions and seventeen subjects in the control condition.

Figure 6 shows voice identification performance for the experimental and control groups over the three days of training. All listeners showed continuous improvement over the three days of training. Both groups identified talkers consistently above chance even on the first day of training and performance rose to nearly 85% correct by the last day of training. A repeated measures analysis of variance (ANOVA) with learning and days of training as factors showed a significant main effect of day of training, $F(2,62) = 74.04$, $p<.001$, and also a significant main effect of group $F(1,31) = 20.27$, $p<.001$. The control group performed significantly better than the experimental group in learning their set of talkers.

Insert Figure 6 about here.

Generalization.

Figure 6 also shows voice identification performance on the word generalization test. Recall that the set of isolated words used at test were familiar voices only for the experimental group. Listeners in the experimental group were 63% correct in the generalization task identifying the voices they had learned during training from sentences. This performance is not significantly different from voice identification performance on sentences at the end of the first day of training.

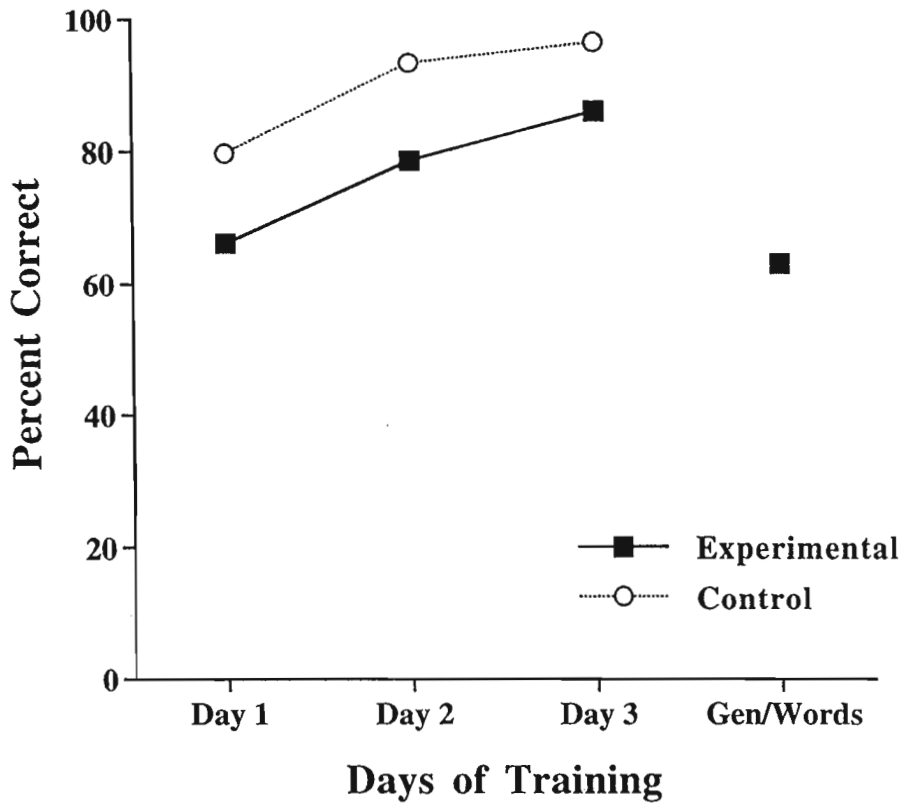


Figure 6. Percent correct voice identification from sentences is plotted for each day of training and for the generalization test given to the experimental group.

Isolated Word Intelligibility.

Figure 7 shows percent correct word identification at pretest and posttest as a function of signal-to-noise ratio and lexical neighborhood structure. The top panel shows the results for the experimental group and the bottom panel shows results from the control group.

Insert Figure 7 about here.

Posttest performance was in general expected to be superior to pretest performance due to mere repetition of the same items. Thus, the magnitude of the difference between pre- and posttest performance for the experimental versus the control group served as our measure of talker-specific perceptual learning. To assess this effect of perceptual learning on word intelligibility, the difference in percent correct word identification from pretest to posttest was calculated for each listener. Figure 8 shows these differences scores for both the experimental and control groups averaged across signal-to-noise ratio for both easy and hard words. Although there is more improvement for subjects in the experimental condition who heard the familiar voices at posttest than for subjects in the control condition, the effects of voice familiarity on word intelligibility were small and did not reach statistical significance [$F(1,31) = 3.33$, $p<.08$]. A repeated measures ANOVA calculated on difference scores averaged across signal-to-noise ratio with training group (experimental vs. control) and word type (easy vs. hard) as factors showed no significant main effects or interactions.

Insert Figure 8 about here.

Discussion

These findings demonstrate that perceptual learning of voices improves dramatically as longer duration utterances are used to familiarize listeners. The majority of listeners in this experiment learned to identify talkers' voices over three training sessions as compared to the nine training sessions needed in the first experiment. Further, a larger percentage (72% for sentences versus 47% for words) of listeners achieved a criterion of 70% voice identification when learning voices from sentence-length utterances even with fewer days of training. There are at least two explanations for improved learning with sentence-length utterances. One explanation is that sentence-length utterances just provide listeners with a larger sample of speech containing the same information they get with word-length utterances (Peters, 1955a). The reason listeners are better able to identify voices from sentence-length utterances is that they receive, in effect, five or so words on which to make their talker identity judgments on each trial rather than just one word which they received when training with isolated words. However, an alternative explanation is that sentence-length utterances also contain additional, qualitatively different information. That is, sentence-length utterances provide listeners both with the voice-specific information found in word-length utterances and also with information about fundamental frequency, duration, and rhythm, which vary over the entire utterance. When listening to sentences, subjects are exposed to prosodic and rhythmic information at the sentence level in addition to the acoustic-phonetic implementation differences among talkers at the word and segmental levels.

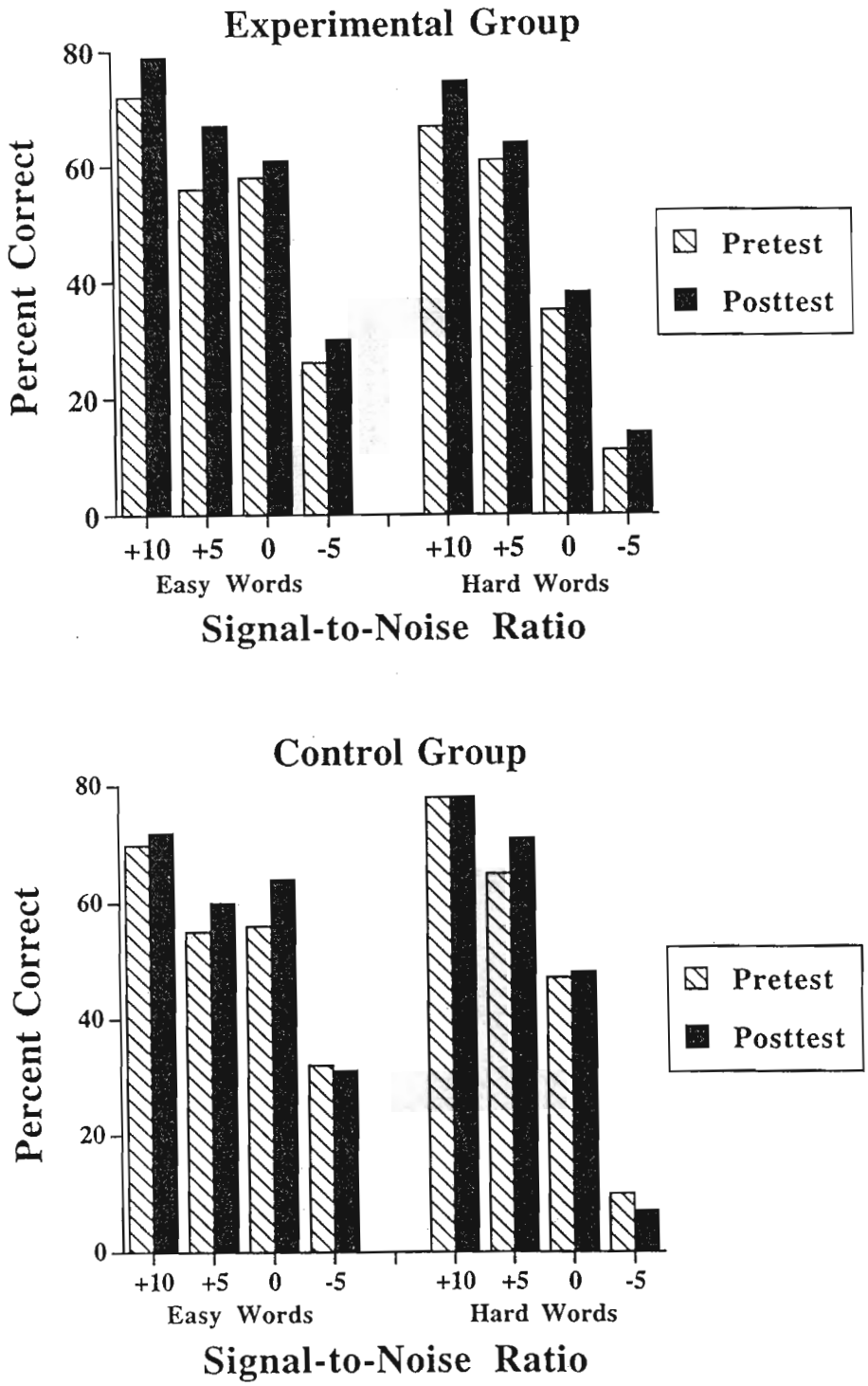


Figure 7. Percent correct pre- and posttest word recognition performance is plotted at each signal-to-noise ratio for easy and hard words. The top panel shows results for the experimental group and the bottom panel show results for the control group.

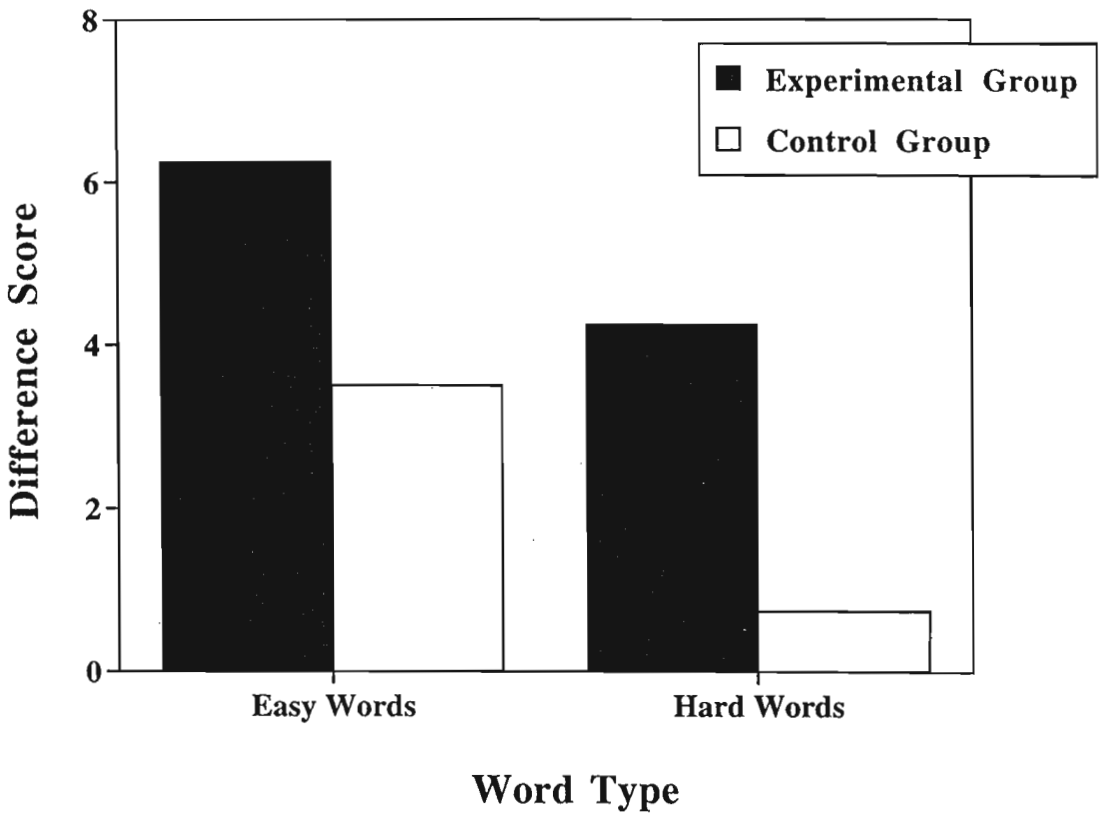


Figure 8. Percent difference scores collapsed across signal-to-noise ratio are plotted for word type and experimental group.

The results of the generalization test suggest which explanation might account for the differences in learning rates between training with words and training with sentences. Generalization of learning using sentences to isolated words was not very good. It appears that listeners learned a qualitatively different set of acoustic properties to identify talkers from sentence-length utterances than from isolated words. Although it could be argued that listeners in this experiment received far less training on the set of voices, it is still the case that listeners were identifying talkers' voices quite well by Day 3 from sentence-length utterances. This finding suggests that in learning to identify voices from sentences, listeners are allocating their attention to several different levels of analysis. In this task, they are not required to attend as closely to the fine acoustic-phonetic details that distinguish voices at the word level. Rather, listeners are learning to distinguish voices along perceptual dimensions in the sentence-length utterances that do not completely overlap with the dimensions used to distinguish voices at the word-length level.

Given that learning voices from sentences does not generalize well to the perception of talker identity in isolated words, it is not surprising that this learning does not significantly improve intelligibility of isolated words. Although listeners who heard familiar voices in the posttest were somewhat better at identifying isolated words than listeners who heard unfamiliar voices, these results fell just short of significance. Listeners appeared to focus on a talker-specific dimension in the sentence-length utterances that was not as useful in the word-length utterances. If this is true, however, listeners should show large effects of talker identification training with sentence-length utterances on the perception the linguistic content of sentence-length utterances. Thus, if there is a match between the type of talker information that is learned during training and the type of linguistic information that is presented at test, then perceptual learning of voices should once again strongly influence the perception of the linguistic attributes of the signal. Experiment 3 was designed to address this issue.

Experiment 3

Experiment 3 was similar to the second experiment except that after training listeners to learn talkers' voices from sentence-length utterances was completed, subjects were given an intelligibility test using sentences produced by familiar and unfamiliar talkers. Two issues were addressed here. First, does specific training on sentence-length utterances generalize to similar test materials? We predicted that the talker-specific information learned from sentences would influence the recognition of words in sentences. Therefore, when a match between information learned during training and information required at test occurred, we expected that the transfer of perceptual learning along the talker identity dimension would increase perceptual sensitivity to the linguistic content as found in the first experiment.

Second, are sentence-length utterances which have higher-level semantic and syntactic constraints susceptible to the effects of familiarity with a talker's voice? This experiment was also designed to determine if talker-specific information could affect linguistic processing when other perceptual constraints might override its influence. Sentences not only contain the phonological and lexical information that influences the recognition of words, but they also contain higher-level syntactic and semantic constraints. Given the redundancy of linguistic information at several levels in sentence-length utterances, our aim was to determine if talker-specific information would influence the recognition of words in sentences or whether this source of information would become relatively unimportant in the context of sentences.

Method

Subjects.

Participants were 20 undergraduate and graduate students at Indiana University. Eleven subjects served in the experimental condition and nine subjects served in the control condition. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. Subjects were paid for their participation.

Stimulus Materials.

Training and test stimuli were drawn from a digital database consisting of 100 Harvard sentences produced by 10 male and 10 female talkers (Torretta, 1995). Sentence identification tests showed greater than 90% intelligibility for all sentences in the quiet. Sentences were digitized on-line at a sampling rate of 20 kHz using 16-bit resolution. The root mean squared (RMS) amplitude levels for all stimuli were digitally equated.

Procedure.

Training. Training was similar to that used in Experiments 1 and 2 except that subjects were trained using a set of 50 sentences. Two groups of listeners completed the three days of training. The experimental group of 11 subjects learned the voices of the same ten talkers that were used for the sentence intelligibility test. The control group of 9 subjects learned the voices of ten different talkers. Listeners were not administered a pretest as in Experiment 2 because it was assumed that the set of 50 sentences used at test would be too memorable if used in a pretest as well.

Sentence Intelligibility Test. In the sentence intelligibility test, 48 novel sentences produced by ten talkers (5 male and 5 female) were presented at either 75, 70, or 65 dB (SPL) in continuous white noise that was low-pass filtered at 10 kHz and presented at 70 dB (SPL), yielding three signal-to-noise ratios: +5, 0, -5. An equal number of sentences was presented at each of the three signal-to-noise ratios. Subjects were asked to transcribe the sentence on a sheet of paper. For subjects in the experimental condition, the sentences were produced by the ten familiar talkers they had learned during training. For subjects in the control condition, the sentences were produced by ten novel talkers they had not been exposed to during training.

Results

Training.

Figure 7 shows talker identification performance for the experimental and control groups over three days of training. All subjects showed continuous improvement over the three days of training. As in Experiment 2, both groups of subjects identified talkers consistently above chance even on the first day of training and performance rose to nearly 85% correct by the last day of training. A repeated measures analysis of variance (ANOVA) with learning and days of training as factors showed a significant main effect of day of training, $F(2,36) = 78.029$, $p<.001$, and no other significant effects.

Insert Figure 9 about here.

Sentence Intelligibility.

Transcription performance was scored in four ways for each sentence. The scoring methods were: (1) *Sentence Correct* -- A response was scored as correct, if and only if, the whole sentence was transcribed correctly (a sentence was still counted correct if the correct verb was used in the wrong form); (2) *Keywords Correct* -- The actual number of key words transcribed correctly out of five possible per sentence was scored; (3) *Total Words Correct* -- The total number of words transcribed correctly per sentence was scored; (4) *Meaning Correct* -- A response was scored as correct when the sentence was correct (as stated above) or when the overall meaning of the sentence was correct using different wording. Because all four scoring methods produced essentially the same pattern of results, only the scoring method using key words correct will be reported here.

Subjects' performance on the sentence intelligibility task was assessed by determining the number of key words correct in each test sentence, adding up the total number of correct key words across sentences and averaging these totals across subjects. Each Harvard sentence contained 5 'key words' and the test set of 48 Harvard sentences contained 240 key words. Figure 10 shows the total number of key words correct as a function of signal-to-noise ratio averaged across subjects for the experimental and control groups.

Insert Figure 10 about here.

A repeated measures ANOVA with signal-to-noise ratio (+5, 0, -5) and training group (experimental vs. control) as factors showed a significant main effect of training group, $F(1,18) = 220.378$, $p<.001$. Subjects in the experimental condition who heard sentences produced by familiar talkers were able to transcribe more key words correctly across all signal-to-noise ratios than the control subjects who heard sentences produced by unfamiliar talkers. A significant main effect of signal-to-noise ratio, $F(2,36) = 286.26$, $p<.001$, was also found indicating better performance at the higher signal-to-noise ratios. Finally, there was a significant interaction between training group and signal-to-noise ratio, $F(2,36) = 44.41$, $p<.001$. As shown in Figure 8, this interaction demonstrates that the effect of talker familiarity became larger as signal-to-noise ratio decreased.

Discussion

These results suggest that perceptual learning of talkers' voices from sentences facilitates the recognition of words in sentences produced by familiar talkers. Learning talkers' voices from sentences generalized to the transcription of similar test materials suggesting that through learning the distinctions among talkers during training, listeners became sensitive to talker-specific linguistic information that was relevant when perceiving sentences in noise. Listeners appear to attend to dimensions when learning voices from sentences that are most relevant when they must extract the linguistic content of sentence-length utterances. That is, learning in this task appears to involve attention to the specific dimensions of talker identity that are relevant at test. These findings replicate and extend the transfer of training results found in

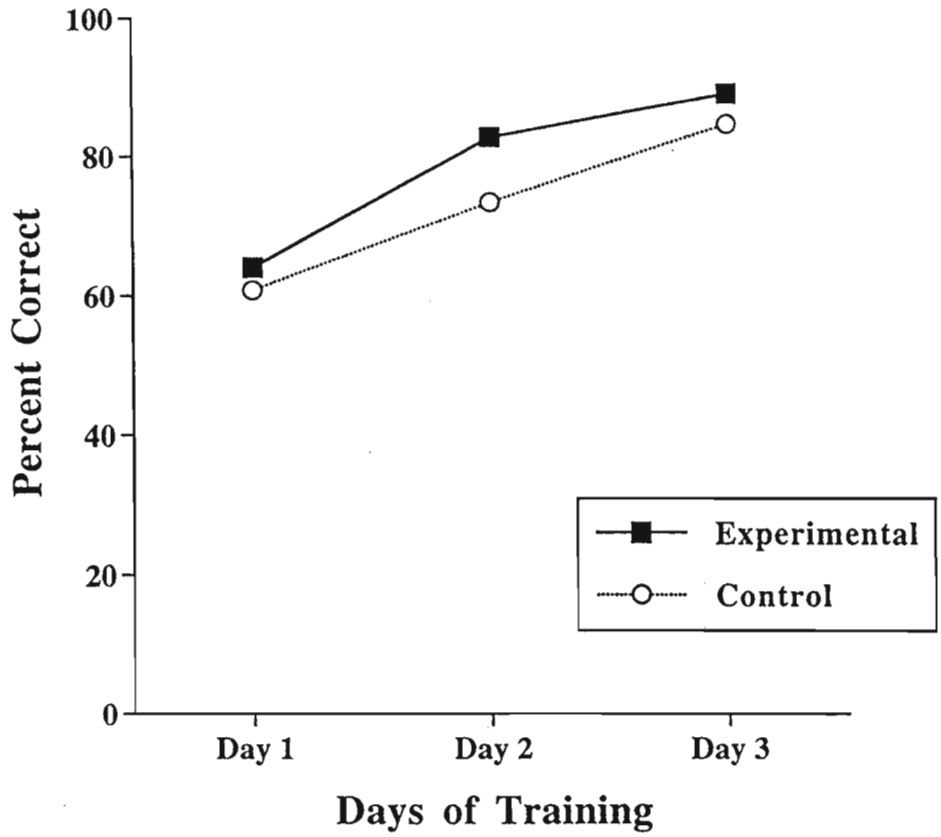


Figure 9. Percent correct voice identification from sentences is plotted for each day of training.

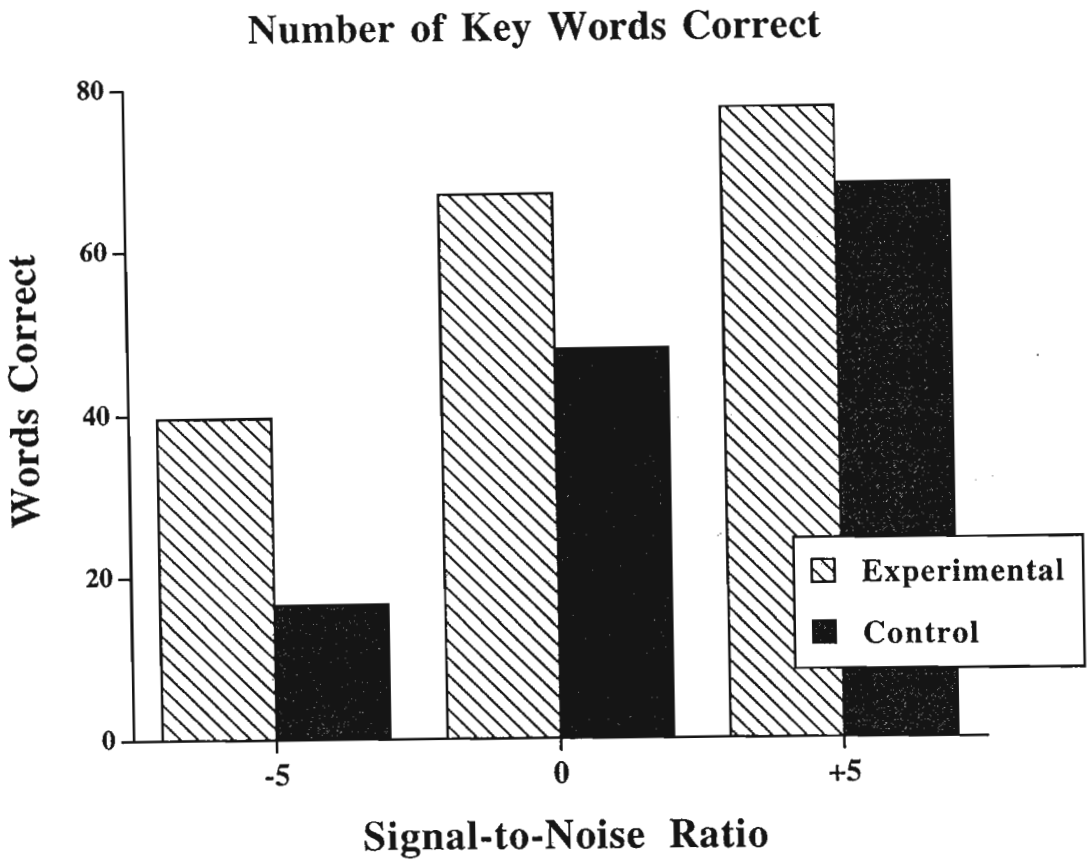


Figure 10. Number of 'key words' correct is plotted as a function of signal-to-noise ratio for the experimental and control groups.

the first two experiments by demonstrating transfer of training from sentences to sentences where none was found in Experiment 2 from sentences to words.

An interaction between familiarity and signal-to-noise ratio was also found suggesting that as listening conditions became more difficult, listeners make greater use of the talker-specific knowledge they acquired in the first phase of the experiment. The difference between the experimental and control groups was larger at lower signal-to-noise ratios. Thus, as overall intelligibility of the stimulus set deteriorated, listeners were more likely to bring to bear talker-specific information to aid in their transcription performance. This finding suggests that listeners may use talker information to a greater extent in listening situations that are degraded. Familiarity with voice information is thus extremely important in most real world listening situations. For example, these findings lead to the prediction that at cocktail parties, on city streets, and in other typical listening environments, listeners are better able to understand talkers whose vocal attributes are most familiar to them. Further, it is well known in clinical audiology that listeners with hearing impairments are better able to understand speech produced by familiar friends and family than speech produced by strangers with unfamiliar vocal tracts and speaking styles.

These results also confirm that learning to identify talkers' voices is much easier from sentences than from isolated words. As in Experiment 2, listeners readily learned to identify the ten talkers over three days of training and all listeners in this experiment reached our 70% criterion level of performance. This finding suggests that sentences are a rich source of talker-specific information and that learners are sensitive to the additional talker information in sentence-length utterances (Peters, 1955a, 1955b). Although we have not provided a direct test, it does appear that sentences provide qualitatively different sources of information about a talkers' voice than do isolated words. That is, sentences appear to provide information about talker-specific acoustic-phonetic implementation strategies in addition to higher-order information about idiosyncratic prosody, rhythm and meter. During training, listeners apparently exploit all sources of information to help learn the set of voices in this task.

Finally, the results confirm the importance of the role of talker information in spoken language processing. Familiarity with talkers' voice was found to affect the perception of sentence-length utterances despite the rich higher-level semantic and syntactic constraints found in these utterances. This finding suggests that perceptual learning of voice and its effect on language comprehension is a general phenomenon that operates in a variety of listening situations. Familiarity with talker-specific information not only aids speech perception when higher-level, top-down strategies are limited, but also when several sources of linguistic information are available to the listener. This finding suggests that the use of talker-specific information is important in general to the perception and comprehension of spoken language and is used in conjunction with other sources of information in spoken language to derive a linguistic interpretation of a talker's utterance.

General Discussion

The results of the present series of experiments demonstrate that perceptual learning of talker's voice facilitates the analysis of the linguistic content of the signal. Listeners who learned to attend to talker-specific attributes of the speech signal were able to use that information to aid in the recovery of the linguistic content in the acoustic speech signal. This finding suggests at the broadest level that the perception of indexical or personal properties in the speech signal and the perception of linguistic properties are not independent, but rather are fundamentally linked in the perception of spoken language. Thus, acquiring sensitivity along the dimension of talker identity also increases perceptual sensitivity for other linguistic dimensions. That perceptual learning of one dimension generalizes to the perception of another

dimension suggests that these dimensions are integral with respect to their perceptual underpinnings (Mullennix & Pisoni, 1990). This demonstration of the influence of perceptual learning of talker identity on linguistic processing has implications not only for theories of speech perception and spoken language processing, but also more generally for theories of perceptual learning and perception.

More specifically, the present series of experiments demonstrates that attention during perceptual learning must be specific to the perceptual task required of the listener. When confronted with an intelligibility task using isolated words, listeners who had attended during training to word level talker-specific attributes showed perceptual facilitation in recognition of isolated words. Listeners who had attended during training to sentence level talker-specific information showed little benefit on a word identification task, but large familiar voice benefits were found in a sentence transcription task. These task-specific aspects of the current investigation suggest that the transfer of perceptual learning of voice to linguistic processing requires that listeners learn about distinctiveness along just those talker-specific dimensions that will be relevant later. The implication of this finding is that there are different kinds of talker-specific information available in different kinds of utterances and that all levels of talker-specific information are susceptible to the effects of perceptual learning.

The proposal that learning talker information can affect linguistic processing, while intuitive, is not presently addressed by any of the contemporary theories of speech perception and spoken language processing (Fowler, 1986; Liberman & Mattingly, 1985; McClelland & Elman, 1986; Stevens & Blumstein, 1978). Either explicitly or implicitly, theories of speech perception have traditionally dismissed talkers' voice in speech perception as a source of noise that must be discarded or separated from linguistic content. To the extent that talker-specific aspects of the signal have been studied, adjustments to variability introduced by talker-specific attributes of the signal have been characterized by the use of normalization procedures in which listeners make short-term automatic compensations for talker variability (Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989). Our finding that learning a talker's voice makes their speech more intelligible suggests a very different interpretation of the role of talker variability in speech perception. The assumption of independence of talker information and linguistic information clearly needs to be abandoned. The fact that attention to talker identity increases sensitivity to phonetic information in the signal suggests that both types of percepts, indexical and linguistic, involve at least some of the same underlying attributes (Remez, Fellowes, & Rubin, in press).

Beyond calling into question traditional assumptions about the role of talker identity in speech perception, the present set of findings suggest several conclusions about the nature of representation and processing within the domain of language. First, our findings confirm that talker-specific information is retained along with linguistic information in long-term memory for linguistic events (Church & Schacter, 1994; Goldinger, 1992; Palmeri et al., 1993; Nygaard et al., 1994). Detailed representations of linguistic events appear to be retained in long-term memory and linguistic categories may consist of collections of instance-specific exemplars (Nosofsky, 1987; Hintzman, 1986), rather than some type of abstract prototypical summary representation in which aspects of spoken language such a talker's voice (and speaking rate, vocal effort, etc., for that matter) are eliminated. However, our findings take this notion one step further. In addition to showing that talker information is retained in memory, these experiments also demonstrate that linguistic processing and the perception of talker identity are linked in a contingent fashion (Nygaard et al., 1994). Not only is talker information retained along with lexical information, but these two dimensions are not separable or independent in perception and attention (Mullennix & Pisoni, 1990). There are important processing consequences for a shared, or detailed representation of linguistic events. One of these consequences is that perceptual learning of voice identity can result in talker-specific sensitivity to linguistic content. Another consequence is that shared, detailed representations take linguistic

representations out of the domain of abstract, symbolic units and into the domain of representation and memory for natural events and specific instances of these events (Brooks, 1978; Jacoby & Brooks, 1984).

Second, the retention of detailed talker-specific information and its effect on linguistic processing has ramifications for the type of processing architecture and perceptual operations that must underlie speech and language perception. One of the most influential ideas in the area of language and cognitive architecture has centered on the notion of modularity (Fodor, 1983). Modules are special-purpose, automatic, serial, cognitively impenetrable structures that process perceptual input quickly and reflexively. As applied to language processing, a modular account of speech perception and word recognition assumes that language is processed by a special-purpose device that is concerned only with the linguistic aspects of spoken language. Higher-level pragmatic or semantic knowledge as well as instance-specific properties of the signal are assumed to be outside the domain of the language module.

As applied to speech perception in particular, Liberman & Mattingly (1985) have argued for a 'phonetic module' that operates exclusively on the linguistic aspects of the signal quickly discarding acoustic information associated with non-linguistic aspects. According to this view, the phonetic module should be impervious to the perceptual learning of talker identity. The perception of a talker's voice is assumed to have separate underlying representations and analyses from the perception of linguistic content. Given the present findings, however, it appears that the phonetic module does 'know' something about the talker's voice. Further, it appears that the phonetic module is susceptible to the general processes of perceptual learning and memory. Therefore, the perception of speech cannot be completely encapsulated with respect to what we know about general cognitive structures, analyses, and representations.

One way to reconcile a modular account of language processing with the present findings is to assume that it is the perceptual normalization process or the set of 'perceptual operations' which discard talker variability that are learned in our task. That is, talker-specific perceptual operations are retained or developed during the course of training and listeners find speech from familiar talkers to be more intelligible than speech from unfamiliar talkers because they are better able to disentangle talker from linguistic information (Kolers, 1979; Kolers & Ostry, 1974). The perceptual operations that are specifically associated with unraveling the variations introduced by particular talkers could be modified to become more efficient. Although this account assumes that learning can affect encapsulated modular operations, it does preserve the speech-specific attributes of the modular account.

This account also preserves the distinction between voice recognition and linguistic processing. Evidence that this interpretation may be appropriate comes from studies investigating voice recognition in brain-damaged individuals (Van Lancker, 1991; Van Lancker, Cummings, Kreiman, & Dobkin, 1988). In a series of studies, Van Lancker and her colleagues have found that perception of the personal characteristics of speech appear to be subserved by the right hemisphere while linguistic processing appears to be localized in the left hemisphere. This anatomical separation predicts a functional dissociation which our data appear to contradict by demonstrating an effect of talker familiarity on linguistic processing. However, if learning voices results in a modification of perceptual compensation operations, then hemispheric differences in identifying a talker's voice and linguistic content could be preserved while at the same time perceptual learning of voice would be shown to have an impact on linguistic processing. Thus, if learning involves facilitation of the unraveling of linguistic and voice information rather than some type of combined, detailed representational system, for example, of indexical and linguistic properties, then the distinctions between the two tasks could be preserved.

If this account is correct, then what listeners are learning during our perceptual learning task is a fine tuning of normalization procedures. As they extract the talker-specific aspects of the signal for talker identification, they are also receiving practice or experience in extracting linguistic content as well. If linguistic processing is assumed to be mandatory, then the specific exposure each listener receives should be sufficient to facilitate linguistic processing. Although our investigation does not provide a direct test of this assumption, the performance differences between 'good' and 'poor' learners suggests that mere exposure to the range of variability found among the talkers is not adequate to produce any subsequent linguistic facilitation. Rather, it appears that it is the categorization training itself and the specific attention to talker-specific aspects of the signal that this training evokes that increases sensitivity to the linguistic content of familiar talkers.

An alternative to this view is to abandon the idea of speech specific processing altogether in favor of more integrated, interactive explanation (Nygaard & Kalish, 1994). In general, such an alternative account would assume that the extraction of talker information and the extraction of linguistic information are at the very least complementary processes and, more likely, comprise a single perceptual ability and perceptual system that is no different from the extraction of surface and object characteristics in other modalities (Fowler, 1986). The reason that familiarity with a talker's voice affects linguistic processing is then a result of a common underlying code for perception (Remez et al., in press) and common perceptual operations for the perception of voice and the perception of phonetic content of the signal. Listeners become highly skilled and attuned in recovering the consequences of dynamic vocal tract events. That is, listeners are assumed to be sensitive to time-varying information in the speech signal that result from articulatory gestures. Sensitivity to time-varying voice information, according to this account, would not be different in principle or kind from sensitivity to time-varying phonetic information. Sensitivity to one aspect of the speech signal would necessarily mean sensitivity to the other aspect. Any perceptual learning that increased distinctiveness or sensitivity to the dimension of talker identity would therefore increase sensitivity to linguistic aspects of the signal as well. The finding that general principles of perceptual learning (Gibson, 1969; Goldstone, 1994) operate within the domain of speech perception and spoken language lends support to this account of the relationship between talker identity and speech perception.

In summary, our findings demonstrate that perceptual learning of a talker's voice influences the intelligibility of spoken language. Familiar voices are more intelligible than unfamiliar voices and this difference suggests that the dimensions along which talker identity varies are the same dimensions that subserve linguistic processing. Our findings of a link in perceptual processing between the indexical and linguistic properties of speech constitutes one of the first demonstrations of the important role talker information plays in the perception of spoken language.

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Chicago, IL: Aldine Publishing Co.

Assmann, P.F., Nearey, T.M., & Hogan, J.T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America,* **71**, 975-989.

Bradlow, A.R., Nygaard, L.C., & Pisoni, D.B. (1995). On the contribution of instance-specific characteristics to speech perception. In C. Sorin et al. (Eds.), *Levels in Speech Communication: Relations and Interactions.* Elsevier Science. Pp. 13-24.

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

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch and B. Lloyd (Eds.), *Cognition and Categorization.* Hillsdale, NJ: Erlbaum.

Church, B.A., & Schacter, D.L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **20**, 521-533.

Cole, R.A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology,* **26**, 1-7.

Costanzo, F.S., Markel, N.N., & Costanzo, P.R. (1989). Voice quality profile and perceived emotion. *Journal of Counseling Psychology,* **16**, 267-270

Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology,* **26**, 274-284.

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

Doddington, G.R. (1985). Speaker recognition: Identifying people by their voices. *Proceedings of the IEEE,* **73**, 1651-1664.

Dupoux, E., & Green, K. (in press). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experiment Psychology: Human Perception and Performance.*

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

Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: MIT Press.

Fodor, J.A. (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics,* **14**, 3-28.

Garner, W. (1974). *The Processing of Information and Structure.* Hillsdale, NJ: Erlbaum.

Garvin, P.L., & Ladefoged, P.L. (1963). Speaker identification and message identification in speech recognition. *Phonetica,* **9**, 193-199.

Geiselman, R.E. (1979). Inhibition of the automatic storage of speaker's voice. *Memory & Cognition,* **7**, 201-204.

Geiselman, R.E., & Bellezza, F.S. (1976). Long-term memory for speaker's voice and source location. *Memory & Cognition,* **4**, 483-489.

Geiselman, R.E., & Bellezza, F.S. (1977). Incidental retention of speaker's voice. *Memory & Cognition,* **5**, 658-665.

Geiselman, R.E., & Crawley, J.M. (1983). Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning and Verbal Behavior,* **22**, 15-23.

Gibson, E. J. (1969). *Principles of Perceptual Learning and Development.* Appleton-Centrure-Crofts: New York.

Gibson, E.J. (1991). *An Odyssey in Learning and Perception.* MIT Press: Cambridge.

Gibson, J.J., & Gibson, E.J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review,* **62**, 32-41.

Goldinger, S.D. (1992). Words and voices: Implicit and explicit memory for spoken words. *Research on Speech Perception Technical Report No. 7.* Indiana University, Bloomington, IN.

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

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General,* **123**, 178-200.

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

Hall, G. (1991). *Perceptual and Associative Learning.* Clarendon Press: Oxford.

Halle, M. (1985). Speculations about the representation of words in memory. In V.A. Fromkin (Ed.), *Phonetic Linguistics* (pp. 101-114). New York: Academic Press.

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

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America,* **37**, 158-166.

Institute of Electrical and Electronics Engineers (IEEE) (1969). *IEEE Recommended Practice for Speech Quality Measurements* (IEEE Report No. 297). New York: Author.

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). New York: Academic Press.

Joos, M.A. (1948). Acoustic phonetics. *Language,* **24**, Supplement 2, 1-136.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America,* **88**, 642-654.

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

Kolers, P.A., & Ostry, D.J. (1974). Time course of loss of information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior,* **13**, 599-612.

Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics,* **50**, 93-107.

Kuhl, P. (1992). Psychoacoustics and speech perception: Internal standards, perceptual anchors, and prototypes. In L.A. Werner and E.W. Rubel (Eds.), *Developmental Psychoacoustics* (pp. 293-332). Washington: APA Press.

Labov, W. (1972). *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language,* **56**, 485-502.

Ladefoged, P., & Broadbent, D.E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America,* **29**, 948-104.

Laver, J. (1989). Cognitive science and speech: A framework for research. In H. Schnelle & N.O. Bernsen (Eds.), *Logic and Linguistics: Research Directions in Cognitive Science. European Perspectives, Vol. 2* (pp. 37-70). Hillsdale, NJ: Erlbaum.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer and H. Giles (Eds.), *Social Markers in Speech* (pp. 1-32). Cambridge: Cambridge University Press.

Lawrence, D.H. (1949). Acquired distinctiveness of cues: I. Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology,* **39**, 770-784.

Legge, G.E., Grossmann, C., & Pieper, C.M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **10**, 1-36.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition,* **21**, 1-36.

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

Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America,* **94**, 1242-1255.

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.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America,* **89**, 874-886.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 122-147). Cambridge, MA: MIT Press.

Markel, N.N., Bein, M.F., & Phillis, J. (1973). The relationship between words and tone-of-voice. *Language and Speech,* **16**, 15-21.

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**, 676-681.

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

Miller, J.D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America,* **85**, 2114-2134.

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

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

Murray, I.R., & Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America,* **93**, 1097-1108.

Nearey, T.M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America,* **85**, 2088-2113.

Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **15**, 700-708.

Nusbaum, H.C., Pisoni, D.B., & Davis, D.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: Indiana University.

Nygaard, L.C., & Kalish, M.L. (1994). Modeling the effect of learning voices on the perception of speech. *Journal of the Acoustical Society of America,* **95**, 2873.

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. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics,* **57**, 989-1001.

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

Peters, R.W. (1955a). The effect of length of exposure to speaker's voice upon listener reception. *Joint Project Report No. 44, U.S. Naval School of Aviation Medicine*, pp. 1-8. Pensacola, FL.

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

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.

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.

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.

Remez, R.E., Fellowes, J.M., & Rubin, P.E. (in press). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance.*

Samuel, A. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics,* **22**, 321-330.

Schacter, D.L. (1990). Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. In A. Diamond (Ed.), *Development and Neural Bases of Higher Cortical Functions*. (Annals of the New York Academy of Sciences, 608, pp. 543-571). New York: New York Academy of Sciences.

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

Shankweiler, D.P., Strange, W., & Verbrugge, R.R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (pp. 315-345). Hillsdale, NJ: Erlbaum.

Shepard, R.N., & Teghtsoonian, M. (1961). Retention of information under conditions approaching a steady state. *Journal of Experimental Psychology,* **62**, 302-309.

Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America,* **96**, 1314-1324.

Stevens, K.N., & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America,* **64**, 1358-1368.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics,* **36**, 131-145.

Summerfield, Q. (1975). Acoustic and phonetic components of the influence of voice changes and identification times for CVC syllables. *Report on Research in Progress in Speech Perception,* **2**, 73-98. Belfast, Northern Ireland: The Queen's University of Belfast.

Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of Speech Research in Progress,* **2**(2), 12-23. Queens University of Belfast.

Torretta, G.M. (1995). The "Easy -Hard" word multi-talker speech database: An initial report. *Research of Spoken Language Processing, Progress Report No. 20.* Speech Research Laboratory, Bloomington, Indiana. Pp. 321-334.

Van Lancker, D. (1991). Personal relevance and the human right hemisphere. *Brain and Cognition,* 17, 64-92.

Van Lancker, D., 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**, 829-854.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I. Recognition of backward voices. *Journal of Phonetics,* **13**, 19-38.

Van Lancker, D., Kreiman, J., & Wickens, T. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics,* **13**, 39-52.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America,* **60**, 198-212.

Weenink, D.J.M. (1986). The identification of vowel stimuli from men, women, and children. *Proceedings from the Institute of Phonetic Sciences of the University of Amsterdam,* **10**, 41-54.

Williams, C.E. (1964). The effects of selected factors on the aural identification of speakers. Section III of Report EDS-TDR-65-153, Electronic Systems Division, Air Force Systems Command, Hanscom Field.

Wohlwill, J.F. (1958). The definition and analysis of perceptual learning. *Psychological Review,* **65**, 283-295.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Assessing Speech Perception in Children[1]

Karen I. Kirk,[2] Allan O. Diefendorf,[2] David B. Pisoni,[2] and Amy M. Robbins[2]

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

[1] This work was supported by NIH NIDCD Grants DC-00064 and DC-00111 to Indiana University.
[2] Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Abstract

This chapter examines the assessment of speech perception abilities in children, with special emphasis on assessment in children with hearing loss. In the first section, a number of behavioral measures are discussed that are commonly used in clinical practice and that can be used to assess a hierarchy of skills in children with both normal hearing and with varying degrees of hearing loss. The second section considers the special case of assessing speech perception performance in children with cochlear implants and describes several assessment batteries that have been developed for use with this population. Finally, several new directions for assessment of speech perception performance in children are examined, including new measures specifically designed to examine the encoding, storage, and retrieval of spoken words from long-term lexical memory by children with hearing loss.

Assessing Speech Perception in Children

Introduction

Hearing is an extremely important sensory mode for many reasons, but most significantly, because a child's ability to develop and use oral language is closely related to the ability to process speech through his or her ears. The crucial role of hearing in spoken language development can be inferred by considering the language delay among children with bilateral hearing loss (Dale, 1974; Lach, Ling, & Ling, 1970; Quigley & Paul, 1984). Developmentally, the reception, perception, and subsequent comprehension of spoken language is a fundamental prerequisite to the verbal expression of language or speech. For many children with hearing loss, the task is difficult yet achievable through exploitation of residual hearing. In order for clinicians to enhance auditory-verbal language for children with hearing loss they must understand the limitations imposed by the hearing loss on auditory perception. When clinicians are faced with evaluating the auditory perception skills of a child with a hearing loss, they usually consider speech perception as the principal skill to be tested. Speech perception must be routinely assessed with a battery of tests so that the prognosis regarding the development of speech, language, reading, and cognitive skills can be made. Moreover, the best overall measure of construct validity is obtained by combining test results from several areas of speech perception performance.

The purpose of this chapter is threefold. First, we describe a number of speech perception measures that can be used to assess a hierarchy of skills in children with normal hearing and children with varying degrees of hearing loss. The tests to be presented here are commonly used in clinical practice to provide descriptive information concerning speech perception performance. Second, we consider the special case of assessing speech perception performance in children who use cochlear implants. Because all of these children are profoundly deaf, traditional clinical assessments may not be sensitive enough to the differences in speech perception performance among this group. We describe several speech perception batteries that have been developed for use with this population. Finally we consider several new directions for the assessment of speech perception performance in these children. Traditional tests have yielded descriptive information concerning speech perception abilities, but have provided little information concerning the underlying perceptual processes and how they may differ between listeners with normal hearing and those with hearing impairment. We describe the development of new speech perception measures specifically intended to examine the way in which children with hearing loss encode, store and retrieve words from long-term lexical memory.

Several assumptions underlie the ideas expressed in this chapter. First, it is assumed that speech perception assessment is an essential feature of a comprehensive audiologic evaluation. Second, information from speech perception testing is fundamental to planning (re)habilitation (including the placement and monitoring of sensory devices) and educational strategies for young children with hearing loss. Finally, assessment of the pediatric population differs substantially from older children, adolescents, and adults and therefore the content validity of speech perception assessment tools with the pediatric population must reflect these differences. We will survey tests that are consistent with our assumptions, and present factors to consider in the administration and interpretation of test outcomes.

Considerations in Pediatric Speech Perception Assessment

Because speech represents the class of sounds most important to the effective daily function of humans, tests utilizing speech stimuli are essential for the evaluation of children with communication

deficits. Determination can be made of the extent to which a hearing loss affects the ability to perceive, recognize, and discriminate speech. Such information is useful not only in the diagnosis of the type and severity of the hearing disorder but also in the approach to prognosis and the monitoring of aural rehabilitation efforts.

Clinicians concerned with assessing speech perception performance in the pediatric population must consider a number of variables, both internal and external, that can influence test outcome and affect the reliability and validity of pediatric speech perception procedures. Internal or subject factors include such things as the child's vocabulary and language competency, chronological age, and cognitive abilities. External factors, such as the designation of an appropriate response task, the utilization of reinforcement, and the reduction or alleviation of memory load inherent in the task, can also affect speech perception performance. Both types of factors must be considered when selecting specific assessment procedures and when interpreting test results. For example, children with either a congenital or early-acquired hearing loss are often delayed in the development of speech and language skills and have restricted vocabularies. This situations presents difficulty in assessing speech perception because children will not be able to recognize unfamiliar words. Therefore, special care must be taken to select tests that are linguistically suitable for young children and to familiarize them with the test items, if necessary. Similarly, when testing young children, one must consider the nature of the behavioral task employed, and the child's ability to participate in testing. As pointed out by Robbins and Kirk (in press), speech perception cannot be directly measured, but only inferred from a child's responses. If the child does not have the cognitive ability to understand a task, or the child is unable or unwilling to attend, then there may be a gap between the observed performance and the child's actual speech perception abilities. We consider some additional factors below.

Taped vs. Live Voice Presentation

The use of recordings versus monitored live voice for the assessment of speech perception has been widely debated. Proponents of recorded materials have stressed that speakers differ and therefore, results are not comparable among clinics or laboratories unless speaker equivalence can be demonstrated. Clinicians have argued that only through use of tape-recorded test materials can consistency in presentation be maintained from one listener to the next or for repeated assessment of the same individual. Yet, there may be as much difference between one recording and another as between two live-voice talkers. These differences suggest the need for standardized recorded versions of the speech perception materials. Those favoring use of monitored live voice presentation of test materials highlight the greater flexibility afforded by this method and also cite reports in the literature demonstrating satisfactory test-retest reliability.

Open-Sets Tests vs. Closed-Set Tests

Open-Set Tests. Open-set tests are those in which the listener theoretically has an unlimited number of response possibilities. On hearing the test item, no response alternatives are given and the listener is free to make any response. Open-set tests are not appropriate for all children. This is true for some listeners because they may lack the ability to give oral responses, for others because they are unwilling or too shy, and for still others because their speech production is so poor that responses cannot be discriminated. For these reasons, a number of closed-set tests have been developed for children.

Closed-Set Tests. Closed-set tests restrict the listener to one of a fixed number of possible responses (e.g., as in a multiple-choice test). One advantage of closed-set tests is that they can be administered without requiring the listener to make either a spoken or a written response, and therefore they are useful for individuals who cannot speak or write well enough to perform with open-set tests. To bypass

these problems and still test speech discrimination ability with young children, picture discrimination tests may be used. The use of pictures rather than printed words adapts the test to those who cannot read. Because the number of potential responses is limited, closed-set tests are easier and yield higher scores than an open-set (unlimited) procedure. Closed-set speech perception performance may be affected by the number of available alternative responses. The amount by which any score may be expected to deviate from the true score due to the effects of the closed set is a function of the number of foils per item. By limiting the number of alternatives with a closed-set test, the listener's need to store and retrieve target messages independently from lexical memory is reduced. Therefore, a disadvantage to closed-set tests is that they may not adequately represent a listener's performance in natural situations.

Task Domain in Closed-Set Test Construction: Sensory vs. Cognitive Contributions

Factors that interfere with the processing of speech from a sensory perspective include competing noise (Johnson, Cosci, Brown, and Scroggins, 1995; Crandell, 1993), reverberation (Nabelek and Robinson, 1982; Nabelek and Mason, 1981), and listener distance from the sound source (Crum, 1976). A sensory perspective on assessing speech perception abilities recognizes a physical (bottom-up) auditory analysis, triggered by the arrival of sensory information to the ear. Noise, reverberation, and distance degrade the primary acoustic speech signal, making spoken word recognition more difficult. Moreover, individuals with hearing loss who sustain physiologic limitations are at greater risk when perceiving speech in difficult listening environments.

Diminished speech perception can result from masking effects and distortion in the auditory-peripheral hearing mechanism. Speech signals which are acoustically/phonetically similar are also difficult for an individual with hearing loss to discriminate due to the frequency, amplitude, and temporal limitations imposed on the auditory-peripheral hearing mechanisms as a result of the hearing impairment.

Unrestricted Task Domain. A task domain is considered unrestricted when target signals are not uniquely specified for the listener, but are embedded among foil items representing selected phonemic confusions. An example would be the Word Intelligibility by Picture Index (WIPI) (Ross & Lerman, 1979) which presents six phonetically similar foils from which the child must select one target item. When items on a test of speech discrimination are embedded among foils that are acoustically/phonetically similar, and when the response set changes for each stimulus presentation (as in the WIPI), children cannot use a "process of elimination" to select a response. Instead, closed-set tasks that utilize an unrestricted task domain are tapping sensory capabilities. The listener with limited information is unable to identify the target on the basis of a conceptually generated "gestalt" and must identify the target in a bottom-up processing strategy.

Restricted Task Domain. While an unrestricted task domain promotes a sensory processing outcome perspective, a restricted task domain promotes a cognitive (conceptual) processing outcome perspective. Cognitive (top-down) processing is facilitated by an individual's general knowledge (including language and vocabulary level) and by expectations based on the context of the sensory event. When a closed set is provided to a listener and the target item stands alone (either because it is phonetically dissimilar from the foils, or because all foils serve as a target and the item can be selected by a process of elimination), the task domain is referred to as restricted. A restricted task domain specifies for the listener what he or she may expect to hear. Even when limited acoustic/phonetic information is available, the listener may be able to complete a conceptual "gestalt," thereby increasing the chance for correct identification.

Some clinicians criticize restricted task domain paradigms in speech discrimination tests and argue that restricted targets may be discriminated on the basis of cognitive abilities in contrast to sensory abilities. Clinicians oriented toward the conceptual processing of speech may prefer restricted task domains with a limited number of target items. Others may lean toward the acoustic/phonetic processing of speech and prefer unrestricted task domains with appropriately selected foil items. The outcome from both approaches can lead to different estimates of an individual's speech perception abilities and help explain why individual differences exist between speech tests that utilize different task domains.

Speech Perception Materials for Use with Children

Open-Set Tests

Monosyllabic Word Lists. Efforts to develop materials for testing speech perception in children date back to the late 1940s. Haskins (1949) developed phonetically balanced (PB) lists composed of monosyllabic words selected from the spoken vocabulary of kindergartners (PB-K). The monosyllabic words incorporated in the PB-K lists were selected on the basis of the International Kindergarten Union vocabulary lists.

The PB-K lists of 50 words each have been used widely with children. Yet, the receptive vocabulary level of a particular child under study is often not ascertained before administering these materials. Clearly, test items must be in the vocabularies of the children tested. If not, the PB-K scores may be depressed and may reflect language and/or vocabulary deficits as well as problems in auditory perception. Based on the findings from Sanderson-Leepa and Rintelmann (1976) which indicated that normal-hearing preschoolers at 3 1/2 years of age obtain substantially lower scores than older children, it is recommended that clinicians exercise caution in administering the PB-K test unless there is relatively good assurance that the receptive vocabulary age of the youngster approaches at least that of a normal-hearing kindergartner.

A similar approach was taken in England by Watson, resulting in the generation of the Manchester Jr. (MJ) lists. These materials consist of four lists of monosyllabic phonetically balanced words considered to be within the vocabulary of children who are 6 years of age or older (Ewing, 1957).

Sentence Lists. Bench, Kowal, and Bamford (1979) have developed a sentence test emphasizing its use for children with hearing loss. The authors advocate that the use of sentences rather than unconnected words gives a more valid indication of how a child copes during natural communication with others. The Standard BKB Sentence lists consist of 21 syntactically and semantically equivalent sentence lists, each containing 16 sentences. Each list contains 50 "key" words which are scored for accuracy (e.g., "The dog sleeps in a basket") to derive a percent-correct score. An alternate form utilizing pictures was developed for use with children who may need additional contextual information while listening to speech materials. There are 11 lists of picture-related BKB sentences; each containing 16 sentences and 50 "key" words.

The impetus for the development of the Bamford-Kowal-Bench Sentence Lists for Children (BKB) came from the authors' interest in working with children with hearing loss. Language samples were taken from over 240 children aged 8 to 15 years with hearing loss. The basic approach was to use vocabulary and grammar which were familiar to these children. Children were asked to describe a set of colored drawings which depicted scenes and events from commonplace play or family environments. For vocabulary, the minimum requirement was that any word used in sentence list construction had to appear at

the phrase level, and the most frequent morphological structures were allowed at the word level. Moreover, sentence length could not exceed 7 syllables. Although the BKB lists were developed using British English vocabulary, an Americanized version of the test has been developed by Kenworthy, Klee, and Tharpe (1990). In the Americanized version of the BKB stimuli, the British English words were converted to their equivalent standard American English forms.

In an attempt to validate the BKB Standard Sentence Test, lists of key words from the same sentences were compared. That is, 50 "key" words from the same selected sentence lists were presented one at a time. Thus, the scored words were the same, but in one condition they were embedded in a linguistic context, in the other they were isolated units. In both cases, the method of scoring was to measure the percentage of correctly repeated key words for each list. Sixteen children with hearing loss (range: 30 to 80 dBHL ISO; averaged over .5, 1 and 2 kHz), age 11 to 13 years were evaluated.

It was hypothesized that the performance-intensity functions for the sentence test should exhibit steeper slopes than those from words presented in isolation. The authors suggest that the steeper the slope of the performance-intensity function, the more valid is the material for assessing hearing for natural speech. Results demonstrated that the slopes for the sentences were significantly steeper than the slopes for the randomized words. The authors conclude that the BKB Standard Sentence Test is appropriate for assessments which relate to natural listening conditions.

Closed-Set Tests: Unrestricted

Monosyllabic Word Lists. Because many children with hearing loss do not have receptive language skills approximating those of a 6 year old child with normal hearing, alternate speech discrimination materials are available for clinical use. The Word Intelligibility by Picture Identification (WIPI) test was developed by Ross and Lerman in 1970 and takes into consideration children who have restricted receptive vocabulary and cannot read. The WIPI test includes picture plates with six illustrations per plate. Although the four lists of 25 monosyllabic words are not phonetically balanced, an attempt was made by the authors to assure that different distinctive features were included in the selection of words. Several of the illustrations that serve as foils are words that rhyme, and other illustrations have phonetic/acoustic similarities. This test construction is consistent with an unrestricted task domain.

On administering the WIPI to 61 children with hearing loss (mean=52 dB; 1964 ISO standard) between the ages of 4 years and 13 years, Ross and Lerman concluded the test-retest reliability was high (reliability coefficients ranged from .87 to .94), and equivalence of the four lists was high (standard error of measurement ranged from 4.7 to 7.7). Moreover, the authors concluded that the WIPI test is suitable for children with moderate hearing losses from ages five or six and for children with severe hearing losses from ages seven or eight. The results reported by Schwartz (1971) and Sanderson-Leepa and Rintelmann (1976), demonstrated that normal hearing children at age 3 1/2 manifest a significant number of errors due to words not in their recognition vocabulary. Thus, the use of WIPI materials is appropriate for those children with receptive vocabulary ages of 4 years and greater.

Hodgson (1973) investigated the relationship between the WIPI words used as an open-set and as a closed-set test for children with normal hearing. In the open-set condition, subjects simply repeated the WIPI words. The closed-set response required the usual picture identification task. Subjects also repeated the words of a PB-K list administered in the conventional open-set fashion. Although there was no difference in the intelligibility of the WIPI and PBK words, use of the WIPI as a closed-set test improved

the discrimination scores by about 10 percent. Therefore it is advisable that the test record include not only the scores obtained but also the particular response format used.

Jones and Studebaker (1974) compared open- versus closed-set test formats using a group of children with hearing loss (pure-tone average re: ANSI 1969: 60-110+ dB; mean=88 dB). They found that the closed-set test paradigm appears more productive for use with severely hearing impaired children whose level of performance is low, but not 0 percent. Jones and Studebaker concluded that data from closed-response sets tend to demonstrate auditory speech discrimination difficulties in a more satisfactory way than data from open-response sets; when used with children who have severe-to-profound hearing losses, open-set tests frequently yield performance scores near the test floor, making it difficult to characterize certain aspects of the children's speech perception abilities.

The Northwestern University-Children's Perception of Speech (NU-CHIPS) test by Elliott and Katz (1980) was developed as a speech discrimination test appropriate for young children. Test materials are limited to monosyllabic words that are documented to be in the recognition vocabulary of children with normal hearing as young as three years. The selection process identified only 67 monosyllabic nouns, represented by pictures, that were within the receptive vocabularies of three-year-old children. Words that were extremely easy or extremely difficult to identify were eliminated and the remaining 50 monosyllabic nouns were selected to constitute the test items. The NU-CHIPS Test is designed to utilize a closed-set, "picture-pointing" response. The test items are representative of the most frequently occurring phonemes of English, with the exception of initial /r/. Four pictures appear on each page of a picture book. Picture foils are as phonemically similar to the test word as is possible within the constraints of the original 67 words.

A number of the statistical analyses for NU-CHIPS were made at sensation levels where performance was below the test ceiling in order to examine test sensitivity. Performance of normal children was evaluated at 6 dB SL while performance for children with hearing loss was evaluated at 12 dB SL in order for the scores to be in approximately the same range. Utilizing this strategy, the four test lists, or forms, demonstrated equivalent means and equivalent variances, thus revealing good reliability. Additionally, inter-test form correlations ranged from .83 to .92 for a group of ten 5-year-old children with normal hearing tested at SL's of 0- and 2 dB.

The validity of the NU-CHIPS was assessed by the authors who looked at performance scores at two receptive vocabulary levels as a function of hearing level. Validity, as assessed, is considered "good." Regression equations in which listener's characteristics such as pure tone sensitivity, chronologic age, and vocabulary level were used to "predict" NU-CHIPS performance, and showed pure tone air conduction sensitivity at one or more frequencies to be a significant "predictor" of the performance of children with hearing loss. Thus, the test demonstrates construct validity. Children with hearing loss and a receptive language of at least 2.6 years (as measured by the Peabody Picture Vocabulary Test [Dunn and Dunn, 1981]) demonstrate familiarity with the words and pictures of the test.

Children with language skills better than the target group for which NU-CHIPS was developed achieve higher scores on NU-CHIPS than on the WIPI Test (Elliott and Katz, 1980). These findings are expected on the basis of the somewhat more difficult vocabulary on the WIPI Test.

Children of different ages with normal hearing demonstrated ceiling effects at sensation levels up to 30 dB. At sensation levels lower than 30 dB, an age effect was demonstrated in that ten-year-olds performed better than five-year-olds who in turn performed better than three-year-olds. The likely

explanation for the age effect observed at low sensation levels is language skill or experience. Because of this developmental effect, it is recommended that NU-CHIPS be administered at 30 dB SL.

Siegenthaler and Haspiel (1966) developed the Discrimination by the Identification of Pictures test (DIP). This measure was designed for young children with pictures selected for easy recognition by preschoolers. The stimuli consist of 48 pairs of familiar monosyllabic words incorporating differences in phoneme distinctive features. The pictures are arranged in pairs where the initial consonants differ by the features of voicing, place of articulation, pressure pattern, or a combination of these features. However, no attempt was made to achieve a phonemic balance with the stimulus words. A major drawback with the DIP is that only two illustrations are pictured on each response plate and thus, chance performance is 50 percent.

DIP test scores vary as a function of age. Siegenthaler (1975) evaluated children with normal hearing between the ages of 3-8 years. The DIP Test was evaluated at SRT plus 5 dB and 10 dB respectively. Scores ranged from 52% for the 3-year-old subjects to 74% for the 8-year-old subjects. DIP test reliability calculated across all ages showed significant test/retest correlations indicating good reliability. The overall standard deviation was 4.84 items (10%) which can be taken as the standard error of estimate for the test. For children with hearing loss between the ages of five and 12 years, the discrimination test has acceptable reliability. If a single test level for children is to be used, Siegenthaler recommends SRT plus 10 dB in view of the high reliability and test scores below the maximum possible.

Environmental Sounds. Finitzo-Hieber, Gerling, Matkin, and Cherow-Skalka (1980) investigated the use of familiar environmental sounds as an alternative to speech stimuli for assessing auditory discrimination abilities of children not capable of differentiating verbal stimuli. A measure of sound effects recognition may serve to categorize children in terms of their ability to utilize nonlinguistic auditory information in daily listening environments. The Sound Effects Recognition Test (SERT) incorporates a closed-set format with a picture-pointing response. The SERT is a standardized approach to using environmental sounds as a measure of auditory recognition and perception.

The SERT is comprised of three equivalent sets, each containing 10 familiar environmental sounds (30 different sounds, most of which are broad-band in spectral content). The forms are equivalent in terms of item-difficulty for children between the ages of 3 years, 0 months and 6 years, 6 months. The test contains four pictures on each of 10 response plates. Three of the pictures are foils with one target item per page. Each target item is also used as a foil on one other response plate. The test will be too easy for a large number of children with hearing loss who have measurable speech discrimination. However, the SERT should be considered when other tests using verbal materials are inappropriate. The authors report that by age 3, a child should be able to identify an average of 25 to 30 environmental test sounds, and by age 5, a mean score of 29 of the 30 test sounds is obtained. Performance-intensity functions reveal that the sensation level needed to achieve maximum sound recognition performance (100%) is achieved with a presentation level of 25 dB relative to the speech awareness threshold for both 6-year-old and adult subjects. Obviously, the presentation level may vary depending upon the child and the purposes of the test.

As part of a pediatric test battery, the SERT may be a viable indicator of a child's auditory capabilities at suprathreshold levels, thus providing clinicians with information about auditory function in youngsters with very limited verbal abilities. Moreover, for children with extremely restricted language competence, for whom a test to monitor auditory development over time is desirable, the SERT can prove useful.

Closed-Set Tests: Restricted

Monosyllabic Word Lists. The Auditory Numbers Test (ANT) is a spectral/pattern perception test that was developed by Erber (1980) for use with children who have severe speech discrimination difficulties. The purpose of developing this test was to establish whether young children with hearing loss are able to perceive the spectral qualities of sounds or only intensity patterns. The ANT utilizes five colored picture cards depicting groups of one to five ants with the corresponding numerals. A child receives one point for each number identified correctly. Thus, a score can range from zero to five, although a score of zero is unlikely if the training procedure is completed successfully.

The ANT procedure is designed to elicit one type of response from children who can hear minimal spectral cues in the test words (e.g., correct identification of each stimulus number), and a different response from children who can perceive only gross intensity patterns in the acoustic stimulus (e.g., "one" reported regardless of which number is the stimulus). The ANT requires only that the child be able to count to five and be able to apply these number labels to sets of from one to five items. Thus, children tested must comprehend the words representing the numbers one through five.

Erber reports outcome data for 39 children ages 3 through 8 years. The following conclusions were drawn from this sample: first, the ANT can be used with children who sustain severe-to-profound hearing losses and are able to demonstrate appropriate comprehension of the numbers one through five; second, a high score (e.g., 3-5) on the ANT implies that a child is capable of perceiving at least some of the spectral qualities of words; third, a low score (e.g., 1) on the ANT suggests that a child's perception is primarily of speech intensity patterns. Because a low score could be the result of a number of other factors however (e.g., insufficient previous auditory experience, poor attention, lack of motivation), repeated testing with the ANT or other appropriate materials is required before clinical/prognostic conclusions can be drawn.

The Pediatric Speech Intelligibility (PSI) test was developed by Jerger, Lewis, Hawkins, and Jerger (1980) for children as young as 3 years of age for evaluating both peripheral and central components of auditory disorders. Both monosyllabic words and sentence materials were generated by normal children between 3 and 7 years of age. Initially, 30 monosyllabic words were chosen to represent an array of English phonemes in the initial and final positions. The responses to the 30 words by children sampled did not differ as a function of chronological age, vocabulary skill, or receptive language ability (Jerger et al., 1980). Yet for 20 words, a correct response was observed in more than 95% of the children. Therefore, only those 20 words were selected for the PSI test. The 20 words are depicted on four response plates (five pictures per plate), and the child uses a picture-pointing response.

Jerger, Jerger, and Lewis (1981), investigated the influence of receptive language on PSI monosyllabic words. Word identification was compared in two groups of children who differed by chronologic age (mean: 4-6 versus 5-11) and receptive language age (Northwestern Syntax Screening Test [NSST; Lee, 1971] mean: 4-2 versus-6-3). Even though the two groups differed significantly, average performance for PSI word materials was not affected. Thus, word materials generated by normal children between 3 and 7 years of age do not differ as a function of chronologic age and receptive language skills. However, average performance for PSI word materials increases as age increases from 3 to 6 years by approximately 10% (Jerger et al., 1981).

In order to establish the reliability of the PSI monosyllabic words, normal hearing children and children with hearing loss were tested on two separate occasions. On both occasions, PSI performance was measured at the same SPL and the same message-to-competing ratio. The reliability of the PSI words, as

determined by correlation coefficients between test-retest measures, was high (.92) in both groups of children tested. Additionally, no significant differences were demonstrated among the four lists of monosyllabic words. Moreover, reported variances suggest that the four lists are homogeneous.

Sentence Lists. The PSI sentence materials incorporate the consistent differences in vocabulary and receptive language function that characterize children between the ages of 3 to 6 years. To represent the different sentence patterns produced by the children sampled, two 10-sentence lists were formed using two difference syntactic constructions. Format 1 consists of the sentence construction "article noun/verb-ing/article noun" and sentences are preceded by the carrier phrase, "show me," such as "Show me a rabbit reading a book." For Format 1 materials, excluding the carrier phrase, the average sentence length is 7 syllables. Format 2 consists of the sentence construction "article noun/auxiliary verb-ing/ article noun," and sentence materials are not preceded by carrier phrase (e.g., "A bear is eating a sandwich."). For Format 2 materials, average sentence length is 8 syllables. The response format for the sentences is similar to that for words, in that the child selects his answer from a five-picture plate. For Format 1 sentences, the correlation coefficient between the two lists is .75; for Format 2 sentences, the correlation coefficient between the two lists is .7. There were no significant inter-list differences for either Format 1 or 2.

In the development of the PSI materials, meaningful sentences were selected for two reasons: First, meaningful material is easier to code, store, and retrieve than non-meaningful messages; second, Jerger et al., preferred to utilize a closed message set rather than an open message set to reduce linguistic, conceptual, and developmental variables that may contaminate response outcomes.

Jerger, Jerger, and Lewis (1981) investigated the influence of receptive language on PSI sentence materials. For children older than 7 years with receptive NSST language scores of about 37 to 40, PSI sentence materials seem inappropriate. In these youngsters, performance remains too high even in the presence of a competing message. For younger children, with NSST scores of 15 to 36, both Format 1 and Format 2 sentence constructions appear necessary. In children with somewhat higher receptive language scores (32 to 36) Format 1 sentences are too easy, thus Format 2 sentences are appropriate. In children with relatively low receptive language scores (15 to 31) Format 2 sentences are too difficult, thus suggesting the need to use Format 1 sentences. The different sentence formats are an effective means of controlling the influence of receptive language ability on speech intelligibility performance for sentences in young children.

In order to establish the reliability of the PSI sentence material, normal hearing children and children with hearing loss were tested on two separate occasions. On both occasions, PSI performance was measured at the same SPL and the same message-to-competing signal ratio. The reliability of the PSI sentences, as determined by correlation coefficients between test-retest measures, was high, ranging from about .82 to .96 in both groups of children tested.

The tests described above are intended to assess speech perception abilities in children with varying degrees of hearing loss. However, relatively few of them (e.g., the SERT and the ANT) were intended for use with profoundly hearing-impaired children who may have minimal auditory capabilities. With the advent of cochlear implants, it has become increasingly important to develop speech perception materials that are sensitive to the individual differences in auditory skills in children with profound loss (see Boothroyd, 1993). In the next section we describe two batteries of tests that were specifically developed to assess speech perception in children with profound hearing loss. The tests within these batteries do not constitute an exhaustive description of the test instruments available for this population (see Tyler, 1993).

However, they were selected because they are among the most commonly used to assess performance in children with cochlear implants or other sensory aids.

Assessing Speech Perception in Children with Cochlear Implants

The assessment of speech perception performance in children with cochlear implants has important implications, both for the clinical management of the children and for evaluating the efficacy of these relatively new medical devices. The evaluation of children's spoken word recognition and speech perception performance is clinically relevant because it allows the clinician to monitor progress following implantation; this in turn, has implications for setting or "mapping" each individual child's cochlear implant signal processor, as well as for determining appropriate auditory training goals (see Robbins and Kirk, in press). From a research standpoint, assessing speech perception skills in children with profound hearing loss who use a given sensory aid (i.e., hearing aids, tactile aids, or cochlear implants) allows clinicians and researchers to compare the effectiveness of these devices; this impacts on issues of cochlear implant candidacy. Such assessments may also provide insight into the considerable variability in speech perception performance that is noted in both children and adult cochlear implant recipients. In both populations, the benefits of cochlear implant use range from lipreading enhancement through open-set speech understanding (Osberger et al., 1991a). Therefore, it is important that the perceptual measures used with these children are sensitive to individual differences among performance.

Because of the wide range of variability in performance noted, it is typical for clinicians and researchers to employ a battery of tests intended to assess a wide range of speech perception abilities. For example, closed-set tests may be used to assess the perception of prosodic cues, speech-feature discrimination, or even word identification. Such tests yield descriptive information concerning a child's perceptual skills and also information about the speech cues that are conveyed by a given implant device, or other sensory aid (i.e., hearing aids or tactile aids). Open-set tests, on the other hand, may be used to assess both word and sentence comprehension, and may be more reflective of real-world performance.

One drawback to many of the current closed- and open-set tests is that data are lacking concerning their test-retest validity and reliability. Without reliability data, it is difficult to determine whether changes in performance over time on a given test are meaningful. The use of a battery of tests can help alleviate this problem because the consistency of performance across measures can be assessed. Below we describe several speech perception batteries designed for use with profoundly hearing-impaired children. Relevant validity and reliability data are presented where available.

Two general approaches have been followed in the development of a speech perception assessment battery for children with profound deafness who use cochlear implants, tactile aids, or hearing aids. In the first approach, espoused by Geers and Moog (1989) at the Central Institute for the Deaf (CID), children are assumed to acquire speech perception skills in a hierarchical fashion ranging from simple detection through spoken word comprehension (see Erber 1982); testing proceeds accordingly, and children are required to reach criterion scores at each level before being administered more difficult measures. The outcome of this testing is used to categorize the children's speech perception abilities and determine auditory training goals. In the second approach, no a priori assumptions are made concerning the sequence of auditory skill development. Rather, children are administered a battery of tests evaluating a range of speech perception abilities, and are assigned scores for each test in the battery. The speech perception batteries employed in the Cochlear Implant Program at Indiana University School of Medicine exemplify this latter approach. Both the CID and the Indiana University (IU) speech perception batteries have been used extensively with profoundly hearing-impaired children to provide comparative data on the relative

benefits of cochlear implants, tactile aids, and hearing aids. Furthermore, reliability and/or validity data are available for some of the tests within these batteries. We will first present a brief description of these two batteries and then summarize the results to date.

The CID Speech Perception Test Battery

The CID speech perception battery, as described in detail by Geers (1994), evaluates the auditory-only speech perception performance of children with profound hearing loss based on 7-point classification scale, with "0" representing no speech detection ability and "6" representing open-set word recognition (see Table 1). It also includes tests to assess lipreading enhancement (i.e., the improvement in lipreading ability in the auditory-plus-visual modality compared to that in the visual-only modality). Tables 2 and 3 list the auditory-only and lipreading enhancement tests, as summarized by Geers (1994).

Insert Table 1 about here

Speech Detection. Aided speech detection thresholds are first determined for each child. Children who cannot detect speech at conversational levels (i.e., 65 dB HL) are assigned to Category 0 and testing is discontinued. Children with aided speech detection thresholds of less than 65 dB HL proceed with speech perception testing.

Insert Table 2 about here

The CID Early Speech Perception Test (ESP) (Moog & Geers, 1990). The ESP is used to assess the closed-set perception of single words through listening alone. According to the authors, the test is intended for use with young profoundly hearing-impaired children who have limited vocabulary and language skills (i.e., those under 10 years of age [Geers & Moog, 1989]). Portions of the ESP were based on earlier tests developed for profoundly deaf children, including the Monosyllable-Trochee-Spondee Test (MTS) (Erber & Alencewicz, 1976), the Auditory Numbers Test (ANT) (Erber, 1980), and the Glendonald Auditory Screening Procedure (GASP) (Erber, 1982). The selection of test items was guided by three criteria: first, the words used should be familiar to most hearing-impaired children by age six years; second, the words should be suitable for representation in picture form so that children who cannot read can be tested; and third, the test should be administered in less than 20 minutes. Both a standard version and a "low-verbal" version of the ESP are available.

Administration of the ESP begins with stimulus familiarization in an auditory-plus-visual modality. If a child does not know the target vocabulary on the standard version, the low-verbal version is used. Stimuli may be presented either live-voice or in a recorded mode. Both an audiocassette version and a computerized version of the ESP are available. It is recommended that live-voice testing be carefully monitored to ensure that stimuli are presented at approximately 70 dB A, and that the stress patterns are accurately conveyed. If live-voice testing is employed, the authors recommend that the clinician first practice presenting the stimuli through a bone oscillator until a listener can identify stress patterns with 100 percent accuracy.

Table 1

CID Speech Perception Categories (Geers, 1994).

Category	Speech Perception Skills
0	No detection of speech (e.g., aided speech detection threshold >65dB HL)
1	Speech detection
2	Pattern perception (discrimination based on temporal or stress cues; e.g., airplane vs. baby)
3	Beginning word identification (closed-set word identification based on phoneme information; e.g., airplane vs. lunchbox
4	Word identification via vowel recognition (closed-set word identification based on vowel information; e.g., boat vs. bat)
5	Word identification via consonant recognition (closed-set word identification based on consonant information; e.g., pear vs. chair)
6	Open-set word recognition (word recognition without contextual cues through listening alone

Table 2

The CID Auditory-Only Speech Perception Battery (Geers, 1994).

Test	Stimulus	Response Format	Perceptual Skill
Speech Detection Threshold	Speech	Closed-set	Detection
ESP[a]	Patterns 1-, 2-, 3- syllable words	Closed-set	Word ID (stress & durational cues)
	Spondees	Closed-set	Word ID (spectral cues)
	Monosyllables	Closed-set	Word ID (vowel cues)
WIPI[b]	Monosyllables	Closed-set	Word ID (consonant cues)
Matrix Test[c]	Phrases	Closed-set	Word ID in phrases
Phonetic Task Evaluation[d]	Syllables	Closed-set	Speech feature discrimination
PBK[e]	Monosyllables	Open-set	Word recognition

[a]Moog & Geers, 1990; [b]Ross & Lerman (1979); [c]Tyler & Holstad (1987); [d]Mecklenburg, Shallop, & Ling (1987); [e]Haskins (1949)

There are three hierarchical subtests on the ESP: Pattern Perception, Spondee Identification, and Monosyllable Identification (see Table 2). The standard version depicts target words using a 12-picture plate (one for each subtest). The low-verbal version uses actual objects rather than pictures, and children select their response from a closed set of four objects. Children must score at significantly higher than chance levels to be given credit for the perceptual skill being tested. Scores of approximately 70-75% on a subtest indicate that the next subtest in the hierarchy should be administered. Scores on the ESP may be assigned to Categories 2-4 in Table 1. If a child scores 75% or higher on the Monosyllable Identification Subtest, their speech perception abilities exceed that of Category 4 (word recognition through vowel identification), and additional testing must be carried out.

Reliability and validity data for the recorded version of the ESP were reported by Geers and Moog (1994). Subjects were drawn from a pool of 49 profoundly hearing-impaired students at the CID. For the standard version of the ESP, 27 children between 8-15 years were tested and re-tested over a 30-day period. Test-retest reliability ranged from .78 for Pattern Perception to .94 for category placement. For the low-verbal version, reliability data were obtained from 24 children aged 4-6 years. Reliability ranged from .75 for Pattern Perception to .89 for category placement. The validity of the ESP standard version was measured by correlating the word-identification performance of 30 children who scored at Category 3 or 4 with their performance on the WIPI. The resulting validity correlation coefficient was .87, suggesting that results across the two measures are consistent. Finally, 26 children were administered both the standard and low-verbal versions of the ESP to assess the validity of category placement. For 24 children, administration of both versions yielded the same category placement.

Additional Word Identification Tests in the CID Battery. The two remaining tests in the CID speech perception battery are the WIPI (Ross & Lerman, 1979) and the PB-K (Haskins, 1949). Both of these tests have been described earlier. The WIPI and PB-K are administered in the standard fashion, but the scores are used for placement in a speech perception category. Children with WIPI scores of at least 28% are considered to have achieved Category 5 speech perception performance (word recognition through consonant identification) and are then administered the PB-K to assess open-set word recognition. Children who can identify at least three PBK words are placed in Category 6, and are considered to understand some conversational speech without lipreading (Geers, 1994). This relatively lax criterion may make it difficult to predict open-set speech understanding under more realistic conditions.

Speech Feature Test. The Phonetic Task Evaluation (PTE) (Mecklenberg, Shallop, & Ling, 1987) is used to estimate a child's speech feature discrimination abilities. This test contains recordings of a male talker producing pairs of nonsense syllables differing in vocal pitch, manner or place of articulation, voicing, or vowel features. Test administration involves a familiarization phase in the auditory-plus-visual modality. Actual testing may be carried out as an identification task (six alternatives) or as a "same-different" discrimination task (two alternatives) depending on the child's ability to imitate the syllabic stimuli. For the CID battery, scores are reported as the percent above chance.

Sentence Perception Test. The MATRIX Test (Tyler & Holstad, 1987) is used to assess the perception of key words in sentences, and is intended for children as young as 4-6 years. The test derives its name from the response cards which depict key words using columns of pictures. The child identifies the target sentence by repeating or by pointing to one picture per column. The test has two levels of difficulty. Level A presents 3-word sentences using a 2 X 3 matrix, and Level B presents 4-word sentences using a 4 X 4 matrix. The test is typically administered via live voice, and is scored as the percent of key words correctly identified. No reliability or validity data have been reported for this sentence test.

The CID Lipreading Enhancement Battery. Lipreading enhancement is the degree to which lipreading improves when audition is added. The CID battery (see Table 3) is based on the assumption that measuring enhancement is difficult if the baseline scores in the lipreading-only modality are either at the top or the bottom of the range of scores. The stimulus materials in the battery are hierarchical in terms of linguistic content and task demands; the assessment of lipreading enhancement is made with the measure that yields a baseline lipreading-only score between 40-75% (Geers, 1994). At the lower levels, lipreading enhancement is assessed using closed-set materials. The Grammatical Analysis of Elicited Language–Pre Sentence Level (GAEL-P) (Moog, Kozak, & Geers, 1983) was originally developed as a measure of receptive and expressive language skills. The GAEL-P contains 30 objects that are identified from a four-choice closed-set. The Craig Lipreading Inventory (Craig, 1992) contains both word and sentence subtests in a four-choice format. If a child scores more than 75% on the Craig sentence subtest, open-set sentence measures are employed. The Monsen Sentences (Monsen, 1978) which were originally developed to assess speech intelligibility, consist of 10 simple sentences. The CID Everyday Sentences (Davis & Silverman, 1978) contain sentences ranging from simple to complex. The most difficult test in the lipreading enhancement battery is the sentence test developed at City University of New York (CUNY) by Arthur Boothroyd and his colleagues (Boothroyd, Hanth-Chisolm, Hanin, & Kisain-Rabin, 1988). The CUNY sentences are presented via laser video disc, but the remaining tests are presented via live voice. Children are tested in a hierarchical fashion until their lipreading score falls between 40-75%. They then receive another form of that test in the auditory-plus-visual modality to determine lipreading enhancement.

Insert Table 3 about here

Indiana University Speech Perception Test Batteries

Two batteries of assessment measures are currently used in the DeVault Otologic Research Laboratory at Indiana University to evaluate the speech perception performance of children with profound hearing loss who use a sensory aid. The School-Age battery was developed for children aged 6 years and older, and the Preschool battery was developed for children between the ages of 2-5 years. Like the CID battery above, both test batteries sample performance ranging from detection through open-set word recognition and sentence identification, and utilize assessments in the auditory-only, visual-only, and auditory-plus-visual modalities. Administration of these batteries differs from the CID approach, in that all procedures within each battery are administered at every testing interval. The rationale for this approach is that it yields information about individual differences in performance, and permits an examination of the rate and pattern of development of speech perception abilities, and allows the clinician to evaluate consistency of performance across the tests in the battery. During test administration, children are first familiarized with the stimulus items in the auditory-plus-visual modality to ensure that they have the target vocabulary. Tests are administered via live-voice at approximately 70 dB SPL. Test items are presented once without repetition.

The School-Age Battery. Table 4 provides a list of the assessment procedures for children aged 6 years or older. This battery has been described in detail by Osberger and her colleagues (e.g., Osberger et al., 1991a; 1991b). The majority of tests in the School-Age Battery utilize a closed-set response format. The Monosyllable-Spondee-Trochee Test (Erber & Alencewicz, 1976) presents words differing in duration and stress: monosyllables, trochees (two-syllable words with stress on the first syllable), and spondees (two-syllable words with equal stress on each syllable). Children respond by selecting the target word from a 12-picture plate. The MTS categorization score represents the percent of stress patterns correctly

Table 3

The CID Lipreading Enhancement Test Battery (from Geers, 1994).

Test	Stimulus	Response Format	Perceptual Skill
GAEL-P[a]	Words	Closed-set	Word ID
Craig Lipreading	Monosyllabic Words	Closed-set	Word ID
Inventory	Sentences	Closed-set	Sentence ID
Monsen Sentences	Sentences	Open-set	Word recognition in simple sentences
CID Sentences[d]	Sentences	Open-set	Word recognition in complex sentences
CUNY Sentences[e]	Stories	Open-set	Connected discourse recognition

[a]Moog, Kozak, & Geers (1983); [b]Craig (1992); [c]Monsen (1978); [d]Davis & Silverman (1978); [e]Boothroyd, Hanth-Chisolm, Hanin, & Kisain-Rabin (1988)

identified (chance = 33%), and the identification score represents the child's ability to select the correct target (chance =8%). If children are able to use stress cues to narrow the set of possible responses, then chance performance on the identification task is also 33% (see Tyler, 1993).

Insert Table 4 about here

The Minimal Pairs Test (Robbins, Renshaw, Miyamoto, Osberger, & Pope, 1988a) evaluates word discrimination on the basis of vowel features (e.g., bee vs. boo) and consonant features (voicing, manner and place of articulation) (e.g., fan vs. van). According to Osberger et al., (1991a), the Minimal Pairs was developed to be similar to the DIP (Siegenthaler, 1975) but using vocabulary that was more suitable for young deaf children. The Minimal Pairs Test consists of 40 word pairs that are presented twice, and the child selects a response from a picture plate (one per word pair). The test is scored as the percent of vowel and consonant features correctly identified, along with a composite score.

The Hoosier Auditory-Visual Enhancement test (HAVE) (Robbins, Renshaw, Miyamoto, Osberger, & Pope, 1998b) was developed to assess the integration of auditory-plus-visual information. This test consists of 40 triplets containing two homophones (e.g., man, pan) and one visually distinct word (e.g., fan) each depicted on a separate picture plate. Test presentation is in the auditory-plus-visual modality, and responses are scored on the basis of visual correctness (i.e., the selection of one of the homophones) and word correctness (i.e., whether the correct homophone was selected).

The final closed-set test is the PSI (Jerger et al., 1980). This test has been described earlier. For administration in the Indiana University battery, the PSI has been modified by adding an extra picture per response card (i.e., six rather than five) so that the child cannot select the word or sentence responses simply by a process of elimination. The PSI is administered in an auditory-only, visual-only, and auditory-plus-visual modality; this allows the assessment of auditory-only word and sentence identification from a closed-set, as well as measuring lipreading enhancement in both a word and sentence context for all subjects.

Two measures have been used to assess open-set speech perception abilities. The PB-K (Haskins, 1949) assesses monosyllabic word recognition, and has been discussed earlier. The second, The Common Phrases Test (Robbins, Renshaw & Osberger, 1995) was developed to assess the understanding of familiar phrases used in everyday situations. According to Osberger et al., (1991a), the test was motivated by the idea that children would be better able to recognize familiar phrases than monosyllabic words in an open-set format. Furthermore, such a test has greater face validity than monosyllabic word tests. The Common Phrases Test consists of six lists of 10 sentences; separate lists are presented in the auditory-only, visual-only, and auditory-plus-visual modalities. Children may respond by correctly repeating all of the words in the sentence or by correctly answering a question. Performance is scored as the percent of phrases correctly understood.

The measures described above are used to assess performance within the confines of a clinic setting. Such information may not accurately predict performance under more naturalistic listening conditions. Therefore, the Indiana University battery also includes a parent-report scale, The Meaningful Auditory Integration Scale (Robbins, Renshaw & Berry, 1991), to assess real-world listening skills. The MAIS consists of 10 probes to examine auditory skills ranging from detection through comprehension in daily living activities. Responses to the probe are assigned a score from 0 to 4 depending on how frequently

Table 4

The Indiana University School-Age Perception Battery

Test	Stimulus	Presentation	Response Format	Perceptual Skill
Monosyllable-Trochee-Spondee (MTS)[a]	1-, 2-, 3-syllable words	• Auditory	Closed-set	Pattern perception Word identification
Minimal Pairs Test[b]	1-syllable words	• Auditory	Closed-set	Word discrimination based on vowel and consonant features
PB-K[c]	1-syllable words	• Auditory	Open-set	Word identification
Common Phrases[d]	2- to 6-word phrases	• Auditory only • Visual only • Auditory + Visual	Open-set	Sentence identification
Hoosier Auditory-Visual Enhancement Test (HAVE)[e]	1-syllable word triplets: two homophones and one visually distinct	• Auditory + Visual	Closed set	Word identification and auditory-visual enhancement
Pediatric Speech Intelligibility Test (PSI)[f]	Single words and sentences	• Auditory only • Visual only • Auditory + Visual	Closed-set	Word and sentence identification
Meaningful Auditory Integration Scale (MAIS)[g]	10 probes	Structured interview schedule	Parent report	Detection through comprehension in daily living situations

[a]Erber and Alencewicz (1976); [b]Robbins, Renshaw, Miyamoto, Osberger, and Pope (1988a); [c]Haskins (1949); [d]Robbins, Renshaw, and Osberger (1995); [e]Robbins, Renshaw, and Osberger (1988b); [f]Jerger, Lewis, Hawkins, and Jerger (1980); [g]Robbins, Renshaw, and Berry (1991)

the child demonstrates a particular listening skill in his or her daily environment. Although parent-report scales may be more subjective than behavioral measures of performance, it is important to assess how skills demonstrated in the clinic are carried over into natural listening situations.

Reliability of Measures in the School-Age Battery. Test-retest reliability data were collected at Boys Town National Research Hospital for some of the measures in the Indiana University School-Age battery including the Minimal Pairs, the Common Phrases, and the PBK. Preliminary data were reported by Carney, Osberger, Miyamoto, Karasek, Dettman, & Johnson (1991), and additional data were obtained from Osberger, Robbins, Todd, Riley, Kirk, and Carney (in press). Twenty-one profoundly deaf children ranging in age from 4.3-11.0 years (mean age = 7.4 years) were tested twice within two weeks. These children all were hearing aid users, with mean unaided and aided pure tone thresholds of 97 and 45 dB HL, respectively. Test-retest reliability correlations ranged from 0.89 for phoneme scores on the PBK to 0.97 for the Common Phrases administered in the auditory-plus-visual modality, indicating very good reliability on these measures.

Indiana University Preschool Battery. Table 5 lists the measures in the Indiana University Preschool battery intended for profoundly hearing-impaired children between the ages of 2-5 years. Materials and procedures in this battery were selected to: sample listening skills along a continuum of difficulty, be sensitive to individual differences among children, require little or no training, use familiar techniques, use real objects where possible, and finally, be administered in one hour or less. The Preschool battery has been described by Robbins & Kirk (in press), and will be briefly summarized here. Of the five procedures included in this battery, four are modifications of previously developed instruments.

Insert Table 5 about here

The Screening Inventory of Perception Skills (SCIPS) (Osberger et al., 1991a) was developed to evaluate children with limited auditory skills. Children must discriminate between pairs of words differing in either duration and stress, or in segmental cues. This discrimination utilizes a "Go-No Go" paradigm in which children are told to listen for one target word. A single word is presented on each trial and the child is to respond if the presented stimulus matches the target word, and to do nothing if the stimulus and target words differ. Responses are scored as the percent correct, with chance performance equal to 50%.

The GAEL-P (Moog et al., 1983) has been adapted for use as a closed-set word identification task. Children are first familiarized with the 30 objects in the auditory-plus-visual modality. During test administration, the 30 stimulus items are presented in sets of four objects, and the child must identify the target word through listening alone. The four-item set changes after each trial to prevent the child from using a process of elimination strategy. The item presentation has been re-ordered from that suggested by Moog et al., 1983) so that the 11 multi-syllabic words are presented first followed by the 19 monosyllabic items. This eliminates syllable number as a cue, forcing the children primarily to use segmental cues for word recognition.

The Mr. Potato Head Task was developed as a modified open-set task (Robbins, 1994). That is, children are asked to carry out commands in assembling a Mr. Potato Head toy through listening alone. Two percent-correct scores are generated: a sentence score for the number of commands correctly carried out, and a word score for the number of key words correctly identified, even if the command was not followed correctly. This test is considered as a modified open-set test because the number of items that can

Table 5

The Indiana University Preschool Speech Perception Battery

Test	Stimulus	Presentation	Response Format	Perceptual Skill
Screening Inventory of Perceptual Skills (SCIPS)[a]	1-, 2-, and 3-syllable words	• Auditory	• Closed-set • Go/No Go paradigm	• Pattern perception • Word identification
Grammatical Analysis of Elicited Language–Presentence Level (GAEL–P)[b]	1-, 2-, and 3-syllable words	• Auditory	• Closed-set • Object selection	• Word identification
Mr. Potato Head[c]	Mr. Potato Head and his 'Bucket of Parts' toy	• Auditory	• Modified open-set	• Key word identification • Sentence comprehension
Pediatric Speech Intelligibility Test (PSI)[d]	Single words and sentences	• Auditory only • Visual only • Auditory + Visual	• Closed-set	• Word and sentence identification
Meaningful Auditory Integration Scale (MAIS)[e]	10 probes	• Structure interview schedule	• Parent report	• Detection through comprehension in daily life

[a] Osberger et al. (1991a); [b]Moog, Kozak, and Geers (1983); [c]Robbins (1994); [d]Jerger, Lewis, Hawkins, and Jerger (1980); [e]Robbins, Renshaw, and Berry (1991)

be used is large (more than 20) but not unlimited. Because children could by chance touch an object representing a key word, 5% has been set as chance performance for key words. No chance score is assigned for sentence comprehension, as children could not complete this task through guessing alone.

The two remaining measures are also used in the School-Age battery and have been described above. The PSI (Jerger et al., 1980) is used to evaluate auditory-only word and sentence identification from a closed-set, and to examine lipreading enhancement. The MAIS (Robbins et al., 1991) is used to evaluate the child's use of listening in real-world environment.

Normative Data for the Indiana University Preschool Battery. The newly developed auditory-only measures in the Indiana University Preschool battery (the SCIPS, GAEL-P, and Mr. Potato Head task) were administered to 40 normal-hearing children, aged 2-5 years, to determine whether they were suitable for very young children (Robbins & Kirk, in press). Performance on each measure was analyzed as a function of age at the time of testing in one-year increments. For each measure, if the normal-hearing children within a given age group (e.g., 2.0-2.9 years) obtained an average score of 80% or higher, the measure was considered developmentally appropriate for children in that age range. Using this criterion, all of the closed-set tasks in the Indiana University Preschool battery are appropriate for children as young as 2.5-3.0 years of age. However, some children under 3 years had difficulty with the "Go-No Go" response paradigm on the SCIPS; when allowed to provide an imitative response they could master the task. Open-set speech perception results were more varied. The average open-set word and sentence comprehension scores of the normal-hearing subjects who were two years old at the time of testing were approximately 50%. Large increases in open-set word recognition were seen for normal-hearing children after 2.5 years, with criterion scores reached by 3 years of age. Performance on sentence comprehension improved through 4 years of age. The results suggest that language or other constraints may influence speech perception performance, especially on the open-set Mr. Potato Head task, and that this measure may not be appropriate for children younger than 3 years of age.

In order to determine whether this test battery was sensitive to individual differences or to changes over time in the performance of young children with cochlear implants, the preimplant and postimplant performance of 11 children implanted between the ages of 2-5 years were examined. At the preimplant interval, only one of the children was able to identify words from a closed set of four items, and the majority had no open-set speech understanding. After one year of cochlear implant experience, closed-set word recognition scores ranged from 23-100% words correct. In addition, eight of the children showed significant improvements in open-set key word recognition, and six in open-set sentence recognition. Furthermore, significant improvements were found for at least some children in each age group (2.0-2.9 years, 3.0-3.9 years, etc.). Based on the results from the normal-hearing and hearing-impaired children, Robbins and Kirk (in press) concluded that the new test battery was appropriate for use with children as young as 2-3 years of age, and was sensitive to individual differences. However, caution must be used when some the tests are administered to children 3 years of age or younger, as poor performance could result from either developmental or auditory factors.

Speech Performance of Children with Sensory Aids. Both the CID and the Indiana University speech perception batteries have been used to assess longitudinally the performance of prelingually, profoundly hearing-impaired children who use either a cochlear implant, a vibrotactile aid, or conventional amplification. On average, children with cochlear implants demonstrate improvements in closed-set speech perception abilities and in lipreading enhancement; furthermore, the majority of pediatric cochlear implant users obtain at least some-open set word recognition through listening alone (Geers & Brenner, 1994; Kirk, Osberger, Robbins, Riley, Todd, & Miyamoto, 1995a; Miyamoto, Kirk, Robbins, Todd, & Riley, in press;

Miyamoto, Kirk, Todd, Robbins, & Osberger, 1995a; Miyamoto, Robbins, Osberger, Todd, Riley, & Kirk, 1995b; Osberger et al., 1991a, 1991b). The acquisition of speech skills in these children is a gradual process. For example, in an investigation of the speech perception abilities of 50 children with the Nucleus multichannel cochlear implant, Miyamoto et al. (1995a) found that closed-set skills did not improve markedly until more than one year of device use, whereas open-set skills continued to improve through five years or more of postimplant experience. Similarly, Geers and Brenner (1994) reported gains in closed-set word recognition after one year of device use, with consonant recognition and auditory-only open-set skills improving over 3 or more years postimplant.

In contrast to cochlear implant users, children who use vibrotactile aids make much more limited speech perception gains over time. Geers and Brenner, (1994), Kirk et al. (1995a) and Miyamoto et al. (1995b) have all reported similar speech perception abilities for children who use a multichannel vibrotactile aid. In general, the majority of children who use a vibrotactile aid achieve the ability to recognize speech patterns with their device, but few acquire the ability to recognize words from a closed-set using spectral information (e.g., vowel and consonant cues) or to understand speech through listening alone. However, tactile aids do appear to provide information that enhances children's lipreading ability.

Osberger and her colleagues (Osberger et al., 1991b; Kirk et al., 1995a; Miyamoto et al., in press) and Geers and Moog and their colleagues (Geers & Moog, 1994; Geers & Brenner, 1994) have compared the performance of children with cochlear implants to that of profoundly hearing-impaired children who use conventional hearing aids (HA). At preimplant intervals, the speech perception skills of CI users were similar to those of HA children with unaided thresholds between 101-110 dB HL (the "Silver" HA group, to use Osberger's term), but poorer than those of HA children with unaided thresholds between 90-100 dB HL (the "Gold" HA group). However, after two or more years of device use, the speech perception abilities of CI users typically exceeds that of the Silver HA users and approaches that of the Gold HA users. These data have had important implications for determining CI candidacy.

Taken together, the above results suggest that multichannel cochlear implants provide substantial auditory information to children with profound hearing impairments who obtain little or no benefit from conventional amplification. However, individual children with CIs vary greatly in their spoken word recognition skills, depending in part on their age at onset of hearing loss and/or the length of their device use (Fryauf-Bertschy, Tyler, Kelsay, & Gantz, 1992; Miyamoto et al., 1994; Staller, Beiter, Brimacombe, Mecklenburg, & Arndt, 1991; Waltzman, Cohen, & Shapiro, 1992). Other factors that have been shown to influence speech perception performance in children with cochlear implants include the type of speech processor employed, duration of deafness, and communication mode. However, Miyamoto et al. (1994) reported that all of these factors together accounted for only 40% of the variance noted in speech perception performance, suggesting that other factors also contribute to individual differences in spoken word recognition. The traditional measures of spoken word recognition in children with hearing loss described above (including those for profoundly hearing-impaired children) provide descriptive information about speech perception performance, but they yield few insights into the nature of individual differences in spoken word recognition or the underlying perceptual and cognitive mechanisms contributing to word recognition. There is a need for theoretically-motivated measures that provide information about the perceptual processes employed by children with sensory aids during spoken word recognition.

New Directions in Pediatric Speech Perception Assessment

Traditional measures of spoken word recognition in children typically assess performance under very constrained listening conditions. Children are tested in sound-treated or quiet rooms, using stimulus

materials presented by a single talker in a carefully articulated way, and often presented in a closed-set format. Such measures may not adequately evaluate the perceptual processes used to perceive speech under more natural listening situations. More seriously, traditional tests have no theoretical motivation in terms of current process models of spoken word recognition and lexical access. Studying the processes of perceptual normalization and lexical discrimination using a few theoretical principals with new stimulus materials might provide important new insights into the wide range of speech perception differences noted among children with sensory aids.

Models of spoken word recognition generally propose several underlying perceptual processes in which the speech signal is converted into an acoustic-phonetic representation, normalized for talker differences, such as vocal tract characteristics or speaking rate, and then identified by matching these transformed internal representations to items stored in the mental lexicon (Pisoni, 1993; Pisoni & Luce, 1986; Studdert-Kennedy, 1974). In normal-hearing listeners, processing at these higher levels influences spoken word recognition and lexical access. For example, as stimulus variability increases, either by increasing the number of talkers or by varying the speaking rate, word identification accuracy decreases and response latencies increase (Mullennix and Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Sommers, Nygaard, & Pisoni, 1994). This decrement in spoken word recognition presumably occurs because the listener is engaged in perceptual normalization processes which consume common processing resources (Mullennix et al., 1989).

Lexical characteristics, such as word frequency (i.e., the frequency of occurrence of words in the language) and lexical similarity (i.e., the number of phonetically similar items) also have been shown to affect the accuracy and speed of spoken word recognition in normal-hearing listeners. One measure of lexical similarity is the number of phonetically similar words or "lexical neighbors" that differ by one phoneme from the target word (Greenberg & Jenkins, 1964; Landuaer & Streeter, 1973). For example, the words "bat, cap, cut, scat, at" are all lexical neighbors of the target word "cat." Words that occur frequently and have few lexical neighbors (i.e., "easy" words) are identified with greater accuracy than words that occur less frequently and have many lexical neighbors (i.e., "hard" words) (Cluff & Luce, 1990; Elliot, Clifton, & Servi, 1983; Luce, Pisoni, & Goldinger, 1990).

There are good reasons for suspecting that a hearing loss could disrupt either the process of perceptual normalization or lexical discrimination. For example, hearing impaired listeners may have difficulty making the fine acoustic-phonetic distinctions among words necessary for lexical selection in dense lexical neighborhoods (i.e., those with many phonetically similar words). At present however, little is known about the perceptual processes employed by children with sensory aids. In this section we describe several new procedures under development at Indiana University that are intended both to describe the speech perception abilities of children with hearing loss, and to provide new information about their underlying perceptual processing.

Theoretical Framework for Test Development

The new tests we have been working on were motivated by recent theoretical developments in spoken word recognition, in particular by a specific model of spoken word recognition know as the Neighborhood Activation Model (NAM) (Luce 1986). NAM offers a two-stage account of how the structure and organization of the sound patterns of words in memory contribute to the perception of spoken words. According to NAM, a given stimulus input activates a set of similar acoustic-phonetic patterns in memory in a multidimensional acoustic-phonetic space, with activation levels proportional to the degree of similarity to the target word (see Luce et al., 1990). Over the course of processing, the pattern

corresponding to the input receives successively higher activation levels, while the activation levels of similar patterns are attenuated. This initial activation stage is then followed by "lexical selection" among a large number of potential candidates that are consistent with the acoustic-phonetic input. Word frequency is assumed to act as a biasing factor in this model by multiplicatively adjusting the activation levels of the acoustic-phonetic patterns. In lexical selection, the activation levels are then summed, and the probabilities of choosing each pattern are computed based on the overall activation level (see Luce, 1986; Luce et al., 1990). Word recognition occurs when a given acoustic-phonetic representation is chosen based on these computed probabilities.

The Lexical Neighborhood Tests

The Lexical Neighborhood Test (LNT) and the Multisyllabic Lexical Neighborhood Test (MLNT) were developed by Kirk, Pisoni, & Osberger, (1995b) to assess word recognition and lexical discrimination in children with hearing loss. A primary goal in the development of these new perceptual tests was to select words that were likely to be within the vocabulary of children with profound hearing losses. Vocabulary items were drawn from a corpus of words obtained from research on child language development (the CHILDES database) (MacWhinney & Snow, 1985). A subset of items produced by normal-hearing children between the ages of 3-5 years served as the pool from which tokens were drawn. These children were in the early stages of language development, and therefore it seemed likely that the words would be familiar to profoundly deaf children with limited vocabularies. "Easy" and "hard" word lists for both the LNT and MLNT were generated from a computational analysis of the frequency of word occurrence and the lexical density of words within the database (Logan, 1992). "Easy" words were selected from those above the median for word frequency and below the median for lexical density, whereas "hard" words had the opposite lexical characteristics. The LNT contains two lists of "easy" and two lists of "hard" monosyllabic words. The MLNT contains one "easy" and one "hard" list of two- to three-syllable words. The tests are administered via live voice using an opens-set response format, and are scored as the percent of words and phonemes correctly identified as a function of lexical difficulty.

Kirk et al. (1995b) examined the effect of lexical characteristics on a group of pediatric cochlear implant users' spoken word recognition, and compared their performance on the LNT and MLNT with their scores on a traditional, phonetically-balanced word list, the PB-K. Results for both the LNT and MLNT demonstrated that word recognition was significantly better on the "easy" than the "hard" lists, indicating that pediatric cochlear implant users are sensitive to acoustic-phonetic similarities among words, and that they organize words into similarity neighborhoods in long-term memory, just as do listeners with normal hearing. No lexical effects on phoneme recognition were noted, suggesting that pediatric cochlear implant users identify words in the context of other words, and not on a phoneme-by-phoneme basis. Finally, word recognition was significantly higher on the lexically controlled lists than on the PB-K. Only 30% of the words on the PBK were contained within the childhood language database. It may be that the restrictions imposed by creating a phonetically-balanced word list result in the selection of test items that are unfamiliar to children with hearing loss.

Perceptually Robust Speech Perception Measures

The results of the previous investigation indicated that it is possible to design speech perception assessment procedures that shed light on the perceptual processing employed by listeners with hearing loss. However, presenting stimulus items via live voice does not assess a listener's ability to deal with speech signal variability introduced by different talkers, as is necessary in natural listening situations. Measures that assess speech perception performance under conditions containing stimulus variability should better

predict performance under real-world listening conditions. In fact, this has been demonstrated for adults who used hearing aids (Kirk, Pisoni, & Miyamoto, submitted). The hearing-impaired subjects' word recognition performance under conditions where the talker varied from trial-to-trial was more strongly correlated with self-ratings of daily listening skills than was performance in single-talker conditions. At Indiana University we are currently piloting computer-based single- and multiple-talker versions of the LNT and MLNT for use with children in our longitudinal study concerning the speech perception performance of children with sensory aids. Our goal is to examine the processes of perceptual normalization and lexical discrimination in this group, as a first step in the development of perceptually robust assessment materials for children. We believe that the development and use of perceptually robust, theoretically-motivated tests is vital to understanding individual differences in speech perception abilities.

Future Developments

We believe that is important to study the cognitive and attentional influences on spoken word recognition that may contribute to individual differences in performance. Future efforts should examine the effects on spoken word recognition of such factors as auditory and visual memory, selective attention, and the integration of auditory and visual information. We also believe that intensive experimental study of "extraordinary" or "superior" hearing-impaired listeners who use sensory aids will provide important new insights into the large individual differences that have been observed, particularly among cochlear implant users. At present, there are few (if any) audiologists and speech-language pathologists using a theoretical framework to study these problems. However, cognitive scientists have examined the issue of superior performers in the context of expert systems (see Ericsson & Lehman, in press).

Future studies should examine ways to assess the spoken language comprehension of children with hearing loss. Many of the speech perception tests in use today assess relatively low-level feature and phoneme discrimination and identification, or the identification of single words under greatly constrained conditions. We need to examine the extent to which children's abilities to make fine phonetic distinctions in open-set word recognition tests is predictive of later real-time comprehension of larger units of spoken language, like sentences and longer passages of connected speech.

Finally, speech perception is not an isolated skill, but an integral part of the communication process. Children with prelingual or early-acquired hearing loss must use the degraded signal they receive through their sensory aid to acquire speech production and spoken language skills. Assessing speech perception alone does not adequately document the nature of communication difficulties, nor does it provide sufficient information to implement aural (re)habilitation activities. We must examine new ways to relate speech perception abilities to speech intelligibility and to language performance (see Robbins, Svirsky, & Kirk, in press). Our goal is to develop predictive measures for the development of these skills. That is, can we predict changes in speech articulation and motor control, or changes in receptive and expressive language skills, from a pattern and constellation of changes in speech perception abilities? Our goal for hearing-impaired children is not to develop isolated abilities in spoken word recognition, but to generalize and incorporate new listening skills into everyday situations, and to use the auditory input from their sensory aid in the development of spoken language skills.

References

American National Standards Institute (1969). *Specifications for Audiometers* (ANSI S3.6-1969). New York: Author.

Bench J. & Bamford, J.M. (1979). *Speech-Hearing Tests and the Spoken Language of Partially-Hearing Children*. New York: Academic Press.

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

Boothroyd, A. (1993). Profound deafness. In R.S. Tyler (Ed.), *Cochlear Implants*. San Diego: Singular Publishing Group, Inc., pp. 1-37.

Boothroyd, A., Hanth-Chisolm, T., Hanin, L., & Kisain-Rabin, L. (1988). Voice fundamental frequency as an auditory supplement to the speech reading of sentences. *Ear and Hearing*, **9**, 306-312.

Carney, A.E., Osberger, M.J., Miyamoto, R.T., Karasek, A., Dettman, D.L., & Johnson, D.L. (1991). Speech perception along the sensory aid continuum: From hearing aids to cochlear implants. In J. Feigen & P. Stemalchowicz (Eds.), *Pediatric Amplification: Proceedings of the 1991 National Conference* (pp. 93-114). Omaha, NE: Boys Town National Research Hospital.

Cluff, M.S. & Luce, P.A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 551-563.

Craig, W.N. (1992). *Craig Lipreading Inventory: Word Recognition*. Englewood, CO: Resource Point.

Crandell, C. (1993). Speech recognition in noise by children with minimal hearing loss. *Ear and Hearing*, **14**(3), 210-216.

Crum, D. (1976). The effects of noise, reverberation, and speaker-to-listener distance on speech understanding. Unpublished doctoral Dissertation. Northwestern University, Evanston, IL.

Dale, D.M.C. (1974). *Language Development in Deaf and Partially Hearing Children*. Springfield, IL: Charles C. Thomas.

Davis, H. & Silverman, R. (1978). *Hearing and Deafness*. New York: Holt, Rinehart, and Winston.

Dunn, L. & Dunn. L. (1981). *Peabody Picture Vocabulary Test - Revised*. Circle Pines, MN: American Guidance Service.

Elliot, L.L., Clifton, L.A.B., & Servi, D.G. (1983). Word frequency effects for a closed-set word identification task. *Audiology*, **22**, 229-240.

Elliott, L. & Katz, D. (1980). *Development of a New Children's Test of Speech Discrimination*. Technical Manual, St. Louis, MO: Auditec.

Erber, N. (1980). Use of the auditory numbers test to evaluate speech perception abilities of hearing-impaired children. *Journal of Speech and Hearing Disorders*, **45**, 527-532.

Erber, N. (1982). *Auditory Training*. Washington, DC: Alexander Graham Bell Association for the Deaf.

Erber, N.P. & Alencewicz, C. (1976). Audiologic evaluation of deaf children. *Journal of Speech and Hearing Disorders*, **41**, 256-267.

Ericsson, K.A. & Lehman, A.C. (In press). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*.

Ewing, A.W.G. (1957). Speech audiometry for children. In A.W.G. Ewing (ed.), *Education Guidance and the Deaf Child*, pp. 278-296, Washington, DC: The Volta Bureau.

Finitzo-Hieber, T., Gerling, I.J., Matkin, N.D., & Cherow-Skalka, E. (1980). A sound effects recognition test for the pediatric audiological evaluation. *Ear and Hearing*, **1**(5), 271-276.

Fryauf-Bertschy, H., Tyler, R.S., Kelsay, D.M., & Gantz, B.J. (1992). Performance over time of congenitally and postlingually deafened children using a multichannel cochlear implant. *Journal of Speech and Hearing Research*, **35**, 892-902.

Geers, A. (1994). Techniques for assessing auditory speech perception and lipreading enhancement in young deaf children. In A.E. Geers & J.S. Moog (eds.), *Effectiveness of cochlear implants and tactile aids for deaf children: The sensory aids study at Central Institute for the Deaf. Volta Review*, **95**, 85-96.

Geers, A. & Brenner, C. (1994). Speech perception results: Audition and lipreading enhancement. In A.E. Geers & J.S. Moog (eds.), *Effectiveness of cochlear implants and tactile aids for deaf children: The sensory aids study at Central Institute for the Deaf. Volta Review*, **95**, 97-108.

Geers, A.E., & Moog, J.S. (1989). Evaluating speech perception skills: Tools for measuring benefits of cochlear implants, tactile aids, and hearing aids. In E. Owens and D. Kessler (eds.), *Cochlear Implants in Young Deaf Children.* Boston: College-Hill Press, pp. 227-256.

Geers, A.E. & Moog, J.S. (1994). Description of the CID sensory aids study. In A.E. Geers & J.S. Moog (Eds.), *Effectiveness of cochlear implants and tactile aids for deaf children: The sensory aids study at Central Institute for the Deaf. Volta Review*, **95**, 1-11.

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

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

Hodgson, W.R. (1973). A comparison of WIPI and PB-K discrimination test scores. Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Detroit, MI.

Jerger, S., Jerger, J., & Lewis, S., (1981). Pediatric speech intelligibility test. II. Effect of receptive language age and chronological age. *International Journal of Pediatric Otorhinolaryngology*, **3**, 101-118.

Jerger, S., Lewis, S., Hawkins, J., & Jerger, J. (1980). Pediatric speech intelligibility test. I. Generation of test materials. *International Journal of Pediatric Otorhinolaryngology*, **2**, 217-230.

Johnson, C.E., Cosci, S.M., Brown, J.W., & Scroggins, A.P. (1995). Reverberation and noise effects on word recognition: Preschool through first grade children. Paper presented at the Annual convention of the American Speech-Language-Hearing Association. Orlando, Fl.

Jones, K. & Studebaker, G. (1974). Performance of severely hearing-impaired children on a closed-response, auditory speech discrimination test. *Journal of Speech and Hearing Research*, **17**, 531-540.

Kenworthy, O.T., Klee, T., & Tharpe, A.M., (1990). Speech recognition ability of children with unilateral sensorineural hearing loss as a function of amplification, speech stimuli and listening condition. *Ear and Hearing*, **11**(4), 264-270.

Kirk, K.I., Osberger, M.J., Robbins, A.M., Riley, A.I., Todd, S.L., & Miyamoto, R.T. (1995a). Performance of children with cochlear implants, tactile aids, and hearing aids. *Seminars in Hearing*, **16**, 370-381.

Kirk, K.I., Pisoni, D.B., & Miyamoto, R.C. (submitted). Effects of stimulus variability on speech perception in hearing impaired listeners.

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

Lach, R.D., Ling, D., & Ling, A.H. (1970). Early speech development in deaf infants. *American Annals of the Deaf*, **115**, 522-526.

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.

Lee, L. L. (1971). *The Northwestern Syntax Screening Test*. Evanston, IL: Northwestern University Press.

Logan, J.S. (1992). A computational analysis of young children's lexicons. *Research on Spoken Language Processing Technical Report No. 8.* Bloomington, IN: Speech Research Laboratory Indiana University.

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

Luce, P. A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, **12**, 271-296.

Miyamoto, R.T., Kirk, K.I., Robbins, A.M., Todd, S., & Riley, A. (In press). Speech perception and production skills of children with multichannel cochlear implants. *Acta Otolaryngologica*, **116**.

Miyamoto, R.T., Kirk, K.I., Todd, S.L., Robbins, A.M., & Osberger, M.J. (1995a). Speech perception skills of children with multichannel cochlear implants or hearing aids. *Annals of Otology, Rhinology, & Laryngology*, **104**(Suppl.), 334-337.

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

Miyamoto, R.T., Robbins, A.M., Osberger, M.J., Todd, S. L., Riley, A.I., & Kirk, K.I. (1995b). Comparison of multichannel tactile aids and multichannel cochlear implants in children with profound hearing impairments. *American Journal of Otology*, **16**, 8-13.

Mecklenburg, D., Shallop, J., & Ling, D. (1987). *Phonetic Task Evaluation*. Englewood, CO: Cochlear Corporation.

Moog, J.S. & Geers, A.E. (1990). *Early Speech Perception Test for Profoundly Hearing-Impaired Children*. St. Louis: Central Institute for the Deaf.

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

Monsen, R.B. (1978). Toward measuring how well hearing-impaired children speak. *Journal of Speech and Hearing Research*, **21**, 197-219.

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

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

Nabelek, A. & Mason, D. (1981). Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. *Journal of Speech and Hearing Research*, **24**, 375-383.

Nabelek, A. & Robinson, P. (1982). Monaural and binaural speech perception in reverberation for listeners of various ages. *Journal of the Acoustical Society of America*, **71**, 1242-1248.

Osberger, M.J., Miyamoto, R.T., Zimmerman-Phillips, S., Kemink, J.L., Stroer, BS., Firszt, J.B., & Novak, M.A. (1991a). Independent evaluation of the speech perception abilities of children with the Nucleus 22-channel cochlear implant system. *Ear and Hearing*, **12**(Suppl.), 66S-80S.

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

Osberger, M.J., Robbins, M.J., Todd, A.M., Riley, A.I., Kirk, K.I., & Carney, A.E. (In Press). Cochlear implants and tactile aids with profoundly hearing-impaired children. In F. Bess (Ed.), *Amplification for Children with Auditory Deficits*. Nashville, TN: Bill Wilkerson Center Press.

Pisoni, D.B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, **4**, 75-95.

Pisoni, D.B. & Luce, P.A. (1986). Speech perception: Research, theory, and the principal issues. In E.C. Schwab & H.C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines: Speech Perception, 1*, New York: Academic Press, pp. 1-50.

Quigley, S.P. & Paul, P.V. (1984). *Language and Deafness*. San Diego: College-Hill Press.

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

Robbins, A.M. & Kirk, K.I. (In Press). Speech perception assessment and performance in pediatric cochlear implant users. *Seminars in Hearing*.

Robbins, A.M., Renshaw, J.J. & Berry, S.W. (1991). Evaluating meaningful auditory integration in profoundly hearing-impaired children. *American Journal of Otology*, **12**(Suppl.), 144-150.

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

Robbins, A.M., Renshaw, J.J., Miyamoto, R.T., Osberger, M.J., & Pope, M.L. (1988b). *Hoosier Auditory-Visual Enhancement Test*. Indianapolis, IN: Indiana University School of Medicine.

Robbins, A.M., Renshaw, J.J., & Osberger, M.J. (1995). *Common Phrases Test*. Indianapolis, IN: Indiana University School of Medicine.

Robbins, A.M., Svirsky, M.A., & Kirk, K.I. (In press). Implanted children can speak, but can they communicate? *Otolaryngology -- Head & Neck Surgery*.

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

Sanderson-Leepa, M. & Rintelmann, W. (1976). Articulation functions and test-retest performance of normal-hearing children on three discrimination tests: WIPI, PBK-50, and NU Auditory Test No. 6. *Journal of Speech and Hearing Disorders*, **41**, 503-519.

Schwartz, D. (1971). The usefulness of the WIPI. A speech discrimination test for pre-school children. Master's thesis. Central Michigan University, Mount Pleasant, MI.

Siegenthaler, B.M. (1975). Reliability of the TIP and DIP speech-hearing tests for children. *Journal of Communication Disorders*, **8**, 325-333.

Siegenthaler, B.M. & Haspiel, G. (1966). Development of two standardized measures of hearing for speech by children. *Cooperative Research Program, Project No 2372, Contract OE-5-10-003.* Washington, DC: US Department of Health, Education, and Welfare, US Office of Education.

Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1994). Stimulus variability and spoken word recognition I: Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-1324.

Staller, S.J., Beiter, A.L., Brimacombe, J.A., Mecklenburg, D.J., & Arndt, P. (1991). Pediatric performance with the Nucleus 22-channel cochlear implant system. *American Journal of Otology,* **12**(Suppl.), 126-136.

Studdert-Kennedy, M. (1974). The perception of speech. In T.A. Sebeok (Ed.), *Current Trends in Linguistics, 2*, New York: Academic Press, pp. 1-62.

Tyler, R.S. (1993). Speech perception by children. In R.S. Tyler (Ed.), *Cochlear Implants*. San Diego, CA: Singular Publishing Group, Inc., pp. 191-256.

Tyler, R.S. & Holstad, B.A. (1987). *A closed-set speech perception test for hearing-impaired children.* Iowa City, IA: The University of Iowa.

Waltzman, S.B., Cohen, N.L., & Shapiro, W.H. (1992). Use of multichannel cochlear implants in the congenitally and prelingually deaf population. *Laryngoscope*, **102**, 395-399

Appendix A

Inventory of Open-Set Speech Perception Tests

Tests of Single-Word Recognition

- Phonetically Balanced Word Lists - Kindergarten (PB-K) (Haskins, 1949)
- Manchester Junior Lists (Ewing, 1957)
- Mr. Potato Head (Robbins, 1994)
- Lexical Neighborhood Test (Kirk et al., 1995b)
- Multisyllabic Lexical Neighborhood Test (Kirk et al., 1995b)

Tests of Sentence Recognition

- The Common Phrases Test (Robbins et al., 1995)
- Mr. Potato Head (Robbins, 1994)
- BKB Sentences (Bench, Kowal, & Bamford, 1979)

Appendix B

Inventory of Closed-Set Speech Perception Tests

Environmental Sounds

- Sound Effects Recognition Test (SERT) (Finitzo-Hieber, Gerling, Matkin, & Cherow-Skalka, 1980)

Tests of Speech Feature Perception

- Discrimination by the Identification of Pictures (DIP) (Siegenthaler & Haspiel, 1966)
- Screening Inventory of Perceptual Skills (SCIPS) (Osberger et al., 1991b)
- Phonetic Task Evaluation (Mecklenburg et al., 1987)
- Monosyllable-Trochee-Spondee Test (Erber & Alencewicz, 1976)
- Early Speech Perception Test (Moog & Geers, 1990)
- Auditory Numbers Test (ANT) (Erber, 1980)
- The Minimal Pairs Test (Robbins et al., 1988a)

Tests of Single-Word Recognition

- Northwestern University-Children's Perception of Speech (NU-CHIPS) (Elliot & Katz, 1980)
- Word Intelligibility by Picture Index (WIPI) (Ross & Lerman, 1979)
- Grammatical Analysis of Elicited Language - Presentence Level (GAEL-P) (Moog et al., 1983)
- Early Speech Perception Test (ESP) (Moog & Geers, 1990)
- Pediatric Speech Intelligibility Test (PSI) (Jerger et al., 1980)
- The Minimal Pairs Test (Robbins et al., 1988a)

Sentence Tests

- Pediatric Speech Intelligibility Test (PSI) (Jerger et al., 1980)
- Matrix Test (Tyler & Holstad, 1987)

Parent Report Scales

- The Meaningful Auditory Integration Scale (MAIS) (Robbins et al., 1991)

Appendix C

Inventory of Tests of Auditory Visual Enhancement

- The Common Phrases Test (Robbins et al., 1995)
- Pediatric Speech Intelligibility Test (PSI) (Jerger et al., 1980)
- Hoosier Auditory-Visual Enhancement Test (HAVE) (Robbins et al., 1988b)
- Craig Lipreading Inventory (Craig, 1992)
- Grammatical Analysis of Elicited Language - Presentence Level (Moog et al., 1983)
- Monsen Sentences (Monsen 1978)
- Central Institute for the Deaf Everyday Sentences (Davis & Silverman, 1978)
- City University of New York Sentences (Boothroyd et al., 1988)

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Recollection of Illusory Voices[1]

Helena M. Saldaña,[2] Kathleen B. McDermott,[3] David B. Pisoni, and Henry L. Roediger III[3]

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

[1] This research was supported by NIH NIDCD Research Grant DC-00111 to Indiana University in Bloomington, IN.
[2] Now at House Ear Institute, Los Angeles, CA.
[3] Psychology Department at Rice University, Houston TX.

Abstract

When subjects are given a list of words that are strongly associated with a nonpresented item they frequently recall the missing word and also recognize it as having occurred in the study list (Roediger & McDermott, 1995). Subjects also report that they remember the actual event of hearing the nonpresented word. The present investigation explores the phenomenology of false recognition in four experiments. In these studies, subjects were presented with auditory lists of related words spoken by male and female voices, and then given a recognition test containing both studied items and highly related nonpresented items. Subjects were asked to judge each test item as "old" or "new" and, if old, whether they remembered the presentation voice. In all four experiments, subjects showed high levels of false recognition and in general the same willingness to attribute the gender of the speaker to the nonpresented items as to the presented items. This pattern emerged for both lists with mixed voices (i.e., male and female) and lists with consistent voices (i.e., male or female); and for auditory as well as visual recognition tasks. The general pattern of results demonstrates that subjects' recognition of critical lures in this paradigm has the same essential characteristics of recognition of studied items, thereby producing a subjectively compelling memory illusion. In addition, the act of recall generally enhanced both veridical and illusory recognition.

Recollection of Illusory Voices

Roediger and McDermott (1995) presented subjects with lists of words that were strongly related to a nonpresented word and showed that subjects frequently recalled the nonpresented word on an immediate recall test. In subsequent recognition tests, subjects falsely recognized the nonpresented item at the same level (around 80%) as they did the studied items. In addition, when subjects were asked to make "remember/know" judgments on recognized items (Tulving, 1985), they claimed to *remember* the nonpresented items at about the same level as the presented items. Roediger and McDermott (1995) concluded that items falsely recalled and falsely recognized seemed indistinguishable to their subjects from other items/information recalled from secondary memory. These results seemed remarkable because subjects were recalling and recognizing word lists (believed, since Bartlett's (1932) pronouncements, to be recalled in a reproductive rather than a reconstructive manner), after short retention intervals and with specific instructions not to guess.

Roediger and McDermott's (1995) results have generally been confirmed by other researchers. Payne, Elie, Blackwell, and Neuschatz (1996) conducted several experiments that were rather close replications of Roediger and McDermott's (1995) Experiment 2, obtaining similar results (with one exception to be mentioned below) and extended the paradigm in several ways. McDermott (1996) showed that after a 2-day delay, false recall exceeded veridical recall; she also showed that the illusory false recall was still maintained, albeit at reduced levels, after five study/test cycles. That is, subjects still falsely recalled words after hearing the lists and being tested on it five times. Norman and Schacter (1996) and Schacter, Verfaellie, and Pradere (1996) used the Roediger-McDermott paradigm to test elderly subjects and amnesiac subjects, respectively. Interestingly, older subjects showed the illusory false recall and false recognition to a greater extent than did younger subjects, despite recalling the lists less well overall (as expected). On the other hand, amnesiac patients' recall of the lists items was very poor, and they failed to show the illusory recall and recognition effects.

These experiments and others (e.g., Read, 1996) generally all confirm the results of Roediger and McDermott (1995), even if leaving them poorly understood theoretically. The primary motivation for the present series of experiments was to examine subjects' phenomenological experience during recognition for both veridical and illusory recognition. In particular, we wished to examine subjects' attributions about the voice used for presentation when the lists of related words were presented in two voices. Previous research on recognition memory has demonstrated that listeners are quite good at remembering not only the specific words that were presented to them during study but also the physical attributes of the voice (Palmeri, Goldinger & Pisoni, 1993), suggesting that memory representations for spoken words contain highly specific details about the event of hearing the word (Goldinger, 1992). In this series of experiments, we were interested in whether the phenomenological experience in false recognition was similarly detailed. Another purpose of the present set of experiments was to gain insight into a more theoretical question about the nature of representations for illusory memories. There is now converging evidence that the phenomenological experience of an illusory memory is very similar to a veridical memory; if so, the suggestion could be made that the representations in memory are similar if not the same. In this set of experiments, we try to determine whether there are fundamental differences between the two types of representations that can be observed during recognition memory experiments.

While conducting this research, we became aware of related work by Payne et al. (1996), who asked a similar question. In their Experiment 3, subjects heard the Roediger-McDermott (1995) lists presented in two voices and then recalled the lists three times. After the third recall, subjects were asked to indicate beside each recalled item whether it had been presented in a male voice, a female voice, or to indicate that

they did not know which voice the item had originally been presented in. Under the conditions in Payne et al.'s experiment, the probability of both veridical and false recall was just over .30. Subjects made voice attributions for between 80 and 90% of the recalled items, both studied and nonstudied, and rarely used the "don't know" category even for items that had not been presented. In the experiments reported below, we asked subjects to assign input modality judgments during a recognition test, rather than after a recall test. Because false recognition in the Roediger-McDermott (1995) paradigm is often higher than 70%, subjects will have more opportunities to judge modality of occurrence for both studied and nonstudied (critical) items.

In all of our experiments, subjects heard lists of words presented in both a male and female voice and then later received a recognition test that requested information about whether test items had been previously heard in one voice or the other voice. Across four generally similar experiments we included manipulations that permitted us to ask the following questions: When experiencing a false recognition would subjects be as willing to attribute the critical lure being spoken in one voice or the other? Would such attributions be as frequent for illusory memories as for veridical memories? Answers to both questions were affirmative in Experiment 1, in which the lists were presented with some words in a male voice and other words in a female voice, and in which items on the recognition test were presented visually. In Experiment 2, we asked if the same outcome would be obtained if subjects received an auditory recognition test instead of a visual recognition test. Matching and mismatching of voice characteristics between study and test might provide an additional clue to recognition and could thereby reduce false recognition. In Experiments 3 and 4, we asked if presenting lists all in one voice or all in the other voice might bias a subject's response for voice attributions. In other words, are subjects more likely to attribute a male voice to a critical lure if the original associated list was presented in a male voice? We also asked whether pure voice lists would reduce false recognition when the test items were presented either visually or auditorily.

A secondary motivation for the current experiments was to confirm (or disconfirm) one aspect of the Roediger and McDermott (1995) results that has not been consistently replicated. They showed that prior recall of lists of words generally increased later recognition, both for veridical (studied) items and for illusory (nonstudied) items. In their Experiment 2 subjects heard 16 lists of words but recalled only half the lists. On the later recognition tests, both veridical and illusory recognition were greater for the previously recalled lists than for the nonrecalled lists. However, others have not consistently obtained these effects. Payne et al. (1996) found such a testing effect in recognition for studied items but not for nonstudied (critical) items. Schacter et al. (1996) did not find statistically significant testing effects for either studied or nonstudied items. Other experiments have reported generally small positive effects of a prior free recall test on later recognition for studied items, but only for items occurring at the end of the list (see Jones & Roediger, 1995; Lockhart, 1975). Therefore, the inconsistencies in the outcome may not be too surprising. However, because repeated testing seems to represent a key aspect to the development of false memories (see Roediger, McDermott & Goff, in press), the effect of recall on later recognition in this paradigm deserves more careful examination. McDermott (1996) has shown that prior recall of both studied and nonstudied items has a powerful effect on their later recall, so the only disputed question is whether prior recall affects later recognition. Therefore, each of the four experiments reported here incorporated the design of Roediger and McDermott's (1995) Experiment 2 in which subjects studied 16 lists and were tested on 8, so that the effect of recall on later recognition could be examined.

In sum, our experiments explored the issues of subjects' willingness to attribute recollection of voice information for both accurately recognized (studied) and false recognized (nonstudied) items and the effect of prior recall on both veridical and illusory recognition.

Experiment 1

The purpose of the first experiment was to replicate and extend the recent findings of Roediger & McDermott (1995). These researchers utilized a procedure developed by Tulving (1985), to try to gain some insight into the phenomenological experience of subjects when falsely recognizing critical lures (Experiment 2). The methodology involved asking listeners to make "remember/know" judgments on items that they had identified as old. A *remember* judgment was to be given if subjects could mentally relive the experience of the presentation of the item (recalling what they were doing when the item was presented, or the physical characteristics associated with the item presentation). A *know* judgment was to be given if the subject was confident that the item appeared on the list but could not remember anything specific about the actual occurrence of the item on the list.

The pattern of results showed that *remember* judgments were just as likely to be given to falsely remembered items as actual remembered items. It was concluded that false memories can be the result of conscious recollection and not only general familiarity. In Experiment 1 of the current investigation we extend this finding by asking subjects to make judgments on the source attributes of the list items. This methodology avoids asking subjects to report directly on their phenomenological experience. Instead, subjects are asked a specific question about the occurrence of a recognized item (whether it was presented in a male voice or female voice)

Method

Subjects

Thirty-two Indiana University undergraduates participated in a one-hour session in partial fulfillment of a course requirement in an introductory psychology course. All subjects were native speakers of English and reported no history of hearing or speech impairments at the time of testing.

Materials

The 24 word lists developed by Roediger and McDermott (1995) were utilized in the present experiment. Each list consisted of 15 words that were highly associated to a nonpresented word in the Russell and Jenkins (1954) norms. The resulting 360 items were digitized on-line at a 20-kHz sampling rate and 16-bit resolution by both a male and a female talker. The root-mean-square amplitude of all stimulus items was equated using a signal processing package. Each list was composed of alternating male and female voices; half of the lists started with a male voice and half of the lists started with a female voice. The ordering of each list was held constant with the highest associates occurring first.

Design

The 24 lists were divided into three sets for the purpose of counterbalancing. Subjects were presented with 16 lists during the study phase of the experiment, with 8 of the lists tested for immediate free recall and the other 8 lists not tested. The remaining 8 lists were not presented during the study phase but appeared on the subsequent recognition test. In the recognition test, subjects were presented with visual words and were asked to determine whether the item had been presented previously. If subjects judged that the item had not been presented during the study phase they were to press a button labeled "new." If the subjects determined that an item was presented during the previous study phase, they were to give one of three possible responses. If the item had been presented previously and the subject remembered which voice it was presented in (male or female) then they were to press buttons labeled "male-old" or "female-old,"

respectively. However, if they determined that the item had been presented but could not remember which voice it had been presented in, they were to press a button simply labeled "old." The labels on the response boxes were changed for each group of subjects.

Procedure

Subjects were tested in groups of six or fewer in a quiet testing room. Stimuli were presented over matched and calibrated TDH-39 headphones at approximately 75 dB SPL. A PDP-11/34 computer was used to present stimuli. The digitized stimuli were converted to 10-kHz 12-bit resolution files for presentation.

Subjects were told that they would be participating in a memory experiment in which they would hear lists of spoken words over the headphones. They were told that after each list was presented they would hear a tone or knock (with examples given), that would indicate whether they should recall the list or perform some math problems. Subjects were instructed to listen carefully to each list because the signal for the task would not occur until after each list was presented; therefore, subjects were unaware until the end of the list whether they would be required to recall the items. The interstimulus interval was 1.5 seconds within lists. Subjects were given 1 minute after presentation of the signal to either recall items or do math problems. Each of these tasks was performed on a piece of paper supplied by the experimenter. After 1 minute, a tone occurred and subjects were instructed to turn over their response sheets so they were no longer in view and prepare for the next word list.

The recognition test occurred about five minutes after the last test or math period. During this time, subjects were given instructions about making old and new judgments. They were told that they would see one item at a time presented on a CRT screen and that they would be required to make judgments on each item. If the item was not presented in the previous study phase, subjects were to press the "new" button. If the item was presented in the previous study phase and the subject could not remember which voice it was presented in, they were told to press the "old" button; however, if they remembered the voice the item was presented in they were told to press the "male-old" or "female-old" buttons.

The recognition test was composed of 96 items, 48 of which had been studied and 48 of which had not. The 48 studied items were obtained by selecting three items from each of the 16 presented lists (always in the serial position 1, 8, and 10). The nonstudied items consisted of the 24 critical lures from all 24 lists (16 studied, 8 not studied) and the 24 items from the nonstudied list (again always in the serial position 1, 8, and 10). The 96 items were randomly presented for each group of subjects. The lists were counterbalanced by having each set of lists serve in each of the three conditions (study-recall, study-math, and nonstudied) across subjects.

Results

Recall

Subjects recalled the critical nonpresented item on 39% of the lists. The serial position curve for studied words is shown in Figure 1. Subjects recalled the critical nonpresented items at about the same rate of studied items presented in the middle of the lists. The false recall in this Experiment was not as high as in Roediger and McDermott's (1995) Experiment 2, perhaps because the test period was only 1 minute in this experiment but was 1.5 minutes in Roediger and McDermott. Still, false recall was robust and approximated recall of words from the middle of the list.

Insert Figure 1 about here

Recognition

The recognition results are presented in Table 1. The hit rate for the study-recall condition (77%) was not significantly different than the rate observed in the study-arithmetic condition (70%), $t(31) = 1.942$, $p = .06$. The false alarm rate for items from the nonstudied list was .19.

Insert Table 1 about here

The recognition results for the critical nonpresented lures are shown at the bottom of Table 1. The recognition for the critical lures is even greater than recognition of presented items, however, this was not statistically significant. Once again there is no difference between the hit rates between recalled lists and arithmetic lists, $t(31) = .955$, $p>.1$.

False recognition was actually somewhat higher than correct recognition; that is, the false alarm rate for critical items was 6% higher than the hit rate, $t(31) = 2.09$, SEM = .028, $p<.05$. Thus, although false recall was a bit lower than the levels reported by Roediger and McDermott (1995), the level of false recognition was remarkably similar overall.

The primary new question asked in this experiment is whether subjects would be as willing to assign a voice of the speaker to items that were never presented as to items that were presented. In general, the rates of voice attribution are rather similar, although they are a bit higher for studied items than for critical lures. In the Study + Recall condition, subjects provided a voice attribution on 51% of the "old" judgments for studied items, whereas the figure for critical lures was 42%, $t(31) = 1.806$, SEM =.067, $p= .08$. The comparable figures for the Study + Arithmetic condition were .44 and .37, $t(31) = 1.983$, SEM= .071, $p=.06$, for the studied items and the critical lures, respectively.

A nonsignificant trend was observed for studied items to have slightly greater rates of voice attribution than nonstudied items; however, subjects would quite frequently provide voice attributions to the critical nonstudied items. This evidence fits with other data described in the introduction showing that subjects' false recollections in this paradigm seem quite similar to their veridical recollections based on presented items.

An analysis on the overall effect of prior recall on recognition reveals a positive effect of recall (79% vs. 74%), $t(31) = 2.290$, SEM = .026, $p<.05$, on later recognition. In other words subjects have higher recognition rates when subjected to a recall task. However the effect only approached significance for the study + recall (77%) vs. the study + math (70%) conditions $t(31)$ 1.942, SEM = .038, $p = .06$; and was not significant for the critical lure + recall (81%) vs. critical lure + math (78%) conditions $t(31) = .955$, SEM = .035, n.s. The overall effect is consistent with previous studies on illusory memory reported by Roediger and McDermott (1995).

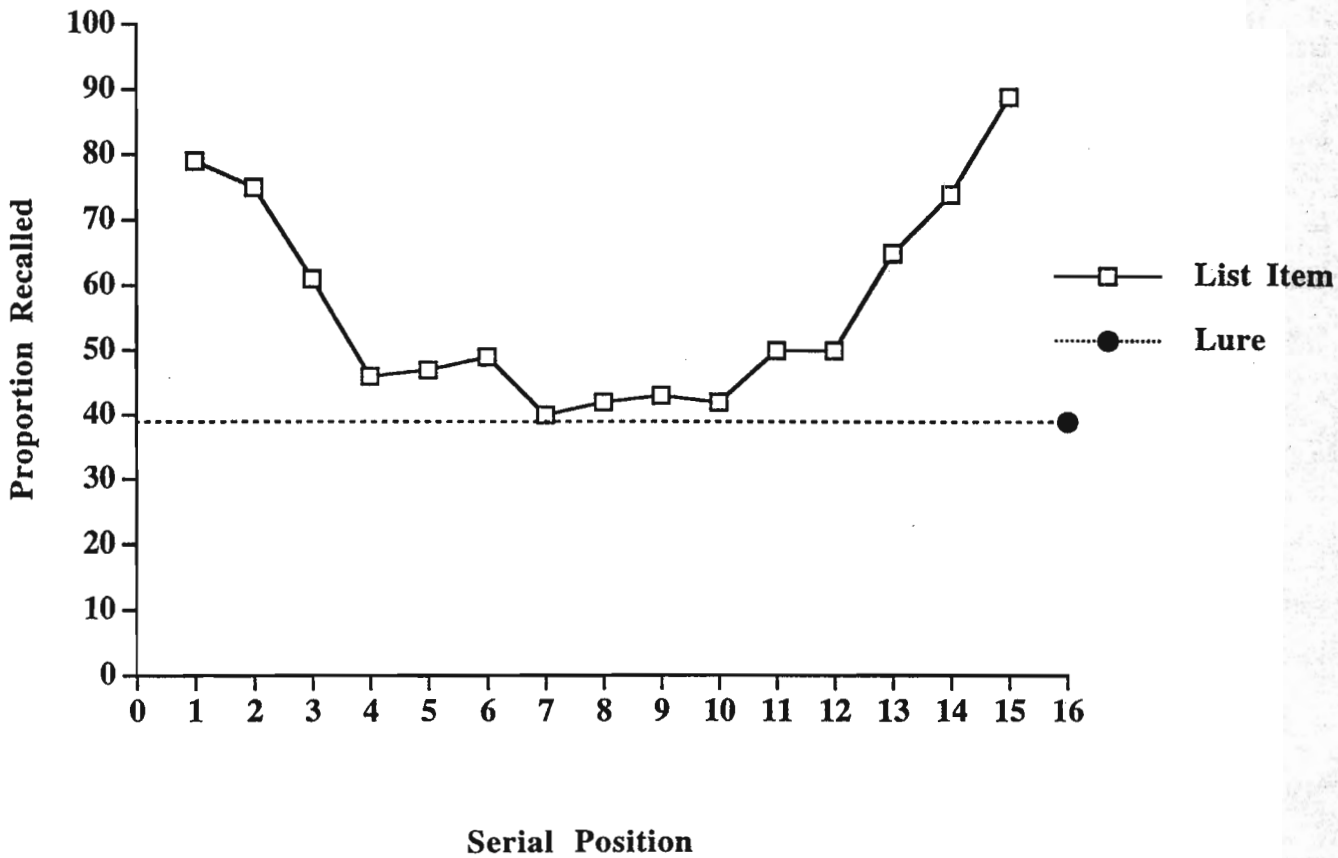


Figure 1. Serial position curve for recall the condition in Experiment 1.

Table 1

Recognition results for studied items (top) and critical lures (bottom) for Experiment 1. M and F represent male and female judgments respectively.

Item Type	Condition	Proportion of Old Responses				
		Overall	Male	Female	Don't Know	%Voice Attribution
Studied						
	Study +Recall	.77	.21	.18	.38	.51
	Study + Math	.70	.17	.14	.39	.44
	Non-Studied	.19	.04	.03	.12	.37
Critical Lure						
	Study +Recall	.81	.16	.18	.47	.42
	Study + Math	.78	.14	.15	.49	.37
	Non-Studied	.19	.02	.03	.14	.26

Experiment 2

The purpose of the second experiment was to replicate and extend our new findings about recognition of source attributes. The same general procedures were used in Experiment 2 with one exception: subjects were presented with an auditory rather than visual recognition test. Further, the voice in which the test item was presented could either match or mismatch the voice used in the study phase. We were interested in seeing if supplying subjects with voice information during the test might reduce their overall false alarm rates to the critical nonstudied items and also reduce the false attributions of voice to the critical nonstudied words.

Method

Subjects

Forty-four Indiana University undergraduates participated in a one-hour session in partial fulfillment of an introductory psychology course requirement. All of the subjects were native English speakers, and reported no history of speech or hearing impairments at the time of testing. The data from four of the subjects were omitted from the final analysis because they failed to follow instructions.

Materials

The speech materials utilized in the present experiment were identical to the first experiment.

Design

The study and recall phase of Experiment 2 was identical to that of Experiment 1; however, the recognition phase differed in that subjects were presented with auditory test items. The recognition test was composed of 96 items, 48 of which had been studied and 48 of which had not. The 48 studied items were obtained by selecting three items from each of the 16 presented lists (always in the serial position 1, 7, and 10). The nonstudied items consisted of the 24 critical lures from all 24 lists (16 studied, 8 not studied) and the 24 items from the nonstudied list (again always in the serial position 1, 7, and 10). The test items were counterbalanced so that half of the items were presented in a male voice and half in a female voice, and, of those, half were in the same voice as the test item and half were in a different voice as the test item. Half of the 16 lures were presented in a male voice and half presented in a female voice.

The 96 items were randomly presented to each group of subjects. The lists were counterbalanced by having each set of lists serve in each of the three conditions (study-recall, study-math, and nonstudied) across subjects.

Procedure

The general procedure was the same as Experiment 1, except in this experiment, subjects listened to the recognition test over the headphones. As in Experiment 1, subjects were instructed to determine whether an item had been presented during the initial study phase. If the subject determined that an item was not presented during the initial study phase, they were required to press a button labeled "new." However, if they judged that the item was presented during the initial study phase they had three possible responses. If the subject remembered that the item was presented during the initial study phase but did not remember which voice the item was presented in then they were to push a button labeled "old." If the subject

remembered that the study item was presented in the same voice as the test item then they were to push the button labeled "old-same." If the test item was in a different voice than the study item then they were to push a button labeled "old-different."

Results

Recall

Subjects recalled the critical nonpresented item on 38% of the lists. Subjects recalled the critical nonpresented items at about the same rate as studied items presented in the middle of the lists.

Recognition

The recognition results are presented in Table 2. The overall hit rate for the study-recall condition (75%) was significantly different from the recall in the study-arithmetic condition (70%), $t(39) = 2.29$, $p<.05$. The false alarm rate for the nonstudied items was 15%.

Insert Table 2 about here

The recognition results for the critical nonpresented lures are shown at the bottom of Table 2. Recognition for the critical lures is even greater than recognition of presented items (78% vs. 73%), and this difference approaches significance, $t(39) = 1.781$ $p=.08$. The difference between recognition of lures between recall and arithmetic was not significant, $t(39) = 1.3$ $p>.1$.

One question of interest in this experiment was whether subjects would be more likely to make a voice attribution when auditory information was available during the recognition phase. We observed a significant difference in the rates of voice attribution between the study + recall items (48%) and the critical lure + recall items (38%), $t(39) = 3.098$, SEM = .062, $p<.05$. But no difference was observed between the study + math items (46%) and the critical lure + math items (44%) $t(39) = .324$, SEM = .065, n.s.

An analysis of the overall effect of prior recall on recognition reveals a positive effect of recall (77% vs. 72%), $t(39) = 2.459$, SEM = 2.0, $p<.05$, on later recognition. This replicates the effect from Experiment 1. The effect was only significant for the studied items; study + recall (76%) vs. the study + math (70%) conditions, $t(39)$ 2.288, SEM = 2.15, $p<.05$; and not for the critical lure + recall (79%) vs. critical lure + math (76%) conditions, $t(39)=1.300$, SEM=2.8, n.s.

The results from Experiment 2 are not as clear-cut as those obtained in Experiment 1. In general the same pattern of source judgments is observed for the studied items and the critical lures; however, when lists have a recall task associated with them it appears that subjects are more likely to make source attributes to studied items than to critical lures. Another pattern that emerges from this experiment is that when subjects do make a source attribution they tend to have a bias for judging an item as occurring in the same voice as the study item regardless of whether it occurred or not. This might suggest that the voice attributions are the result of some general feeling a familiarity with an item and not the result of an actual recollection of an event.

Table 2

Recognition results for studied items (top) and critical lures (bottom) for Experiment 2.

Item Type	Condition	Proportion of Old Responses				
		Overall	Same	Different	Don't Know	%Voice Attribution
Studied						
	Study + Recall Same	.74	.27	.09	.38	.49
	Study + Recall Diff	.77	.23	.14	.40	.47
	Mean	.76	.25	.12	.39	.48
	Study + Math Same	.72	.24	.08	.40	.45
	Study + Math Diff	.69	.24	.08	.37	.46
	Mean	.70	.24	.08	.38	.46
	Nonstudied	.15	.03	.02	.10	.33
Critical Lure						
	Study + Recall	.79	.21	.09	.49	.38
	Study + Math	.76	.23	.10	.43	.44
	Mean	.78	.22	.10	.46	.41

Experiment 3

The purpose of Experiments 3 and 4 was to ask more specific questions about the source judgments and the recollections of our subjects. In the first two experiments, all of the lists that were presented consisted of mixed voices. This might explain why we did not observe any consistent patterns of source judgments. In the following two experiments, we replicate the previous studies except that now the voices in the lists are blocked (either all male or all female voices). This allows us to look at whether listeners are encoding specific attributes of the list items and whether this affects the recollection of the lures. In general, we would expect that subjects will attribute a source to the lure that is the same as the associated list. In Experiment 3, we presented subjects with a visual recognition test and asked them whether they remembered the source of recognized items. In Experiment 4, we presented subjects with an auditory recognition test and asked them to tell us whether the items were presented in the same or different voices.

Method

Subjects

Thirty-four Indiana University undergraduates participated in a one-hour session in partial fulfillment of an introductory psychology course requirement. All the subjects were native English speakers and reported no speech or hearing impairments at the time of testing. The data from one of the subjects were omitted from the final analysis because of his failure to follow instructions.

Materials

The speech materials utilized in the present experiment were identical to those used in Experiments 1 and 2.

Design

The design of Experiment 3 was identical to that of Experiment 1 with one exception; each of the 16 word lists was presented in either a male voice or a female voice.

Procedure

The instructions were identical to those of Experiment 1. The recognition test occurred about five minutes after the last test or math period. During this time, subjects were given instructions about making old and new judgments. They were told that they would see one item at a time presented on a CRT screen and that they would be required to make judgments on each item. If the item was not presented in the previous study phase, subjects were to press the "new" button. If the item was presented in the previous study phase and the subject could not remember which voice it was presented in they were told to press the "old" button; however, if subjects remembered the voice the item was presented in they were told to press the "male-old" or "female-old" buttons.

The recognition test was composed of 96 items, 48 of which had been studied and 48 of which had not. The 48 studied items were obtained by selecting three items from each of the 16 presented lists (always in the serial position 1, 7, and 10). The nonstudied items consisted of the 24 critical lures from all 24 lists (16 studied, 8 not studied) and the 24 items from the nonstudied list (again always in the serial position 1, 7,

and 10). The 96 items were randomly presented to each group of subjects. The lists were counterbalanced by having each set of lists serve in each of the three conditions (study-recall, study-math, and nonstudied) across subjects.

Results

Recall

Subjects recalled the critical nonpresented item on 29% of the lists.

Recognition

The recognition results are presented in Table 3. The overall hit rate for the study-recall condition (74%) was not significantly different from the hit-rate in the study-arithmetic condition (70%), $t(32) = .807$ $p>.1$. The false alarm rate for the nonstudied items was 16%.

Insert Table 3 about here

The recognition results for the critical nonpresented lures are shown at the bottom of Table 3. The recognition for the critical lures was higher for the study + math condition ,75%, than the study + recall condition, 67%, however, this difference was not statistically significant, $t(32) = 1.646$, $p>.1$.

There was no effect in Experiment 3 of prior recall on recognition $t(32) = .613$, SEM = 3.7, n.s. This is inconsistent with the effects found in both Experiment 1 and 2. Those experiments showed a small but significant positive effect of recall on later recognition. It should be noted however, that overall subjects in this experiment did much poorer on the recall portion of the experiment than subjects in the first two experiments. The overall recall rate in Experiment 1 (56%) and Experiment 2 (49%), is much higher than that found in Experiment 3 (40%). This is surprising given that the stimuli for all three experiments were similar (with only the voicing within lists changing). However, there is evidence that additional voicing cues provided within lists can aid in serial recall (see Goldinger et al., 1993).

The finding of most interest in this experiment was that subjects are quite accurate at correctly identifying the source attributes for studied items. In other words, subjects are more likely to attribute a female voice to an item if it was presented in a female voice and a male voice if it was presented in a male voice. This indicates that detailed voice information is being encoded with the specific word in memory. Furthermore, the subjects' voice attributions to critical lures are influenced by the voice of the associated word list. In fact, the results indicate that subjects tend to attribute the voice of the associated list to the critical lure with the same frequency as they attribute the correct source to an actual presented item.

Experiment 4

The results of Experiment 3 demonstrate the subjects can remember specific details about the physical attributes of a word list and that illusory memories are influenced by those specific details. The purpose of Experiment 4 was to investigate the effect that voices have during the recognition phase of the experiment. The general design of Experiment 4 was the same as Experiment 3 with one exception: in Experiment 4, the recognition list was auditory instead of visual. This allows us to look at recognition rates for consistent

Table 3

Recognition results for studied items (top) and critical lures (bottom) for Experiment 3.

Item Type	Condition	Proportion of Old Responses				
		Overall	**Male**	**Female**	**Don't Know**	**%Voice Attribution**
Studied						
	Study + Recall Male	.73	.22	.09	.42	.42
	Study + Recall Female	.75	.10	.22	.43	.42
	Mean	.74	.16	.16	.43	.42
	Study + Math Male	.70	.18	.07	.45	.36
	Study + Math Female	.70	.06	.19	.45	.36
	Mean	.70	.12	.26	.45	.36
	Nonstudied	.16	.02	.02	.12	.25
Critical Lure						
	Study + Recall Male	.67	.20	.07	.40	.40
	Study + Recall Female	.67	.09	.18	.40	.39
	Mean	.67	.15	.12	.40	.40
	Study + Math Male	.73	.20	.08	.45	.38
	Study + Math Female	.78	.05	.21	.52	.33
	Mean	.75	.12	.14	.49	.35

voice lures versus inconsistent voice lures. Previous research on recognition memory for spoken words has shown that voice information aids in veridical recognition of list items (Palmeri, Goldinger, & Pisoni, 1993). Therefore, we would expect that listeners should be more likely to identify a lure as old if it were presented in the same voice as the associated list.

Method

Subjects

Thirty-two Indiana University undergraduates participated in a one-hour session in partial fulfillment of an introductory psychology course requirement. All of the subjects were native English speakers and reported no speech or hearing impairments at the time of testing. The data from two of the subjects were omitted from the final analysis because they failed to follow instructions.

Materials

The study materials utilized in the present experiment were identical to the previous three experiments. The voices for each of the lists were presented in a blocked format.

Design

The study, recall, and recognition phases of Experiment 4 were identical to those of Experiment 2. The test items were counterbalanced so that half of the items were presented in a male voice and half in a female voice, and, of those, half were in the same voice as the test item and half were in a different voice as the test item. The 16 critical lures were presented in both male and female voices, half, were presented in the same voice as the associated study list and half were presented in a different voice as the associated study list.

Procedure

The procedure for Experiment 4 was identical to that of Experiment 2.

Results

Recall

Subjects recalled the critical nonpresented item 37% of the time. Subjects recalled the critical nonpresented items at about the rate of studied items presented in the middle of the lists.

Recognition

The recognition results are presented in Table 4. The overall hit rate for the study-recall condition (75%) was significantly different from the hit rate observed in the study-arithmetic condition (68%), $t(29) = 2.833$, $p<.05$. The false alarm rate for the nonstudied items was 11%.

Insert Table 4 about here

Table 4

Recognition results for studied items (top) and critical lures (bottom) for Experiment 4.

Item Type	Condition	Proportion of Old Responses				
		Overall	Same	Different	Don't Know	%Voice Attribution
Studied						
	Study + Recall Same	.76	.32	.10	.34	.55
	Study + Recall Diff	.73	.12	.28	.33	.55
	Mean	.75	.22	.19	.34	.55
	Study + Math Same	.69	.30	.07	.32	.54
	Study + Math Diff	.67	.14	.15	.38	.43
	Mean	.68	.22	.11	.35	.49
	Nonstudied	.11	.02	.02	.07	.36
Critical Lure						
	Study + Recall Same	.75	.17	.22	.36	.49
	Study + Recall Diff	.80	.05	.27	.48	.40
	Mean	.78	.11	.25	.42	.46
	Study + Math Same	.74	.21	.17	.36	.51
	Study + Math Diff	.72	.09	.25	.38	.47
	Mean	.73	.15	.21	.37	.49

An analysis on the overall effect of prior recall on recognition reveals a positive effect of recall (76% vs. 70%) $t(29) = 3.118$, SEM = 2.3, $p<.01$, on later recognition. The effect was significant only for the studied items; study + recall (75%) vs. the study + math (68%) conditions $t(29)$ 2.838, SEM = 2.3, $p<.01$; and not for the critical lure + recall (78%) vs. critical lure + math (73%) conditions $t(29) = 1.408$, SEM = 3.4, n.s. This replicates the effect from Experiments 1 and 2, but not Experiment 3. A look at the recall results in this experiment reveals that subjects recall levels (46%) are more similar to the levels observed in Experiments 1 and 2.

The recognition results for the critical nonpresented lures are shown at the bottom of Table 4. The recognition for the critical lures is even greater than the recognition of presented items (75% vs. 71%), however, this difference was not statistically significant, $t(29) = 1.546$, $p>.1$. Of interest, however, is the pattern of responses for the items that were presented in the same voice at study and test. Subjects tended to judge an item as old-same more often than old-different when the studied item and the test item were in the same voice for both the recalled lists (32% vs. 10%), $t(29) = 4.115$, SEM = 3.5, $p<.001$ and non-recalled lists (30% vs. 7%), $t(29) = 6.541$, SEM = 2.75 $p<.001$. However, this pattern was not observed for the critical lures in either recalled (17% vs. 22%), $t(29) = .865$, SEM = 3.85, n.s, or non-recalled lists (21% vs. 17%), $t(29) = .166$, SEM = 2.5, n.s. This result suggests that although illusory memories act like veridical memories in many ways, there are important fundamental differences that are assessed with different tasks. It might be the case that illusory memories are based on abstract representations, whereas veridical memories are based on episodic traces of events (Hintzman, 1986, 1988). Further research is necessary in order to determine how these two types of memories differ.

General Discussion

The four experiments reported here provide consistent evidence for false recall and false recognition. A small, but significant, positive effect of previous recall on recognition was observed in three of the four experiments. We also demonstrated that subjects are willing to attribute source judgments to nonpresented items. We discuss these results in relation to previous findings in illusory memory and implication for future research.

First, we obtained false recall effects that ranged from .29 to .38. Although these effects are impressive, the magnitude is slightly below the levels originally reported by Roediger and McDermott (1995) and others more recently (McDermott, 1996; Payne et al., 1996; Norman & Schacter, 1996; Schacter et al., 1996). One possible reason for the lower levels of false recall is that we reduced the amount of recall time after the 8 tested lists in our experiments from 90 seconds to 60 seconds. Because false recall often occurs towards the end of the subjects' responses (Roediger & McDermott, 1995), the shorter recall period may have had the unfortunate effect of reducing both veridical and false recall, but also possibly disproportionately harming the latter quantity. In retrospect, this reduction of recall time was unfortunate, but nonetheless, we did observe relatively high levels of false recall across the four experiments, with the weighted average being 36%. False recall in the Roediger-McDermott (1995) paradigm is robust, especially, in light of instructions to subjects not to guess and of the generally low level of other intrusions obtained in these experiments.

A second issue examined in the present experiments was whether recall (of studied and nonstudied items) affected later recognition. In general, the answer is yes, confirming the results of Roediger and McDermott (1995). This was the case in three of the four experiments. It should also be noted that overall recall in Experiment 3 (where no effect was observed) was greatly reduced compared to recall in the other three experiments. This low level of recall might account for the lack of an effect on recognition. A meta

analysis over all four experiments on the factors of recall (recalled x not recalled) and study (studied x not studied), revealed only a marginal effect of recall, $F(1,536) = 3.31$, $p = .06$. In summary, prior recall of a list seems to have a moderate but relatively consistent effect on later recognition. In addition, McDermott (1996) has shown that three tests have a more powerful effect than one test in boosting later recall of both studied and nonstudied items. The act of retrieval does seem to play a critical role in the development of false recall and probably of false recognition, too (see Roediger, et al., in press, for further elaboration).

The present experiments were specifically designed to give us further insight into our subjects' phenomenological experience during false recognition. Overall we found that listeners were as likely to attribute a voice to an illusory memory as a veridical memory. Furthermore, an analysis over all four experiments, reveals that subjects are more likely to label as "old" a critical lure than an actual studied item, $F(1,536) = 4.8$, $p<.05$. In general, false recognition was equal to or exceeded veridical recognition and the source judgments occurred with about the same frequency for both types of items. Further, this pattern held up over manipulations at encoding (presenting lists purely in one voice or mixed, within-list, as to voice) and over manipulations at retrieval (testing either with visual or auditory recognition tests). The finding that changing neither encoding nor retrieval factors greatly affected the levels of false recognition or voice attribution shows how powerful the memory illusion arising in the Roediger-McDermott (1995) paradigm is at promoting false recognitions. One might expect that subjects would be able to use differences in voice information (either when whole lists are presented in one or the other voice, or when test items are presented in the same or a different voice rather than being visually presented) as powerful retrieval cues to aid recognition. Yet, subjects apparently cannot use this information to lower levels of false recognition. However, it should be noted here that although voice information did not change the overall level of illusory recognition, patterns emerged which indicated that there is something fundamentally different between illusory and veridical memories. This was most evident in comparisons of Experiments 3 and 4. The results from Experiment 3 demonstrate that subjects illusory memories are influenced by physical attributes of the associated list. Subjects in the third experiment tended to identify the voice of the nonpresented item to be the same as the associated word list. However, the results from Experiment 4 show that when faced with an auditory recognition test subjects were less likely to make the same attribution. Further research is necessary to tease apart the differences between false and veridical recognition.

The levels of voice attribution observed here (voices were attributed for 30-40% of items called old) were much lower than those reported by Payne et al. (1996). However, as noted previously, Payne et al.'s paradigm differed considerably from the present recognition procedure, because they had subjects make judgments after the third of three successive recall tests and to respond to each recalled item by responding male, female or don't know to indicate their source attributions. In addition, our experimental procedures may have biased subjects to place their response criterion more conservatively to saying that an item was heard in a particular voice. We gave subjects the option of responding "old" with no voice attribution and it may have been that subjects used this option unless they were highly confident that they remembered the voice of presentation. Other testing procedures, such as having subjects always respond "old" for recognized items and then always to give the judgment of male, female, or don't know might cause them to be more willing to provide judgments of voice. This procedure would more closely approximate the methods used by Payne et al. We plan to adopt these procedures in future recognition experiments dealing with source attributions of voice of speaker.

Taken together with prior results, the present findings replicate and extend the conclusions that subjects' false recall and false recognition in the Roediger-McDermott (1995) paradigm are robust and that these false responses have, to the subjects, the same phenomenological characteristics as veridical recall and recognition of studied material. False recollections are often judged as remembered (Payne et al., 1996;

Roediger & McDermott, 1995) when Tulving's remember/know paradigm is used, meaning that subjects believe they remember features about the actual occurrence of test items that were not presented. Further, Norman and Schacter (1996) have shown recently that even when subjects are asked to write down some feature they remember from the study event, the levels of false recognition do not drop (relative to a condition in which subjects did not have to justify remember judgments in this way). In our experiments, and those of Payne et al. (1996), subjects attribute a presentation voice to critical lures at the same general level as actual presented items. This suggests that, at the very least, subjects reatained a general knowledge about base rates and frequency of occurrence of different voices during study. The challenge for future studies remains to specify why these items that are not presented are remembered as well as items that are presented.

References

Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: The University Press.

Goldinger, S.D. (1992). Words and voices: Implicit and explicit memory for spoken words. *Research in Speech Perception Technical Report No. 7,* Indiana University, Bloomington, IN.

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

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

Jones, T.C. & Roediger, H.L (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology*, **7**, 65-80.

Lockhart, R. (1975). The facilitation of recognition by recall. *Journal of Verbal Learning and Verbal Behavior*, **14**, 253-258.

McDermott, K.B. (1996) The persistence of false memories in list recall. *Journal of Memory and Language*, **35**, 212-230.

Norman, K. & Schacter, D.L. (1996). False recognition in younger and older adults: Exploring the characteristics of illusory memories. Manuscript submitted for publication.

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 & Cognition*, **19**, 309-328.

Payne, D.G., Elie, C.J., Blackwell, J.M., & Neuschatz, J.S. (1996). Memory illusions: Recalling, recognizing and recollecting events that never occured. *Journal of Memory and Language*, **35**, 261-285.

Read, J.D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review*, **3**, 105-111.

Roediger, H.L, & McDermott, K.B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.

Roediger, H.L., McDermott, K.B., & Goff, L. (in press). Recovery of true and false memories: Paradoxical effects of repeated testing. In M.A. Conway (Ed.) *Recovered memories and false memories*. Oxford: Oxford University Press.

Russell, W.A., & Jenkins, J.J. (1954). *The Complete Minnesota Norms for Responses to 100 Words from the Kent-Rosanoff Word Association Test* (Technical Report No. 11, Contract N8 ONR 66216). University of Minnesota.

Schacter, D.L., Verfaellie, M. & Pradere, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesic patients. *Journal of Memory and Language*, **35**, 319-334.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist*, **62**, 1-12.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Treatment Effects on Phonological Acquisition in a Cochlear Implant Recipient[1]

Amy McConkey Robbins[2] and Steven B. Chin

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

[1] This work was supported in part by NIH-NIDCD Training Grant DC00012 to Indiana University in Bloomington, NIH-NIDCD Research Grant DC00111 to Indiana University in Bloomington and NIH-NIDCD Research Grant DC00423 to Indiana University Medical Center in Indianapolis. The authors wish to thank Allyson I. Riley for her invaluable assistance in data collection and David B. Pisoni and Karen I. Kirk for helpful suggestions during manuscript preparation. We are also grateful to Susan Todd for help in the preparation of stimulus materials and to G.K.'s mother for her commitment to this project.

[2] Also DeVault Otologic Research Laboratory; Department of Otolaryngology–Head and Neck Surgery; Indiana University School of Medicine; Indianapolis, Indiana 46202-5200.

Abstract

This paper presents a case study of phonological development in a young child wearing a multichannel cochlear implant. The purpose of the paper is twofold; first, a description is given of the training program carried out with this child. Hallmarks of the program included: a "motor-loaded" emphasis in the early stages of training; a cognitive-linguistic approach in the later stages of training, including the use of speech contrasts which change meaning (Elbert, Rockman, & Saltzman, 1980); utilization of a modified Cycles approach (Hodson & Paden, 1991); and systematic exposure to many different talkers. The paper also presents longitudinal data reflecting the subject's speech production skills prior to implantation and at regular intervals thereafter. Evidence is given of phonological patterns similar to those found in the developing phonologies of normally-hearing children.

Treatment Effects on Phonological Acquisition in a Cochlear Implant Recipient

Introduction

Cochlear implants are viewed primarily as devices which aid speech perception in profoundly deaf individuals. The role implants play in enhancing speech production has begun to receive considerable attention as well (Tye-Murray, Spencer, & Woodworth, 1995; Tobey, Geers, & Brenner, 1994; Osberger, Robbins, Todd, & Riley, 1994). Of particular interest is the question of whether multi-channel cochlear implants provide sufficient information about the acoustic speech signal to influence the developing speech production skills of deaf children wearing these devices. One method of doing this is to identify in implanted children the presence of developmental phenomena which are known to be dependent on audition. If one were to find similar auditory learning trends in normally-hearing and implanted children, this would provide compelling evidence that cochlear implants do replicate, albeit imperfectly, certain aspects of normal hearing.

The focus of this paper is the development of speech production skills in a child wearing a multi-channel cochlear implant. The paper is divided into two sections. In the first, a description is given of the training program carried out with this child. The second section presents data reflecting the subject's phonological development prior to and following implantation. The unique contributions of audition and training to this child's phonological development are reviewed.

Subject

The subject of this study, G.K., was a profoundly deaf (see Appendix) youngster whose etiology of deafness was Waardenberg Syndrome. She was three years, five months at the onset of the study, three years, ten months when her cochlear implant was first activated, and six years, four months at the end of the study. Her hearing loss was identified at age 11 months, and at 12 months of age she was fitted with amplification and began a parent-infant training program using Total Communication. Limited progress was reported in speech development, in spite of her family's faithful participation in the training program and G.K.'s acceptance of appropriate and powerful binaural amplification. G.K.'s participation in our laboratory began in May 1992. As traditional amplification did not seem to provide adequate auditory information for her, G.K. was fitted with a seven-channel Tactaid 7 device, and utilized this device on a full-time basis for six months. The *TARGO* (*Tactaid 7 Reference Guide and Orientation*; Robbins, Hesketh & Bivins, 1993) was used as a structured tactile curriculum for two one-hour sessions per week.

When G.K. entered our study, she was able to produce on demand no recognizable English consonants and only two English vowels: /ɑ/ and /u/. She was not aware when her voice was on, and therefore frequently communicated by using articulatory postures without voicing. At her initial assessment, several speech production measures were taken. First, a six-minute spontaneous sample of her speech was video- and audio-taped and then analyzed according to a procedure described by Osberger, Robbins, Berry, Todd, Hesketh, and Sedey (1991). The results showed a preponderance of tokens in the "Other" category, which was described as articulatory behaviors that are non-productive and should be eliminated from the child's repertoire. These included blowing, lip smacking, use of ingressive plosion, clicks, and articulatory posturing without voicing.

A second speech production measure was also administered. The Nonsense Imitative Test using Syllables (NITS; see description below) required G.K. to imitate CV syllables following an examiner's model while an audio-tape was recorded. No subject in our laboratory (out of 180 subjects) ever produced fewer vocalizations as did GK during the administration of the NITS. She attempted to imitate mouth posture and produced several non-speech attempts, but no she had no notion of her voice being turned on.

In November 1992, G.K. received a Nucleus-22 multiple-channel cochlear implant (Cochlear Corporation). Full electrode-insertion was achieved, and activation took place in December 1992. From that point, speech training was designed to take advantage of the newly introduced auditory signal from the cochlear implant.

Training Program Characteristics

A multi-sensory approach to therapy was utilized with this subject, and consisted of many of the traditional speech and listening techniques commonly used in aural rehabilitation. These will not be described in this paper as they are well known to readers. The foundation of the training program, however, was characterized by four components not typically combined for use with hearing-impaired children. Table 1 summarizes the four components of the training program.

Table 1: Components of the Phonological Training Program

1. Early motor-loaded emphasis: motor approach used to establish basic speech production system.
2. Later emphasis on speech contrasts: emphasizes that differences in sound segments can signal differences in meaning.
3. Use of a modified Cycles approach: guides selection and ordering of target speech sounds for training.
4. Use of multiple talkers during training sessions: exposes child to natural variability occurring across talkers.

1. Early motor-loaded emphasis

In the earliest stages of therapy with G.K., emphasis was placed on the production, rather than the perception aspects of speech. Using a method described by Osberger (1983), speech targets were addressed first using imitation, generally with a visual prompt of some sort. Shortly thereafter, production-on-demand and perception tasks were introduced using the same speech targets. This approach diverges from that which has often been prescribed for hearing-impaired children, the integration of production and perception tasks, an approach which these authors often advocate. For example, Robbins (1994) recently outlined speech training guidelines for implanted children which emphasized the intertwining of speaking and listening activities. Our decision to utilize a different approach, the motor-loaded approach, with G.K. was based upon the extremely limited speech skills she demonstrated initially, and upon her lack of prior experience with sound. A compelling rationale for using such an approach is the outcome of speech training studies (Osberger, 1983; Osberger, Johnstone, Swartz, & Levitt, 1978; Lieberth & Subtelny, 1978; Subtelny, 1983) which suggests that the ability to produce a speech pattern may precede and assist in the hearing-impaired child's perceptual decoding of the corresponding auditory signal. Thus, the auditory-feedback loop may be more readily established if the initial emphasis is placed on the motor, rather than the perceptual aspects of speech. This seems to be especially true for very young deaf children who have no auditory experiences prior to implantation. These youngsters do not appear to have the prerequisite listening skills upon which to build a perceptually-based speech program. Rather, they first need practice with the production aspects of speech sounds. Once some competence is gained in producing certain speech

elements, more balance is needed in therapy, and perception and production tasks should be intertwined, an approach which makes efficient use of therapy time because several purposes may be accomplished with the same set of materials. Table 2 outlines the rationale for choosing a motor-loaded over an integrated approach in the initial stages of speech training.

Table 2: Rationale for Motor-Loaded Approach to Training

1. Child has no prior experience with sound.
2. Child does not demonstrate vocalization on demand.
3. Phonetic repertoire contains few or no recognizable vowels or consonants.

Motor-based approaches have been advocated by speech pathologists as treatment for speech errors which are articulatory, rather than linguistic in origin (Bernthal & Bankson, 1988). Our term "motor-loaded" is used to imply an approach which emphasizes the motoric aspects of phonology only on the front-end or early stages of training speech targets. The goal is to move on to a more cognitive-linguistic approach as soon as possible. The transition to a meaning-based approach was gradually made with G.K., as described below.

2. Later emphasis on speech contrasts that change meaning

This Cognitive-Linguistic approach was employed with G.K. later on in her rehabilitation program, once she had gained sufficient production practice with speech sounds. It represented a shift from the "motor-loaded" approach, described above, which was employed in the early stages of training. The use of speech contrasts in phonological training has been described by others, including Winitz (1975) who referred to it as "conceptualization training". The thrust of this approach is the child's use of speech sounds in a *contrastive* way. Therapy items include pairs of words with a target sound and a contrast sound. The target sound is the adult "form" of a word (such as 'him'), and this is contrasted to the child's incorrect production (such as 'hip'). This approach emphasizes the meaning that is conveyed by small changes in the sounds in words. Its premise is that the child will be more motivated to correct his speech errors if he realizes, through natural consequences, that those errors actually change the meaning of his message.

We began to employ meaningful contrasts within therapy as soon as G.K. was able to produce and perceive both the target and the contrast sound of a cognate pair. We used this method even while working on a "Motor-Loaded" approach with other sounds that needed production emphasis. Thus, the two approaches were not mutually-exclusive, but were used selectively depending on G.K.'s level of competence with the target sounds. For example, about six months post-implantation, we were addressing /d/ in therapy using a "motor-loaded" approach, as G.K. could not accurately and consistently produce this sound in a CV syllable on demand. At the same time, we were working on /m/ vs. /b/ using a contrastive approach ('Mom' vs. 'bomb'; 'my' vs. 'bye'; 'mall' vs. 'ball') because she had demonstrated the ability to produce both words in the pair, and to perceive the difference between them most of the time.

This type of contrastive approach would not have been practical to use with G.K. in the early stages of training due to the extremely limited nature of her phonetic repertoire. The approach is a logical one to use at a point in time when a child has the motoric skills to produce speech contrasts intentionally. As Bernthal and Bankson (1988) have noted, this type of approach does not dismiss the motoric requirements of the phonological system, but acknowledges that many speech errors are not motorically based. This was true in G.K.'s case in the area of final consonant deletion, for example. She failed to use a

word-final consonant in certain words, but used that same consonant word-initially, or even word-finally in other words. For example, 'tea' and 'teeth' were used in therapy as a contrastive pair because G.K. produced both as an open syllable, /ti/. Useful clinical materials for addressing this included *Language Approach to Open Syllables* (Young, 1981), *Contrasts: The Use of Minimal Pairs in Articulation Therapy* (Elbert et al., 1980) and *Contrastive Word Pairs* (Kiernan & Zentz, 1985)

3. Use of a modified Cycles approach

Hodson and Paden (1991) described an approach to therapy in normally-hearing children with delayed phonology. An underlying premise of this approach is that children actively engage in absorbing new information, generalize this, and slowly alter their phonological system over time. Although the approach was not designed specifically for hearing-impaired children, many of its characteristics make it a suitable approach for this population. These include: (a) auditory bombardment; (b) the development of accurate kinesthetic images for speech (what Ling refers to as an oro-sensory-motor code); and (c) the use of time periods or cycles of training which guide the selection of targets. These three aspects of the Cycles approach were employed with G.K. as follows:

Auditory Bombardment. When a new sound was chosen as a production target, we first exposed G.K. auditorially to many examples of the sound, via rhymes, repetitive babbling and single words. According to Hodson and Paden (1991), this initial step of auditory bombardment draws a child's attention to a new speech sound and isolates it to enhance its salience. For example, this technique was used with the high front vowel /i/, which G.K. produced as a central, nasalized vowel. Correct production of /i/ is a critical goal for deaf children, as it is considered a point vowel, involving maximum excursion of the tongue (to the high front position). In addition, it is an important sound for speech intelligibility, because of its high frequency of occurrence in English. Early on in therapy, G.K. could not approximate this vowel, so we avoided choosing words for practice which contained it. Rather, we used auditory bombardment to expose her auditorially to the sound, requiring her to focus on that specific vowel. Auditory bombardment continued even after G.K. was stimulable for /i/ as a way of providing acoustic highlighting of target patterns.

Kinesthetic Images. This aspect of the Cycles approach emphasizes the importance of how speech "feels" as well as what it sounds like. Hodson and Paden (1991) state that a speaker's ability to use correct sound patterns in connected speech is dependent upon continual self-monitoring, not only by listening to one's own productions, but also by the kinesthetic sensations associated with articulation. In the case of profoundly deaf children, this kinesthetic feedback is critically important because the children's limited residual hearing may not allow them full auditory access to speech sounds. Ling (1976) referred to this mechanism as the oro-sensory motor code. He postulated that auditory feedback of one's own voice, as well as awareness of what speech feels like, were vital to developing intelligible speech in deaf children. This is one step in the process of self-monitoring which eventually allows children to self-correct and to prevent incorrect productions.

We approached the introduction of each new speech sound in G.K.'s training as an opportunity to develop a new kinesthetic pattern, or to correct a faulty one. If G.K. was unable to produce an approximation of a target after numerous attempts, we temporarily abandoned work on that sound, so as not to give her "negative practice" producing incorrect patterns. Only after G.K. demonstrated stimulability for a given speech sound was it addressed in training. Our emphasis was always on the *correctness* of her productions of a given sound. For example, when G.K. did become stimulable for /i/, she could produce it correctly only in the syllable /bi/. Therefore, that was the only context in which we addressed /i/, until she became stimulable for it in other contexts. In this way, her practice of /i/ was always 100% accurate, reinforcing the development of a correct kinesthetic pattern.

Use of time periods or cycles for training. Although the goal of the Cycles approach is to change the phonological patterns of a child, what must be targeted are specific phonemes or phoneme sequences

that represent those patterns. The Cycles concept is meant to capitalize on a child's normal ability to generalize and takes into account that phonological acquisition is gradual. For these reasons, sounds are introduced for training, worked on, and then left, and other sounds are addressed. During "off" training times, it is hypothesized that some aspects of the target patterns may be incorporated into the child's phonology and/or may generalize to other targets. Sounds are generally "recycled" and addressed in future cycles until they are mastered. As stated earlier, only sounds for which the child is stimulable become targets for production training. In addition, phonetic contexts for targets are carefully chosen to enhance the likelihood of production accuracy.

These principles guided the selection and ordering of target sounds for G.K. and the contexts in which they were trained. For example, when G.K. was first stimulable for /m/, we avoided eliciting it in syllables with oral labials, such as /bɑm/ or /mæp/, as the nearby non-nasal consonants appeared to cause denasalization of the /m/. Similarly, her productions of /m/ were correct when adjacent to the rounded (labial) vowel /u/ but incorrect when adjacent to /ɑ/.

4. Use of multiple talkers during training sessions

The rationale for varying the talker was based upon the work of Logan, Lively, Pisoni, and colleagues (Lively et al., 1992, 1993, 1994; Logan et al., 1991, 1993). These investigators examined the progress of Japanese subjects who were learning to perceive and produce distinctions between /r/ and /l/. Subjects who listened to multiple talkers during their course of training showed better accuracy and generalization of skills than subjects trained with a single talker. These results suggested that variability during perceptual training was important for helping the listener establish broad, rather than narrow and rigid, phonetic categories.

Therapy sessions were set up so that a relatively large number of different people were involved, on a rotating basis. These included two therapists, the subject's mother, and her four siblings. A number of other laboratory staff also participated in the therapy sessions from time to time, thereby exposing G.K. to different talkers whose speech varied from each other across a number of dimensions, including rate, loudness, and fundamental frequency. As part of the "dialogue" rather than "tutor" format (Blank & Marquis, 1987; Robbins, 1994) which was utilized in therapy, the roles of speaker and listener were switched often. This meant that stimuli presented during listening activities were sometimes spoken by G.K.'s mother, other times by a clinician, still other times by a sibling, and finally, by G.K. The latter occurred during activities in which G.K. assumed the role of "teacher", where she produced the stimulus and a listener responded. Thus, this subject experienced the considerable variability that occurs across talkers by virtue of listening to so many different speakers. It was hypothesized that this experience with variability would make G.K.'s perception skills more flexible due to the breadth of her phonetic categories.

Phonological Measures

Nonsense Imitative Test using Syllables

Two measures were used to assess G.K.'s phonetic-phonological ability. The first, the Nonsense Imitative Test using Syllables (NITS, developed at the DeVault Otologic Research Laboratory, Indiana University School of Medicine) assesses the ability to imitate 68 CV syllables three times each. Productions are assessed first in terms of their speech characteristics ("speech, speech-like, non-speech") and second in terms of syllable shape and correctness of production of constituent sound segments. The first 11 items on the NITS consist of syllables formed of the voiced bilabial stop /b/ followed by the monophthongs /i ɪ ɛ æ u o ɑ/ and the diphthongs /eɪ aɪ au ɔɪ/. The remaining 57 items on the test consist of 19 English phonemes followed by the point vowels /i ɑ u/. The second, described in more detail below, assessed production on demand through presentation of pictures of people familiar to the child.

The NITS was administered twice pre-cochlear implant and at regular intervals post-implant. It was first administered 6 months before implantation; at that interval, G.K.'s renditions of syllables were characterized almost exclusively as "non-speech", although there were isolated and apparently random occurrences of vowel-like productions approximating the extremely open /ɑ/. There was a uniform lack of voicing (necessary for correct production in every case, given the vocalic nucleus). G.K. appeared able to imitate bilabial articulation for a few target segments, but these postures were either released ingressively or were otherwise unaccompanied by either egressive pulmonic airflow upon release or by a following voiced vowel. In sum, there appeared to be lack of awareness of an articulation/sound relationship (speech), including voicing, and thus of syllable structure.

During the same month as, but subsequent to, the first administration of the NITS, G.K. was fitted with a seven-channel Tactaid-VII device, which she used for approximately six months. The second administration of the NITS took place subsequent to this six-month period and within a week prior to implantation. Rating of this administration revealed some improved knowledge of the articulation/sound relationship, with very few productions assessed as "non-speech", some as "speech", and most as "speech-like". Still lacking was evidence of knowledge of syllable structure, since almost all speech productions were isolated consonants or vowels. Production of consonants was generally limited to visually-salient ones, primarily labials, while vowels were uniformly low (open) ones (/æ ʌ ɑ/). Some productions still consisted solely of a completely unimpeded pulmonic egressive airstream forced through protruded lips, and all vowels were produced with creaky voice.

Implantation took place six days subsequent to the NITS administration just discussed, and the first post-implant NITS was administered six months later. Productions at this administration were almost all characterized as "speech", with very few (approximately 6%) characterized as "speech-like"; importantly, no productions were characterized as "non-speech". Nonspeech gestures such as blowing and smacking in place of speech productions, as well as creaky voice, were eliminated. Consonant productions consisted mainly of voiced bilabial stops, although there were isolated occurrences of stops at other places of articulation, as well as some instances of fricatives. Vowel productions included both low (open) vowels (as occurred pre-implant), but also some high (close) vowels such as /i ʊ/. Correct production of gross syllable structure (with segmental substitutions) obtained in virtually all cases for syllables initiated by labials; for other places of articulation, however, incorrect productions occurred with greater frequency.

Administration of the NITS at 2.0-years post-implant showed further development of speech, segmental, and syllabic characteristics. All productions were characterized as "speech", and except for some isolated cases, the CV syllable shape was produced correctly. Consonant production heavily favored voiced bilabial and alveolar stops, with some instances of fricative production. Vowel production showed evidence of awareness of height distinctions, so that the point vowels /i ɑ u/ were correctly produced in most cases. The situation 2.5 years post-implant was similar, with production of the correct syllable shape, of both voiced and voiceless consonants, of consonants at various places of articulation (including visually non-salient ones), and correct point vowels.

Picture-Naming Task

Steady improvement in performance on the NITS from pre-implant to post-implant intervals and through the post-implant intervals served as evidence of a growing awareness of the relationship between phonatory/articulatory gestures and sound production. Although the NITS is thus a good indicator of the degree of motoric ability to imitate speech, it nevertheless leaves aside the linguistic question of the relationship between sound and meaning. By 2.0 years post-implant interval, it was clear that G.K.'s speech production abilities had sufficiently progressed, so that another measure, of phonological production ability, was administered. Full details regarding this instrument, its administration, and results can be found in Chin, Pisoni, and Svec (1994), which is summarized here.

For this production-on-demand task, a set of 23 pictures was made, which depicted the faces of persons known to the child, including friends, teachers, and family members; these pictures served as stimulus items to elicit speech productions of each of the names (see Table 3) associated with the pictures. The use of names, rather than common nouns, addressed the question of whether the child's phonological (as opposed to sign) lexicon was of sufficient size to permit testing. It was predicted that at least names of familiar persons would be phonologically represented and available for speech production. This prediction was borne out at the first elicitation session.

During regular therapy sessions, the full set of 23 pictures was displayed one at a time to the child, who was asked to say aloud the name of the person depicted in the picture. When necessary, the child was prompted to say the name with a question such as "Who's that?" or "Can you tell me who this is?". No direct training on the names, except acknowledgment repetition during sessions or normal use of the names in everyday use was provided. The child's productions were tape-recorded and narrowly transcribed phonetically by a trained clinical phonologist. Five elicitation sessions were held near the 2.0-year interval, the first at 647 days and the last at 693 days post-implant.[3]

Table 3: Names for Picture-Naming Task

Orthography	Target Transcription	Orthography	Target Transcription
Alfred	['æl.frɨd]	Kris	[kʰrɪs]
Alice	['æ.lɪs]	Kristin	['kʰrɪ.stĩn]
Allyson	['æ.lɪ.sɨ̃n]	Marge	[mɑrdʒ]
Amy	['ẽĩ.mi]	Nick	[nɪk]
Carly	['kʰɑr.li]	Patty	['pʰæ.ɾi]
Cathy	['kʰæ.θi]	Sarah	['sɛ.rə]
Debbie	['dɛ.bi]	Sawyer	['sɒɪ.jɚ]
Diana	[daɪ.'ɐ̃.nə]	Shanan	['ʃæ.nɨ̃n]
Dwayne	[də.'wẽĩn]	Tara	['tʰɛ.rə]
Haley	['heɪ.li]	William	['wɪl.jə̃m]
Josh	[dʒɑʃ]	Yvonne	[jɨ.'vɑ̃n]
Kim	[kʰĩm]		

[3]Less accurately reported in Chin et al. (1994) as 685 days and 731 days post-implant.

Although the nature of the stimulus set did not permit probing of all English segments, as Table 4 shows, for consonants, there was fairly good coverage of voicing and various manners and places of articulation. Table 5 shows the vowel segments contained in the probe list; although front and central vowels were well represented, there were, unfortunately, no occurrences of back rounded vowels in the names contained in the list. In addition to a relatively wide variety of segment types, the probe list of names also contained a variety of syllable structure types; importantly, it contained both opened and closed syllables. Results from the NITS had shown good control over production of CV syllables, and the picture-naming task further tested G.K.'s ability to combine consonants and vowels. Table 6 contains schematics and examples of the various syllable types contained in the names.

Table 4: Target Inventory of Consonants in Picture-Naming Task

Stops	pb			td		k[]		
Fricatives		fv	θ[]	s[]	ʃ[]			
Affricates					[]dʒ			
Nasals	m			n		[]		
Liquids			l	r				
Glides	w				j		h	

Note: Phonetic symbol indicates that a segment was contained at least once in the name elicitation probe. Empty brackets indicate that the English segment at that position was absent from the probe.

Table 5: Target Inventory of Vowels in Picture-Naming Task

	Front	**Central**	**Back**	**Diphthongs**		**Rhotacized**
High	i ɪ	ɨ				
Mid	ɛ	ə			eɪ	ɚ
Low	æ	ɑ		aɪ	ɒɪ	

The inventories of consonants and vowels produced in the five sessions are shown in Tables 7 and 8. At this early stage of development, the criterion used for inclusion of a segment in the phonetic inventory was one occurrence in any of the five elicitation sessions. Given the fairly close temporal range from first to last elicitation session (approximately 45 days, with from 5 to 22 days between sessions), it is possible to consider these sessions as repeated trials within a single interval rather than as longitudinal intervals.

Table 6: Target Syllable-Structure Types in Picture-Naming Task

Syllable Structure Type		Example
Open:	V	[æ] in 'Allyson'
	CV	[dɛ] in 'Debbie'
	CCV	[kʰrɪ] in 'Kristin'
	CVV	[daɪ] in 'Diana'
Closed:	CVC	[wɪl] in 'William'
	CCVC	[krɪs] in 'Kristin'
	CVVC	[wẽĩn] in 'Dwayne'
	CVCC	[mɑrdʒ] 'Marge'

Consonant inventory. Table 7 indicates that G.K.'s consonant inventory exhibited a voicing distinction, as well as a full range of segments of different places and manners of articulation. Although the consonant inventory was relatively large, it was also true that it was not precisely an English inventory. One reason was the absence of specific segments contained in the probe list (e.g., /r/, which, however, also presents problems for hearing children). The lack of some segments from the probe list makes it somewhat difficult to assess their absence from G.K.'s productions (although she did produce some English segments that were absent from the probe list); this is especially true in the case of voiced fricatives. Second, there were a number of segments present in the inventory that do not occur in the English inventory. These included stops [ɓ ɗ], central fricatives [ʂ ʐ ʂ ç], and a lateral fricative [ɬ].

Vowel inventory. As indicated in Table 8, the situation for G.K.'s vowel inventory was similar to that of the consonant inventory. Specifically, there was a wide range of vowels produced, including vowels of different tenseness, height, and backness. In general, G.K. was able to produce most of the vowels that were contained in the probe list (although not always in the right place); again, the absence of some vowels from the probe list (particularly back vowels) makes it difficult to assess their absence from G.K.'s inventory. As was the case with the consonants, there also occurred some non-English vowels (including nasalized vowels produced in the absence of following nasal consonants) and a front rounded vowel.

Table 7: Production Inventory of Consonants from Picture-Naming Task

Stops	bb̪															k	ʔ
Fricatives			fv		θ			s̪		s	z̪				ʃʒ	ç	
Lateral Fricative										ɬ							
Affricates		pf	bv					t̪s̪		ts	dz		tʂ		tʃ	dʒ	
Nasals			m								n						
Liquids											l						
Glides			w	w̃											j		h

Table 8: Production Inventory of Vowels from Picture-Naming Task

	Front	**Central**	**Back**	**Diphthongs**
High	i ɪ			
Mid	e ẽ ɛ œ	əə̃	ɔ	eɪ əɪ
Low	a	ɑ		aɪ ao

Syllable structure. Phonological systems of hearing children at early stages of development very often have limits on the types of syllable structures that can occur, and very often the only syllable structure type that does occur is a consonant followed by a vowel (CV syllable, such as probed in the NITS). In addition to this simple syllable structure type, relatively well-developed phonological systems also exhibit more complicated structures, especially (1) syllables that end in consonants (i.e., "closed" syllables) and (2) syllables that either begin or end with clusters of consonants (CCV– or –VCC). As Table 6 indicates, the names on the probe list contained both relatively simple syllable structures and relatively complicated syllable structures.

In fact, G.K. was able to produce both simple and complicated syllable types, as indicated in Table 9. The forms in Table 9 indicate that G.K. was able to produce both open and closed syllables (e.g., CV vs. CVC), both light and heavy syllables (e.g., CV vs. both CVV and CVC), and both singletons and clusters (e.g., CV vs. CCV, and CVC vs. CVCC).

Table 9: Production Syllable-Structure Types from Picture-Naming Task

Syllable Structure Type		Example
Open:	V	[ɑ] in 'Allyson'
	CV	[dɑ] in 'Debbie'
	CCV	[kwɪ] in 'Kristin'
	CVV	[dəɪ] in 'Diana'
Closed:	CVC	[nɪç] in 'Nick'
	CCVC	[fwɪs̪] in 'Kristin'
	CVVC	[feɪʃ] in 'Dwayne'
	CVCC	[mɔʒʃ] 'Marge'

Substitution patterns. During the course of phonological acquisition by hearing children, there are normally a number of apparent substitutions of one sound for another (or "correspondences" between these sounds). In many cases, these are systematic, in that (1) all instances of a target sound have a common substitute sound (e.g., [t] for [k] in all cases, and (2) substitutions affect natural classes of sounds rather than individual sounds (e.g. stops substituting for fricatives). In addition, many substitution patterns are common across a number of children (e.g., alveolars for velars), while others are less common (e.g., velars for alveolars). G.K. exhibited a number of substitution patterns that were both systematic (to a degree) and common. These included unaspirated stops for aspirated ones, stops or affricates for fricatives, an alveolar stop for a lateral. One substitution pattern showed a voiceless alveopalatal fricative [ʃ] for a number of target nonlabial segments, including [s t k].

Lexical representations. A clear indication of acquisition of an ambient phonological system is consistently correct production of forms in a variety of linguistic environments. Just two years post-implant, G.K.'s phonological system could not be expected to be English-like in all aspects, but some signs of awareness of a sound-meaning relationship was the ability to produce forms without imitation and the relative stability of at least some lexical representations across trials. Additionally, a number of productions were very close to being correct. Thus, of the 23 names elicited, the following were close to correctly produced and furthermore appeared to be fairly stable: 'Amy, Debbie, Haley, Nick, Patty, Sarah, Tara'. Productions of these forms showed fairly small distances from the intended target in terms of marking of segmental slots (maintenance of syllable structure) and phonetic and phonological similarity between target and response. Other productions were fairly stable across trials, but segments showed greater phonological distances from targets; these were 'Alfred, Alice, Cathy, Dwayne, Josh, Kris, Kristin, Marge, William'. The remaining names of the list were produced in such a way as to indicate either relatively unstable lexical representations or greater phonological distance between target and response.

Thus, on a number of measures (segment inventory, syllable structure, substitution patterns, lexical representations), G.K.'s phonological system appeared to be developing in the desired direction.

Treatment Effects

The goal of the treatment program incorporating the four characteristics discussed previously (early motor-loaded emphasis, later cognitive-linguistic approach, modified 'Cycles' approach, and use of multiple talkers) was to facilitate speech production in such a way as to enhance the overall development of a target-appropriate phonological system, as well as the development of communicative abilities. Although it is probably too early to judge the overall effects of the treatment program on phonological and communicative development, specific components of the treatment program appear to be at least conceptually related to various characteristics of G.K.'s present speech production abilities. Evidence for such relationships comes from both observations made during treatment sessions as well as data collected from the NITS and picture-naming task. We review some of these relationships below:

1. Motor-loaded emphasis

Recall that this approach was chosen early on in therapy due to the extremely limited phonetic repertoire that G.K. demonstrated. Its goal was to allow G.K. extensive and purposeful motor practice with new speech targets, in order to expand her phonetic repertoire. Evidence of the usefulness of this approach may be seen when comparing the number of sounds (9 vowels, 14 consonants) G.K. produced on imitation at her 6-months post-implant interval to those she produced prior to the initiation of training and fitting of the implant (3 vowels, 6 consonants).

2. A cognitive-linguistic approach in later stages of training

Evidence of the effectiveness of this technique may be found in G.K.'s ability to learn, relatively quickly, new contrastive pairs with similar patterns to earlier-trained pairs. For example, the contrast 'row' vs. 'rope' was introduced within the first year of training and the presence of a word-final consonant was tied to the meaning conveyed. Once this pair was mastered, G.K. required very little training to distinctly produce 'hoe' vs. 'hope', 'sew' vs. 'soap', and later, 'toe' vs. 'taupe'. The word-final /p/ also generalized to pairs with different preceeding vowels (e.g., 'Sue' vs. 'soup' and 'key' vs. 'keep'), suggesting that the sound was not constrained to appear only in the context of certain vowels.

Although minimal pairs such as those just cited are a traditional indicator of the use of sounds or features contrastively, other evidence can be adduced to show linguistic use of various phonological structures. In the picture-naming task, for instance, there were near-minimal pairs such as 'Cathy/Patty' and 'Amy/Debbie', as well as pairs of names differing in the number of syllables, such as 'Kris/Kristin'. Differential production of the members of these pairs serve as evidence of the awareness of the need to mark contrasts (even if not target-appropriately). Thus, across the five repetitions of 'Cathy', productions of the initial consonant differed from those of renditions of 'Patty'. Similarly, 'Amy' was always produced with a nasal intervocalic consonant, whereas 'Debbie' was produced with an oral consonant. Finally, 'Kris' was always produced as a monosyllable, while the related form 'Kristin' was always produced with two syllables.

3. Use of a modified "Cycles" approach

Elements from the "Cycles" approach were utilized as a systematic way of determining which sounds to introduce into training, and when. As stated earlier, Hodson and Paden (1991) recommend that sounds not be addressed in drills until the child can at least produce a correct approximation of the sound. This avoids the negative practice that may result from drilling sounds that are incorrectly produced by the child. The Cycles approach is time-efficient in that no time is wasted in working on sounds the child cannot produce.

G.K. showed excellent retention of correct productions over time, which appeared to be a result of the use of the Cycles approach. Some speech training methods involve introducing sounds for training based upon a pre-determined schedule, regardless of the child's ability to approximate those sounds. This often results in a child's being able to produce a sound during the time it is being directly drilled, but yielding no carry-over over time. Thus, speech sounds are "lost" from the child's repertoire anytime they are not being actively trained.

Evidence for retention of phonological knowledge was to be found in both the NITS and picture-naming task results. Successive post-implant administrations of the NITS showed not only the addition of new sounds but also the retention of old ones. In the picture-naming task, correct productions of sounds contained in the names persevered across sessions, indicating stability of both the sounds themselves and the lexical representations in which they were contained.

4. Use of multiple talkers

The benefit of this technique with G.K. may be seen both in informal observation and in empirical test results. Informally, G.K. is observed to adjust easily to new and different speakers, including those with foreign accents and different dialects of English. Her flexibility with different talkers is seen empirically in her periodic testing, conducted every 6 months. This testing was always performed by an examiner who did not work regularly with G.K. and the examiner varied from test interval to test interval. That G.K.'s speech perception and production scores improved steadily over time (with no evidence of a plateau in performance) is testimony to the fact that her skills were not tied to one talker. The set of names used in the picture-naming task were introduced simply in the course of daily living and, importantly, were introduced from multiple sources by different people. G.K.'s ability to associate names with faces and to produce them, in many cases, correctly attests to the viability of using multiple talkers in training.

A recently-designed stimulus set will be administered to G.K. in the near future. The set consists of two types of recorded word lists. On the first, the child hears one talker say all the stimuli, whereas on the second, she hears multiple talkers (Kirk, Pisoni, Osberger, 1995). Comparison of G.K.'s performance on the single- vs. multiple-talker lists will further define the extent of her flexibility with different talkers.

Conclusions

In this paper, we have outlined a therapeutic program designed to advance development of a phonological system by a child with a cochlear implant. This program contains four main component characteristics, each of which targets a specific area of the phonological system: (1) early motor-loaded emphasis to establish a basic speech production system, (2) later emphasis on speech contrasts to increase awareness that different sound segments can signal differences in meaning, (3) the use of a modified Cycles approach to guide the selection and ordering of specific training targets, and (4) the use of multiple talkers to training sessions to expose the child to the natural variability that occurs across speakers in the community. Although the child described here is still at a relatively early stage in the development of a spoken linguistic system generally and a target-appropriate phonological system specifically, preliminary speech production data from periodic intervals following implantation indicate that phonological abilities are developing in the desired direction, in terms of phonetic inventory and syllable structure. Additionally, this child's lexical representations indicate an awareness of the basic linguistic principle that specific combinations of sounds correspond to specific meanings.

It would perhaps be premature to advance here the strong claim that the treatment program we have described is solely (with the exception of the prosthesis itself) responsible for whatever advancements in phonological ability this child has achieved; we would hope, however, that the descriptive report we offer here is sufficiently detailed to allow other researchers to compare both their own treatment methods and measures of phonological ability with ours.

References

Bernthal, J.E., & Bankson, N.W. (1988). *Articulation and Phonological Disorders.* Second edition. Englewood Cliffs, NJ: Prentice-Hall.

Blank, M. & Marquis, M.A. (1987). *Directing Discourse: 80 Situations for Teaching Meaningful Conversation to Children.* Tucson, AZ: Communication Skill Builders.

Chin, S.B., Pisoni, D.B. & Svec, W.R. (1994). An emerging phonetic-phonological system two years post-cochlear implant: A preliminary linguistic description. *Research on Spoken Language Processing*, **19**, 253–270.

Elbert, M., Rockman, B., & Saltzman, D. (1980). *Contrasts: The Use of Minimal Pairs in Articulation Training*. Austin, TX: Pro-Ed.

Hodson, B.W. & Paden, E.P. (1991). *Targeting Intelligible Speech*. Second edition. Austin, TX: Pro-Ed.

Kiernan, K.L. & Zentz, B.W. (1985). *Contrastive Word Pairs*. Baltimore, MD: K.Z. Associates.

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

Lieberth, A. & Subtelny, J. (1978). The effect of speech training on auditory phoneme identification. *The Volta Review*, **80**, 410–417.

Ling, D. (1976). *Speech and the Hearing-Impaired Child: Theory and Practice*. Washington, DC: Alexander Graham Bell Association for the Deaf.

Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, **94**, 1242–1255.

Lively, S.E., Pisoni, D.B., & Logan, J.S. (1992). Some effects of training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 175–196). Tokyo: Ohmsha Publishing.

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.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, **89**, 874–886.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1993). Training listeners to perceive novel phonetic categories: How do we know what is learned? *Journal of the Acoustical Society of America*, **94**, 1148–1151.

Osberger, M.J. (1983). Development and evaluation of some speech training procedures for hearing-impaired children. In I. Hochberg, H. Levitt, & M.J. Osberger (Eds.), *Speech of the Hearing Impaired: Research, Training, and Personnel Preparation* (pp. 333–348). Baltimore, MD: University Park Press.

Osberger, M.J., Johnstone, A., Swartz, E., & Levitt, H. (1978). The evaluation of a model speech training program for deaf children. *Journal of Communication Disorders*, **11**, 293–313.

Osberger, M.J., Robbins, A.M., Berry, S.W., Todd, S.L., Hesketh, L.J., & Sedey, A. (1991). Analysis of spontaneous speech samples of children with cochlear implants or tactile aids. *The American Journal of Otology*, **12**(Supplement), 151–164.

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

Robbins, A.M. (1994). Guidelines for developing oral communication in children with cochlear implants. *The Volta Review*, **96**, 75–82.

Robbins, A.M., Hesketh, L.J., & Bivins, C. (1993). *Tactaid 7 Reference Guide and Orientation*. Somerville, MA: Audiological Engineering Corporation.

Subtelny, J.D. (1983). Patterns of performance in speech perception and production. In I. Hochberg, H. Levitt, & M. J. Osberger (Eds.), *Speech of the Hearing Impaired: Research, Training, and Personnel Preparation* (pp. 215–230). Baltimore, MD: University Park Press.

Tobey, E., Geers, A., & Brenner, C. (1994). Speech production results: Speech feature acquisition. *The Volta Review*, **96**, 109–129.

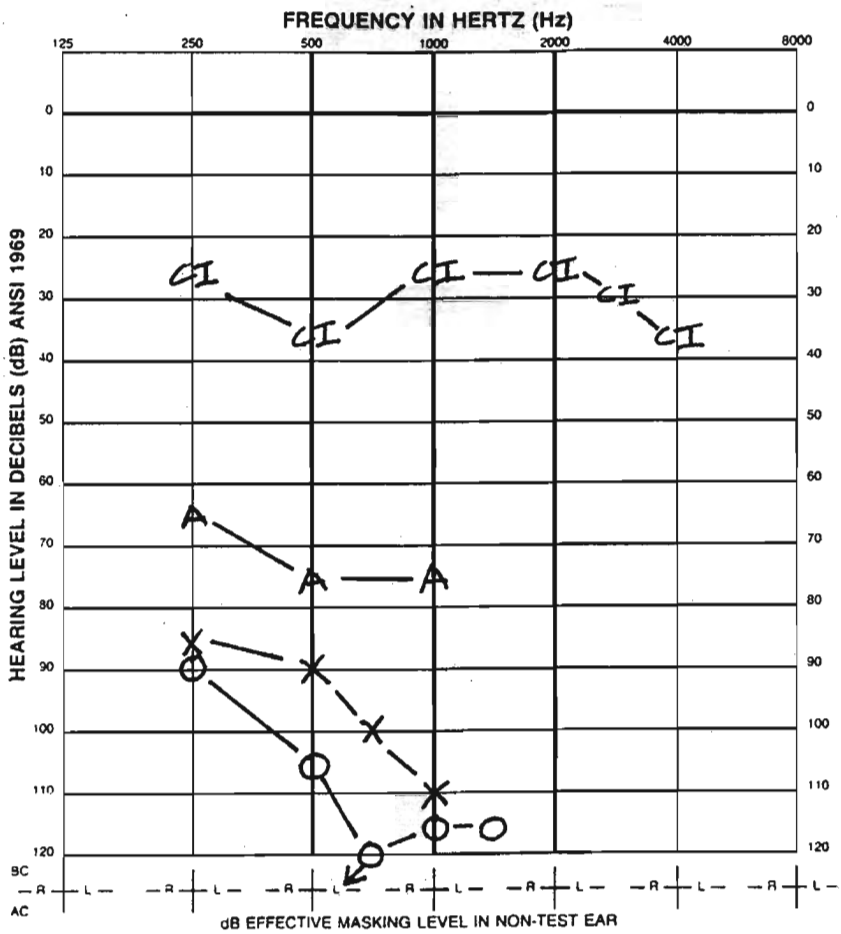
Tye-Murray, N., Spencer, L., & Woodworth, G. G. (1995). Acquisition of speech by children who have prolonged cochlear implant experience. *Journal of Speech and Hearing Research*, **38**, 327–337.

Winitz, H. (1975). *From Syllable to Conversation*. Baltimore, MD: University Park Press.

Young, E.C. (1981). *Language Approach to Open Syllables*. Tucson, AZ: Communication Skill Builders.

Appendix

Combined audiogram for unaided (X, O; headphone response), hearing-aided (A), and cochlear-implant (CI; sound-field response) conditions for G.K.



RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Normalization of Vowels by Breath Sounds[1]

Douglas H. Whalen[2] and Sonya M. Sheffert

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

[1] This research was supported by NICHD Grant HD-01994 to Haskins Laboratories and NIH-NIDCD Training Grant DC00012 to Indiana University. We thank Julia Irwin and Larry Brancazio for help with the experiments. Carol A. Fowler and Leigh Lisker provided helpful comments.

[2] Haskins Laboratories, 270 Crown Street, New Haven, CT 06511

Abstract

Listeners use multiple sources of information to normalize vowels for speaker characteristics, both within the utterance and outside of it. The present study was designed to test whether the sound of a breath intake (inspiration) could serve as another source for normalization. Breath sounds are external to the utterance, but can potentially convey two kinds of information: speaker vocal tract characteristics and the upcoming shape of the vocal tract (to the extent that formants might be realized in the inspiration). We tested effects of natural inspiration noise from two speakers (male and female) and three vowel environments on the perception of two synthetic vowel continua. We found that speaker sex affected one continuum, while vowel context affected the other. Although further exploration of this phenomenon is needed, these results suggest that even inspiration noise can provide linguistic information.

Normalization of Vowels by Breath Sounds

> "Last night, the phone rang just before midnight, and only half awake, she answered it. There was only breathing on the other end, but instead of hanging up she lay there in the dark, listening to it. She was sure it was Philippe, but did she know the sound of his breathing that well? Could the sound of one person's breathing be that recognizable? Is it like fingerprints--no two breaths are the same?" (Davis, 1994:237).

The acoustic realization of speech sounds depends greatly on the vocal tract that produces them. Adult speakers differ in their vocal anatomy, especially across the sexes, and every listener must take such differences into account. Additionally, every adult begins life as a child ("I started out as a child," as Bill Cosby used to say), and productive language begins as early as one year of age, well before the head has reached its full, adult size. So each speaker must be able to normalize his or her parents' speech in order to know what they are supposed to learn. And one's own speech must be normalized for the current size of the vocal tract, not to the size during the early stages of acquisition. Given such variation, humans have had to make use of a communicative system that allows listeners to make compensations for changes in speakers. Listeners clearly do make adjustments, most clearly in the case of vowels (Nearey, 1989), but also with stops (Rand, 1971) and fricatives (May, 1976; Strand, 1995). Thus speaker normalization must be considered a feature of speech perception.

The sources of normalization have been found to be both intrinsic to the signal and extrinsic to it. Intrinsic factors include the ratio of the formants (Chiba & Kajiyama, 1941; Potter & Steinberg, 1950), F0 of the vowel (Syrdal & Gopal, 1986; Miller, 1989), the spacing of F3 and F4 (Fujisaki & Kawashima, 1968), and the shape of the upper range of the spectral envelope (Kitamura & Akagi, 1994). Extrinsic factors include a speaker's overall range of formant values (Gerstman, 1968), presentation of point vowels (Ainsworth, 1975), the sentence preceding the item to be identified (Broadbent & Ladefoged, 1960; Remez, Rubin, Nygaard, & Howell, 1987), and even the assumed sex of the speaker as given visually (Strand, 1995). As Nearey (1989) points out in his overview of this issue, the existence of intrinsic effects has often led to the insistence that only intrinsic effects should be considered, despite the experimental evidence of extrinsic effects. This tendency is heightened by the fact that the influence of extrinsic factors may be limited to ambiguous stimuli, in light of the lack of an effect of speaker information on error rates in identifying naturally produced vowels (Verbrugge, Strange, Shankweiler, & Edman, 1976). Intrinsic methods of normalization also have the advantage for computational treatments that they can be applied to any stretch of the speech signal without taking the rest into account. But the evidence is fairly clear that extrinsic factors must play a role, and so some combination of extrinsic and intrinsic normalization must occur.

A further question that remains unresolved is whether normalization is carried out on every vowel (and perhaps on accompanying consonants) or if the listener sets up a working model of the speaker in order to interpret what is heard. Recent evidence from short-term memory studies shows that it is more difficult to remember lists spoken by several speakers than those spoken by a single speaker (Martin, Mullennix, Pisoni, & Summers, 1989; Goldinger, Pisoni, & Logan, 1991). One possible interpretation of this effect is that the need to construct models of the different speakers is more time-consuming that constructing one for a single speaker, and thus the resources available for the memory task are reduced. More explicit testing for such "model construction" is necessary, including testing whether intrinsically ambiguous vowels (i.e., ones that cannot be classified without knowing which vocal tract produced them) are treated unambiguously in a single speaker context, and whether the model can be retained for some

length of time and reactivated either automatically or by explicitly cueing. The alternative to model construction is a more general set of context effects, in which speech is adapted to its current environment but in a way that relates various acoustic properties to each other, not to a proposed speaker. Such models would not predict that any gain would accrue from earlier exposure to a speaker.

Many of the normalization effects rely on features of the acoustics that are not available to consciousness on the part of the listener. The intensity differences due to vowel quality, for example, are fairly automatically taken into account in perception (Lehiste, 1970). Other features, such as speaker sex, are fairly easily reported, even if the exact effect of the normalization is still unavailable for listener report. If the model construction approach is correct, we would expect that it would be easy to manipulate the level of consciousness attainable by manipulating how much the listener "knows" about the speaker, including visual appearance, name, and "personal" facts (however fictitious!). Again, these experiments remain to be done, but there is an interesting discrepancy between what listeners can report and what they can actually do.

One interesting example of this discrepancy between performance and report is in the judgment of speaker height and weight based only on acoustic speech information. Listeners are quite willing to perform this task, and generally report a moderate level of confidence in their answers (Lass, Phillips, & Bruchey, 1980). Even novelists have noted the expectation of a correlation: "She walked past Sara with a quick 'Hi,' hugged Anthony, and said, 'Hello, darling,' in a voice that had more height than she did" (Davis, 1994: 263). However, subjects are terrible at the task. One reason for that poor performance, it turns out, is an almost complete lack of correlation between any measurable speech characteristic and height or weight (Künzel, 1989; Dommelen, 1995). The basis for this confidence is fairly clear. There are physical reasons to expect that large bodies will have large resonances and lower pitches, given the constraints on vibrating bodies. So there are ample opportunities for human listeners to associate low sounds with big bodies and vice versa. Within the species *Canis familiaris*, for example, there is a great tendency for large breeds (e.g., German Shepherds) to have low fundamentals, while the smaller breeds (e.g., Chihuahuas) have high ones. However, those examples lead listeners astray in two ways. First, the size of the human body does not seem to be terribly well correlated with the size of the vocal tract, and especially so in the case of the vocal folds themselves. While it is true that heavier folds (as seen, for example, in post-pubescent males) result in a lower F0, the weight of the folds does not seem to correlate with body weight by itself. Nature is full of examples of creatures that have exploited the apparent connection between F0 and size to allow some species (certain frogs and toads, for example) to sound much larger than they are. The fact that this trick works is evidence of its abiding appeal to perceptual systems, even our own, in which we can learn from experience that the association is a weak one. We are fooled by the apparent link, and we continue to think that we can judge body size by voice quality.

We normalize successfully, even when we are misled about body size. This is due to the fact that vocal tract size is what is important, and we are better attuned to that feature than to other, irrelevant features. Vocal tract normalization is so important, in fact, that we must assume its operation for every act of speech perception. Speech perception itself has been found to be sensitive to virtually every acoustic property that covaries with speech production (Studdert-Kennedy, 1976; Repp, 1982; Liberman & Mattingly, 1985; Lisker, 1986). If speech perception makes use of all available information, as it seems to, and normalization is a normal part of speech perception, then we would expect that it too would make use of all information available, even if it is somewhat unusual.

With that thought in mind, we set out to see whether normalization could be affected by a nonspeech sound that nonetheless is a typical component of the speech process, namely, the sound of

inspiration. Since speech is generated primarily by controlling the outgoing breath stream, it follows that inspiration is a necessary precondition to speaking. Sometimes that inspiration will be silent, or nearly so. But quite often it is easily audible. We have already found (Whalen, Hoequist, & Sheffert, 1995) that inclusion of such breath sounds can enhance the memorability of synthetic speech, although the exact mechanism of that improvement is unclear. One possibility is that the inspiration helps to normalize the speech, and thus makes the maintenance of the model of the speaker that much easier. However, we currently have no direct evidence that inspiration noise provides any information for normalization. The present study was designed to produce some of that evidence.

We looked for normalization effects of natural inspiration noises preceding synthetic utterances that formed two vowel continua. We made use of the speaker characteristic that has provided the most consistent normalization effects, namely, speaker sex. Would male and female breaths shift the boundary between the synthetic vowels? Additionally, we tested whether there is coarticulatory information in the inspiration noise. The noise is presumably generated primarily at the larynx and at various points in the lungs themselves, but it propagates through the vocal tract and thus will be shaped by it. If there is anticipation of the upcoming vowel in the vocal tract during the inspiration, then it may provide direct coarticulatory evidence of the vowel to be produced. If so, listeners might be able to use that information and so shift the boundary in the direction of the vowel that the breath was originally produced with.

Experiment

The experiment made use of natural inspiration noises, placed as a precursor to synthetic syllables (similar to Broadbent and Ladefoged, 1960). Two continua were used so that a variety of contexts could be tested.

Method

Stimuli

The inspiration noises were produced by two native speakers of English, one male and one female (the second author). Each read a randomized list containing five repetitions of the words "bead," "bad" and "bud." They were instructed to be sure that they took a breath through the mouth before each utterance. From these, a good exemplar was chosen. They were roughly equivalent in duration and amplitude. Spectrograms of all six are presented in Figure 1. Although the spectral characteristics are quite weak, given the source of the noise, there are noticeable differences among the breaths. The male noises have a lower overall frequency than the female ones. The vowel context did not seem to give rise to different resonances, but there were differences in the relative amplitude of the two main resonances, as can be seen in panels d and e of Figure 1.

Insert Figure 1 about here

We also included a speech precursor, namely, the phrase "And the next word is ____." One token of this phrase from each speaker was used.

The synthetic speech stimuli were created on the serial resonance synthesizer SYN, written by Ignatius Mattingly. There were two continua, one going from /bæd/ to /bəd/and the other going from /bɪd/ to /bɛd/. Formant values were intermediate between those of the male and the female speakers, as

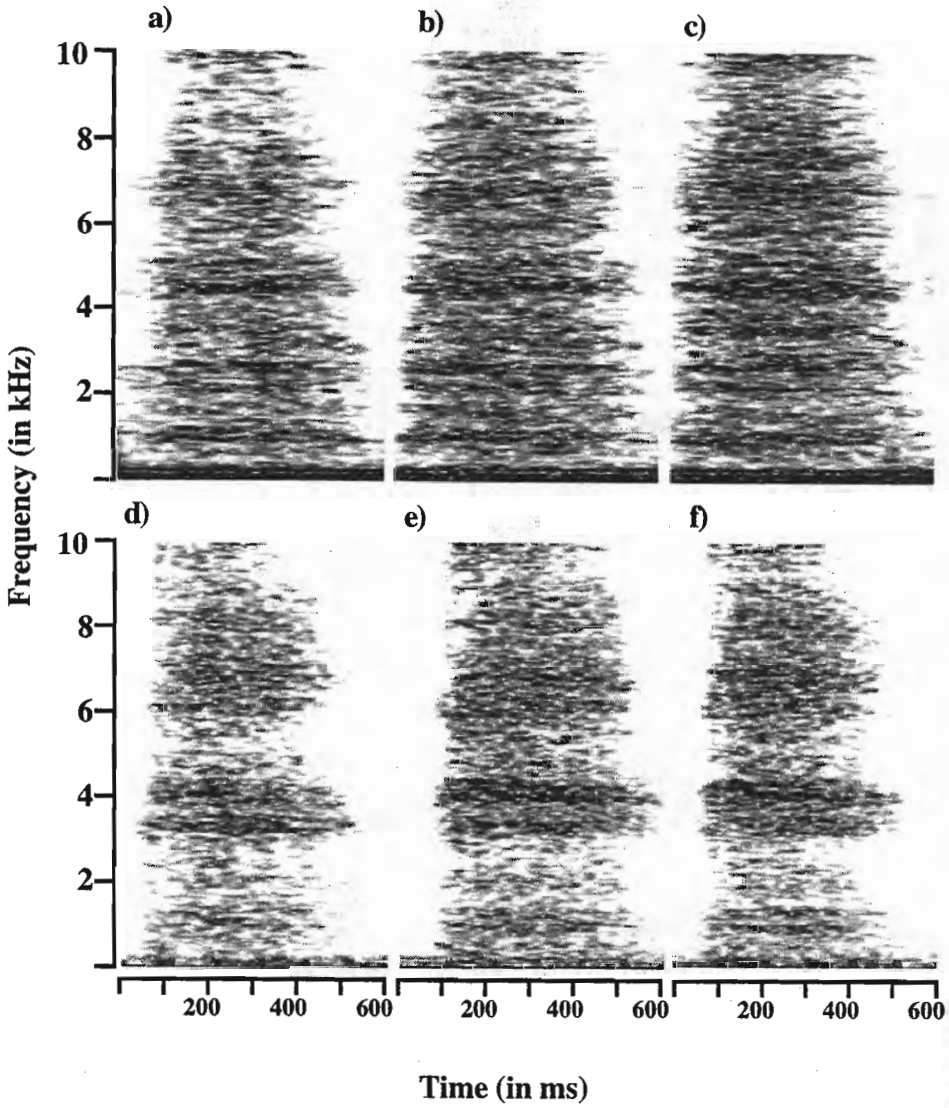


Figure 1. Spectrograms of the six inspiration noises used in the experiment. Panels (a) - (c) are for the male speaker, (d) - (f) for the female speaker. Panels (a) and (d) are from the "bud" utterances, (b) and (e) from the "bad" utterances, and (c) and (f) from the "bead" utterances.

measured from their productions. (We also collected the speakers' productions of "bid" and "bed" for just this purpose.) For the æ/ə continuum, F1 began at 300 Hz and F3 began at 2050 Hz. We found that shifting the onset of F2 along with the steady-state value improved the intelligibility of the token. The onsets are listed in Table 1. There was a short release burst simultaneous with the first 10-20 ms of the syllable, centered at 2000 Hz. The formant center frequencies, listed in Table 1, were attained after 50 ms. There was a steady-state of 250 ms, after which F1 changed linearly to 350 Hz and F2 changed to 1700 Hz. We had better results with a variable F3 ending frequency; those are listed in Table 1. There was a small burst for the final stop simultaneous with the closure. Such "closure bursts" are seldom commented on in the literature, but are readily visible in the speech waveform for many utterances. Inclusion of a very weak closure burst greatly improved the perception of the final stop. F0 was a constant 160 Hz for the first 250 ms, then it fell linearly to 147 Hz. This value is midway between the F0 of the male and that of the female speaker.

Insert Table 1 about here

For the ɪ/ɛ continuum, the formants began at 250, 1200 and 2275 Hz for F1, F2 and F3 respectively. F3 was the same for all five stimuli, but it did rise from 2500 Hz to 2550 Hz during the "steady-state" portion of the syllable (50-250 ms). There was a release burst simultaneous with the first part of the transitions, centered at 2000 Hz and lasting 10-20 ms. The steady-state values for F1 and F2 are given in Table 2. For the final stop, the transitions lasted 50 ms and ended at 175, 1850 and 2400 Hz for F1, F2 and F3 respectively. There was a small "closure" burst at the end of the syllable. As before, F0 was a constant 160 Hz for the first 250 ms, then it fell linearly to 147 Hz.

Since these formant and F0 values were based on the average of the actual formant and F0 values used by our two speakers, the sex of the synthetic "speaker" was maximally ambiguous, which would presumably allow the speaker information in the inspiration noise to have an effect. It would not be surprising if relative weak information, such as that presumably contained in the breath sound, would be unable to overcome more robust cues in the speech itself.

Insert Table 2 here

Procedure

Each of the five members of each continuum were combined with each of the precursors. For the æ/ə continuum, the breaths produced before "bad" and "bud" were used. For the ɪ/ɛ continuum, the breaths produced before "bead" and "bad" were used. The first case is therefore one in which the vowels at the endpoints of the continua were identical to the ones produced with the inspiration noise. The second case is one in which the vowels produced with the breath were more peripheral in the vowel space than the vowels of the test items. In this way, we might perhaps see a coarticulatory effect in the exaggerated case even if we did not find one in the case in which the exact items were used. The full speech precursors (male and female versions of "And the next word is _____") were the same for both continua.

The continua were recorded on DAT and presented to one subject at a time over headphones. Ten repetitions of each combination of synthetic syllable and precursor were randomized and presented for identification as "bad" or "bud" or, for the other continuum, "bid" or "bed." There was a 200 ms pause between the precursor (speech or breath) and the syllable, and a 2.5 s pause in which the answer was

Table 1

Formant values for the æ/ə continuum.

		F1	F2		F3	
Vowel	Cont. #		onset	steady-state	steady-state	offset
ə	1	640	850	1450	2850	2900
	2	660	900	1530	2830	2825
	3	685	975	1610	2810	2750
	4	705	1015	1740	2790	2675
æ	5	725	1100	1875	2775	2600

Table 2

Formant values for the steady-states of F1 and F2, ɪ/ɛ continuum.

vowel	cont. #	F1	F2
ɪ	1	470	2280
	2	490	2260
	3	510	2240
	4	530	2220
ɛ	5	550	2200

written. There was a 5 s pause after every ten items, corresponding to the end of a line on the answer sheet. The order of presentation of the two continua was randomly selected for each listener.

Subjects

The subjects were 15 undergraduate students at the University of Connecticut, who received course credit for their participation, along with seven colleagues from Haskins Laboratories and the two authors. All were native speakers of English with no reported hearing problems.

Results

Our expectations for the directions of the shifts in identification were as follows. For the coarticulatory information in the breaths, we predicted that the identification would shift toward the vowel that the breath had been produced with. This was the same vowel of the response in the æ/ə continuum, and the vowel that was nearest the endpoint vowel for the ɪ/ɛ continuum (that is, /i/ for /ɪ/, and /æ/ for /ɛ/). Our expectations for the effects of speaker sex on the responses were based on the changes in the significance of F1 for the two speakers. With the smaller vocal tract, the female speaker should have higher F1s than the male. Thus this should lead to more /ə/ responses for a female breath or precursor in the æ/ə continuum and more /ɪ/ responses for a female breath or precursor in the ɪ/ɛ continuum.

The mean percentage of "bud" responses (across the whole continuum) to the æ/ə continuum are shown in Table 3. The mean percentage of "bid" responses (across the whole continuum) to the ɪ/ɛ continuum are shown in Table 4. The effects were assessed with a series of ANOVAs, using the percentage of responses across the whole continuum. (With a small number of continuum steps and a relatively small number of judgments per step, our experience has been that PROBIT analysis overemphasizes random variation at the endpoints of the continuum.) For each continuum, three analyses were performed. The first, main analysis included both the speech precursors and the breath precursors, with factors Sex of Speaker (two levels, male and female) and Precursor (three levels, speech and two phonetic contexts of breath). Because the breaths include two effects (speaker normalization and coarticulation), we can expect that this factor will need further elaboration without regard to the speech precursor. Thus, the second analysis examined only the breath precursors, with the factors Sex and Precursor (two levels each). The third analysis was a one-way examination of Sex for the speech precursors.

Insert Tables 3 and 4 about here

For the æ/ə continuum, Sex was not significant in the main analysis ($F(1,23) < 1$, n.s.). The Precursor factor just missed significance ($F(2,46) = 3.11$, $p = .054$), and the interaction was not significant ($F(2,46) = 2.07$, n.s.). In the analysis of just the breaths, Sex was again not significant ($F(1,23) = 2.38$, n.s.). The Precursor factor was significant ($F(1,23) = 6.30$, $p < .05$), while the interaction was not ($F(1,23) < .1$, n.s.). As can be seen from Table 3, the /ə/ breaths elicited a significant 2.2% more "bud" responses, indicating that the coarticulatory information was effective in shifting the perception of these ambiguous synthetic vowels. In the analysis of just the speech precursors, Sex was not significant $F(1,23) < .1$, n.s.). Thus the speaker characteristic of sex did not influence the perception of these stimuli.

For the ɪ/ɛ continuum, Sex was not significant in the main analysis ($F(1,23) = 1.49$, n.s.). The Precursor factor was significant ($F(2,46) = 9.96$, $p < .001$), perhaps due to the overall difference between the speech and breath precursors. The interaction just missed significance ($F(2,46) = 3.05$, $p = .057$). In the analysis of the breaths alone, Sex was significant ($F(1,23) = 4.29$, $p < .05$). The Precursor factor was not significant ($F(1,23) < 1$, n.s.), while the interaction was marginal ($F(1,23) = 3.00$, $p < .10$). As can be seen

Table 3

Percent "bud" responses to the æ/ə continuum. The prediction was that the female speaker and the /ə/ breath would elicit more "bud" responses.

	speech	/æ/ breath	/ə/ breath	marginal	without speech
male	51.4	53.9	55.4	53.6	54.7
female	53.0	51.2	54.1	52.8	52.7
marginal		52.6	54.8		

Table 4

Percent "bid" responses to the ɪ/ɛ continuum. The prediction was that the female speaker and the /i/ breath would elicit more "bid" responses.

	speech	/æ/ breath	/i/ breath	marginal	without speech
male	52.4	44.6	43.3	46.8	44.0
female	51.5	45.1	48.0	48.2	46.6
marginal		44.9	45.7		

from Table 4, the female breaths elicited 2.6% more "bid" responses than the male breaths, an effect in the direction predicted by the likely effect of normalizing for the speaker's vocal tract. Although the /i/ breaths elicited 0.8% more "bid" responses, as we would expect on coarticulatory grounds, this difference was not significant. In the analysis of just the speech precursors, Sex was not significant $F(1,23) = 1.95$, n.s.), despite a 2.6% difference in the predicted direction. Thus, the speaker characteristic of sex did not influence the perception of these stimuli when the precursor was speech, but it did do so when the precursor was an inspiration noise.

After the completion of the experiment, it seemed of interest to determine how much of the information in the breath sounds was consciously available to listeners. To test this, we ran a small study with eight staff members of the Speech Research Laboratory at Indiana University. They had not participated in the main experiment. These subjects listened to 10 repetitions of each of the six breath sounds to see whether they could identify the vowel it had preceded. We blocked the breaths into the same pairs that had been used in the perception test, so that subjects judged either /i/ or /æ/, or /ə/ or /æ/. Subjects were 49.5% correct on the first comparison, and 47.5% correct on the other. Neither of these figures differs from the chance level of 50%. Four of the subjects also made judgments of speaker sex. These judgments were 98.3% correct for both sets of breaths.

Discussion

Although the sound produced by inspiration is nonlinguistic, it is nevertheless shaped by the vocal tract. Thus such breath sounds have the potential to provide information about the dimensions and configuration of that vocal tract. In the present experiment, we have found evidence that such information is used in the perception of synthetic speech, leading to the expectation that this information is available in more natural situations. Whether there will be instances in which the contribution of such sounds is apparent is a different question. But these results raise the possibility that they might be used in automatic speech recognition as a converging source of information about the speaker's characteristics. Additional research will be needed in order to determine the conditions that result in the contribution of such sounds.

The degree of conscious awareness of the information seemed to play no role in the effectiveness of that information. Speaker sex was readily perceived from the inspiration sounds, while vowel quality was not. Yet speaker sex influenced the judgments in one continuum, while coarticulatory vowel information affected the other. This indicates that the effect of the breath sounds is unlikely to be a strategic one, as might arguably be the case in the example of the specification of speaker sex by vision alone (Strand, 1995).

Our findings suggest that the coarticulatory information in the breaths may need to match exactly the vowels that are to be perceived. In our data, the coarticulatory influence was present in the continuum in which the breaths were produced with just the vowels that constitute the endpoints of the continuum. In the other continuum, we used more peripheral vowel articulations, in the hopes that this might exaggerate the influence that the coarticulatory information would have. For that case, however, there was no significant influence of that information. Perhaps exactly appropriate information is, in fact, more easily used in this context. We are currently conducting an experiment to test this further.

The natural speech precursors did not affect the perception of the continua, in contrast to results found with synthetic precursor sentences (Broadbent & Ladefoged, 1960; Remez, et al., 1987). We suspect that this lack of an effect is due to the discrepancy between the natural speech and the synthesis. The previous studies used synthesis for both the precursor and the target. This may make it easier for the listeners to integrate the two into a single, coherent percept. The clear difference between the natural speech and the synthesis in the present experiment, on the other hand, may have led to a lack of integration and thus to a lack of an effect. Various manipulations of the stimuli, such as synthesizing the precursor or creating a more natural sounding synthetic continuum, would be needed to fully resolve this issue.

The fact that speaker information extrinsic to the signal, as shown here and in Strand (1995), can influence the perception of speech clearly indicates that normalization procedures should not be exclusively intrinsically based. There is enough evidence now to indicate that the listener may construct a "model" of the speaker as a means of perceiving speech, given the large number of differences in vocal tract size that any listener will confront in a lifetime. Such "model making" leads to the prediction that individual characteristics outside the speech realm (such as a personal name, a habitual gait, or even a well-known belonging) could influence perception of speech in a way appropriate to a known individual. As an alternative to making a model of an individual speaker, extrinsic effects might result from all the information available about a vocal tract as selected from among a range of vocal tract types that are not associated with individuals. In that case, only speech information would be expected to play a role. The present study shows, at least, that the sound of inspiration must be included in the range of speech-relevant information available to the listener.

References

Ainsworth, W.A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 103-111). New York: Academic Press.

Broadbent, D.E., & Ladefoged, P. (1960). Vowel judgments and adaptation level. *Proceedings of the Royal Society*, **151**, 384-399.

Chiba, T., & Kajiyama, M. (1941). *The vowel: Its nature and structure*. Tokyo: Tokyo-Kaiseikan Publishing Co., Ltd.

Davis, P. (1994). *Bondage*. New York: Pocket Books.

Dommelen, W.A. (1995). Speaker and listener sex for speaker height and weight identification. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (pp. 738-741). Stockholm: Stockholm University.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, **AU-16**, 73-77.

Gerstman, L.J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, **AU-16**, 78-80.

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**(1), 152-162.

Kitamura, T., & Akagi, M. (1994). Speaker individualities in speech spectral envelopes. In *International Conference on Spoken Language Processing*, 3 (pp. 1183-1186). Yokohama: Acoustical Society of Japan.

Künzel, H.J. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica*, **46**, 117-125.

Lass, N.J., Phillips, J.K., & Bruchey, C.A. (1980). The effect of filtered speech on speaker height and weight identification. *Journal of Phonetics*, **8**, 91-100.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech*, **29**, 3-11.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 676-684.

May, J.G. (1976). Vocal tract normalization for /ʃ/ and /s/. *Haskins Laboratories Status Report on Speech Research*, **SR48**, 67-74.

Miller, J.D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114-2134.

Nearey, T.M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088-2113.

Potter, R.K., & Steinberg, J.C. (1950). Towards the specification of speech. *Journal of the Acoustical Society of America*, **22**, 807-820.

Rand, T.C. (1971). Vocal tract size normalization in the perception of stop consonants. *Haskins Laboratories Status Report on Speech Research*, **SR25/26**, 141-146.

Remez, R.E., Rubin, P.E., Nygaard, L.C., & Howell, W.A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance*, **13**, 40-61.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**, 81-110.

Strand, E.A. (1995). The role of visual information in speaker normalization of fricatives. *Journal of the Acoustical Society of America*, **97**(5), 3285.

Studdert-Kennedy, M. (1976). Speech perception. In N.J. Lass (Eds.), *Contemporary issues in experimental phonetics* (pp. 243-293). New York: Academic Press.

Syrdal, A.K., & Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086-1100.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, J.R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, **60**, 198-212.

Whalen, D.H., Hoequist, C.E., & Sheffert, S.M. (1995). The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America*, **97**, 3147-3153.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Encoding of Visual Speaker Attributes and Recognition Memory for Spoken Words[1]

Helena M. Saldaña,[2] Lynne C. Nygaard,[3] and David B. Pisoni

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

[1]This work supported in part by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University–Bloomington.
[2]Now at House Ear Institute, Los Angeles, California.
[3]Now at Department of Psychology, Emory University, Atlanta, Georgia.

Encoding of Visual Attributes and Recognition Memory for Spoken Words

Abstract. An experiment was designed to assess the extent to which visual articulatory information is encoded in memory. In a recent study, Palmeri, Goldinger, and Pisoni (1993) reported that listeners were more accurate at recognizing previously presented words when they were presented in the same voice as at test than when they were presented in a different voice. This result suggests that detailed voice information is not stripped away by a normalization process during the early stages of spoken word recognition; instead information about the talker's voice is encoded into long-term memory and may later facilitate recognition of spoken words. The present investigation was designed to determine whether detailed cross-modal linguistic information is also retained in long-term memory. Subjects were presented with two audio-visual speakers producing lists of isolated words. The words were presented at three signal-to-noise ratios ranging from +5 dB to –5 dB. In the low signal-to-noise conditions, listeners are forced to attend more carefully to the visual information presented in order to extract the linguistic content. It is proposed that this manipulation will cause dynamic visual information to be encoded along with the spoken words in memory. If it is the case that detailed visual information about the talkers articulation is retained in long term memory, then we would expect that listeners will be better at recognizing "old" words when they are presented with the same visual speaker at test. No significant effect of visual articulation was found on recognition memory, however a consistent trend was observed. The present results have implications for current theories of spoken word recognition and the nature of the representation of words stored in the mental lexicon.

Introduction

Speaker Variability

The speech signal is highly variable across individual talkers. This is largely the result of differences in the shape and length of the vocal tract (Carrell, 1984; Fant 1973; Summerfield & Haggard 1973), glottal source function (Carrell, 1984), individual differences in articulation (Ladefoged, 1980), and dialect. Traditional theories of speech perception have often treated such variability as noise in the signal that must be filtered out by the listener (e.g., Blandon, Henton, & Pickering, 1984; Disner, 1980; Green, Kuhl, Meltzoff & Stevens, 1991). Many of these theories assume some kind of talker-normalization process, in which talker specific information is stripped away from the signal and the remaining phonetic cues are matched to idealized representations in memory. This view would predict that only an abstract symbolic linguistic code, and not its carrier are retained in long term memory.

Recent evidence suggests that detailed information about the speaker's voice is encoded into long-term memory. Craik & Kirsner (1974) utilized a continuous recognition paradigm and demonstrated that recognition memory was facilitated when the test item was presented in the same voice as the study item. Palmeri, Goldinger, & Pisoni (1993) extended this finding by presenting listeners with twenty different voices and increasing the lag between study and test items. They replicated Craik & Kirsner's finding by

showing a clear voice effect even when there were 64 intervening items between study and test. When a word was repeated in the same voice, subjects showed higher levels of recognition performance than when the same word was presented in a different voice.

In a similar study using visual stimuli, Kirsner (1973) found that recognition memory was better when test items were presented in the same typeface as study items. There is also evidence in the literature showing that surface information is retained when subjects read passages of text. Kolers & Ostry (1974) presented subjects with inverted passages of text and had them practice reading them. In a subsequent session, they had subjects read text in the same inverted form or in a different inverted form. They found that reading times were best when subjects were given the same inverted form at test. In fact, savings was found even one year after the original presentation. All of these findings indicate that surface features of auditory and visual stimuli are not discarded during the perceptual process, but rather are encoded in the memory trace and retained in long term memory.

Audio-Visual Speaker Variability

In a recent study, Sheffert & Fowler (1995) designed an experiment to determine whether visual information about the speaker was encoded along with the phonetic representation of the word. The question that they explored was the following: When listeners are able to view the speaker talking do they also retain in memory physical details about the speakers face along with the lexical representation? The authors utilized a continuous recognition procedure like the one used by Palmeri et al. with audio-visual tokens. They wanted to show that observers would be better at recognizing words when the same visual speaker was presented at study and test. In a series of four experiments, the authors consistently replicated the voice effects found by Palmeri et al., but they failed to show an implicit effect of face information on word recognition. The authors concluded that voice information shared a "privileged" status in relation to phonological information in memory for spoken words.

In this study, Sheffert & Fowler assumed that any information about the speaker should be encoded along with the phonetic representation. For instance, in one condition the authors presented listeners with the same speaker at test but changed an item of clothing on the speaker (i.e., hat or scarf). In all previous studies which demonstrated the encoding of surface features with the linguistic code, the surface features shared an integral relation with the spoken message. In other words, some processing of the surface features had to be performed to retrieve the linguistic message. It is clear that listeners under normal listening conditions do not require visual speaker information to retrieve linguistic content. Therefore, the lack of an implicit face effect in the Sheffert & Fowler study might be due to the fact that it was not necessary to process the face information in a mandatory way to perceive the word.

Current Investigation

In the present investigation, we manipulated the signal-to-noise ratio in a continuous recognition task, in an attempt to force listeners to process the visual articulatory information in a mandatory fashion. If the encoding of surface features is due to an integral relationship between surface features and the linguistic code, it is possible that this manipulation will cause visual surface features to be encoded in memory along with the lexical items. For half of the repeated items the visual speaker was presented in a dynamic display. For the other half of the repeated items, the visual speaker was presented in the form of a static image. Therefore, face information is available to the subject but no dynamic articulatory information is available for the subject to process. We propose that listeners will only show a benefit in recognition when the dynamic information from the same talker is available at both study and test.

Method

Subjects. One-hundred-nineteen subjects were paid $5.00 an hour for participating in the experiment. All subjects were native speakers of English with normal hearing and normal or corrected vision.

Stimulus Materials. The stimuli were a list of isolated words spoken by a male and a female talker. Two different talkers were videotaped uttering three hundred monosyllabic words. Talkers were recorded in a sound attenuated studio. Each word was presented to the talker on a CRT screen and the talker was instructed to say each word while looking into the camera. The stimuli were digitized on a Macintosh Quadra 950. The sound was sampled at 20.5 kHz with 16 bit resolution. The video image was captured at 30 frames per second. The length of each word varied from 2 to 3 seconds. Some items were randomly selected to serve as audio-alone trials. These trials did include an image of the actors face, however, the actor was not articulating the word. A static image was constructed by placing the same frame in consecutive positions for the duration of the originally articulated utterance.

Noise-Embedded Conditions. A General Radio 1381 random-noise generator was attached to the Macintosh Quadra 950. The noise generator was controlled by a voice key. Any time a sound was detected from the sound board at the initiation of a word, white noise was output. When the sound ceased, the white noise was deactivated. Therefore, the noise was only present during the auditory presentation of each word.

The test list was constructed from a subset of two hundred words from each talker (the same words for each talker). Each word was presented and repeated once. The repetition of any given word occurred after a lag of 2, 8, 16, or 32 intervening items. Each lag value was used an equal number of times in each list. One hundred forty-four pairs of presented and repeated items were used in the list. All original items were articulated by each speaker, however, half of the repeated items consisted of a static image of the talker's face in place of the visually articulated utterance (72 items). Each subject was given an initial practice list of 15 words to become familiar with the task. None of these words were repeated in the experiment. The next 30 items were presented to establish a memory load and were not considered in the analysis. Twenty-one filler items were randomly distributed throughout the test portion of the list. All of the filler items consisted of a static visual presentation. The total number of words in the list equaled 354 items.

Each talker was presented an equal number of times on the list. On the initial presentation of the word, one of the talkers was randomly selected. The probability of the same talker or different talker repetitions was equal.

Design. The lag between the initial presentation and the repeated word was a within subjects variable (2,8,16,32), as was the talker who produced the repetition (same talker vs. different talker) and type of repetition (static vs. dynamic).

Procedure. Subjects were tested in groups of 3 or fewer. Stimulus presentation was controlled by a Macintosh Quadra 950. After hearing and seeing each word, the subject was allowed a maximum of four seconds to respond on a numbered response sheet. The subject's task was to simply circle the word OLD if they had heard the item previously or NEW if they had not. Subjects were told it was very important to watch as well as listen to each presentation. An experimenter was seated in the room to make sure subjects looked directly back at the monitor after each response.

Groups of subjects were randomly selected to serve in one of four conditions: The following number of subjects were assigned to each condition: Clear, N=29; SNR +5, N= 27 ; SNR 0, N=34; and SNR -5, N=29.

Results

All the data to be presented here are expressed in terms of hit rates (the number of items identified as "old" given that they were previously presented). We first examine the performance of the listeners in the clear. There was a significant effect of talker. Subjects were more likely to correctly identify an item as "old" if it was presented by the same speaker $F(1, 84) = 35.727$, $p<.001$. This was true regardless of whether the visual speaker was articulating the word or not. There was also a significant effect of lag ($3, 84) = 30.257$, $p<.001$. This result indicated that recognition memory performance decreased as the lag increased.

Insert Figure 1 about here

An overall 3 x 2 x 2 x 4 ANOVA was conducted on the between subjects factor of signal-to-noise ratio (+5, 0, -5), visual-display (articulated, non-articulated), talker (same talker, different talker) and lag (2, 8, 16, 32). The main effect of signal-to-noise ratio was highly significant $F(2, 87) =9.433$, $p<.001$, demonstrating that signal-to-noise ratio had an overall effect on the recognition of items. The effect of talkers was also highly significant, $F(1,87) =360.53$, $p<.001$. Subjects were better at recognizing an item as "old" when it was presented by the same talker used in the original presentation. However, the interaction of articulation x talker was not significant. This finding indicates that dynamic articulation did not differentially help recognition for same versus different talker.

Separate ANOVAs were conducted on each signal-to-noise ratio. In all of the conditions, we observed significant main effects for talker and stimulus lag. There was no main effect of articulation, again, suggesting that seeing a dynamic visual image of the talker's face during the repetition did not affect recognition performance.

Insert Figure 2 about here

Discussion

The present findings replicate the implicit voice effect reported by Palmeri et al (1993). Subjects showed a clear advantage on item recognition when the repeated item was presented in the same voice as the original item. This implicit voice effect was not eliminated when words were embedded in noise.

We also observed an overall decrease in hit rate as SNR decreased. However, we did not observe a difference in performance between dynamic and static stimuli. It is well known that the presence of dynamic visual information about the talker's face increases intelligibility of degraded speech. Based on these results, we might have expected that subjects would do better when presented with a dynamic face regardless of voice (simply due to better intelligibility), however, this prediction was not observed. It is

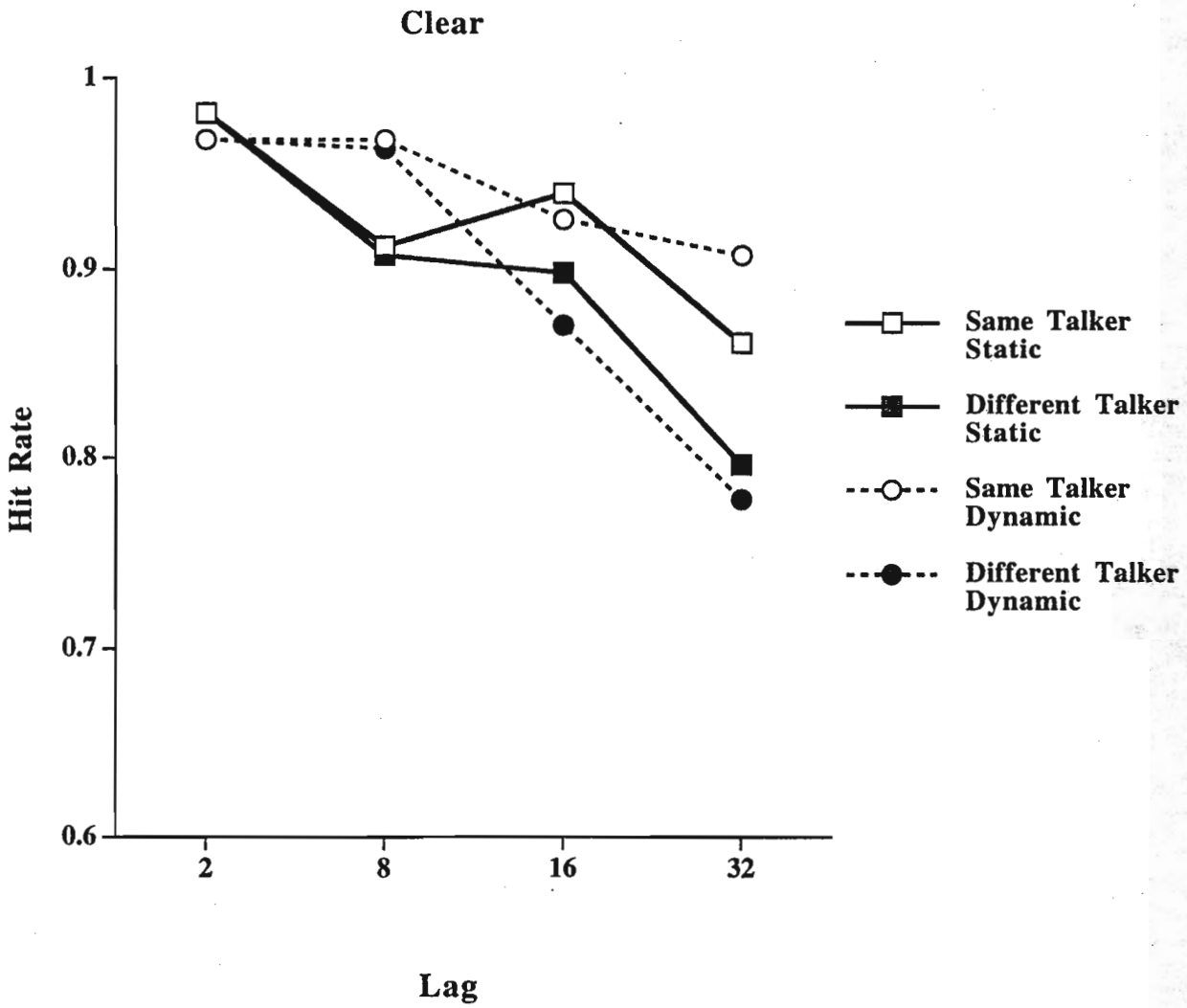


Figure 1: Hit rates for words in the clear speech condition

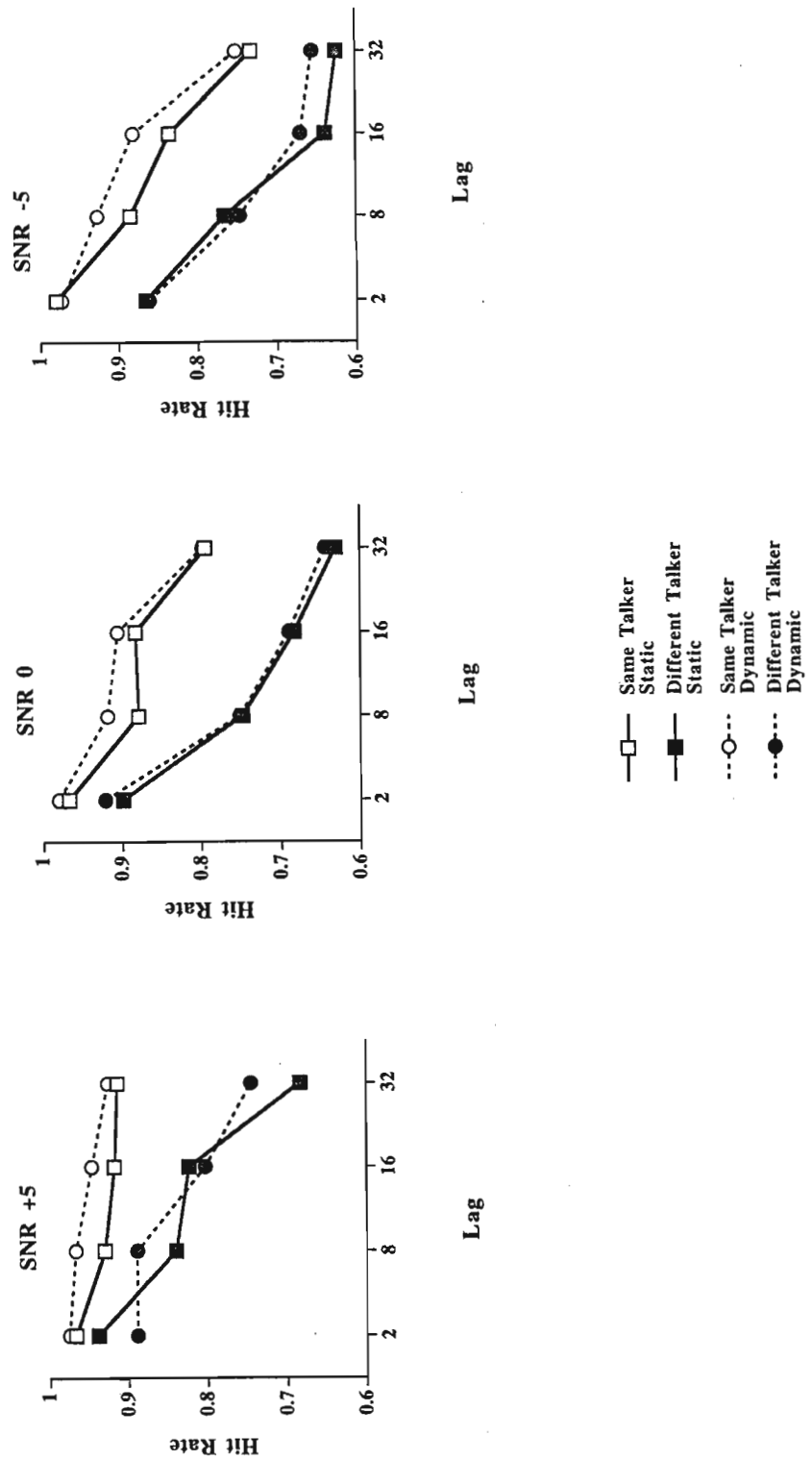


Figure 2: Hit rates for words in each of the three signal-to-noise conditions

possible that the noise levels were not high enough to force listeners to process the visual information in a mandatory way. It is also possible that the words were too distinctive therefore preventing confusion at the lower SNR ratios.

General Discussion

This experiment was designed to investigate whether visual surface features of spoken words are encoded in memory along with a symbolic linguistic code. We hypothesized that forcing listeners to use visual information to extract the linguistic message would cause them to encode the complementary visual information in a mandatory fashion. This hypothesis was not borne out. However, an interesting pattern emerged from the data. Across all signal-to-noise ratios, the function for the same talker-articulated (dynamic) visual stimuli was higher than the same talker-non articulated (static) visual stimuli. This observation suggests that visual articulatory variability might be encoded in long term memory. Recall that in the same talker condition the voice remained the same regardless of the visual display, the only difference between the same talker stimuli was the dynamic visual information. It is possible that the highly robust voice effect washed out the effects of dynamic visual information. In subsequent studies we plan to use a cross-splicing technique in order to completely disassociate the auditory and visual information and force observers to rely on only one source of information or the other in encoding the stimulus array.

Studies are also planned which compare the results from our normal listeners with performance obtained with hearing-impaired and good and poor speechreaders. It is possible that the absence of a robust effect in the present study is due to the fact that our normal-hearing listeners routinely rely on auditory information for speech communication. If so, we might observe a visual effect when we use a group of observers who rely heavily on visual information for speech communication.

In summary, this study was designed to investigate the nature of the representations encoded during spoken word recognition. It was demonstrated that voice information is encoded during word recognition and that dynamic visual talker information is also encoded. However, further research is required to determine to what extent the visual information is useful in subsequent recognition memory.

References

Blandon, R.A.W., Henton, C.G., Pickering, J.B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication*, **4**, 59-69.

Carrell, T.D. (1984). Contributions of fundamental frequency formant spacing and glottal waveform to talker identification. *Research on Speech Perception Technical Report No. 5*. Bloomington: Indiana University, Department of Psychology.

Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274-284.

Disner, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**, 253-261.

Fant, G. (1973) *Speech Sounds and Features*. Cambridge, MA: MIT Press

Green, K.P., Kuhl, P.K., Meltzoff, A.N., & Stevens, E.B. (1991). Integrating speech information across talkers, gender, and sensory modalities: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524-536.

Kirsner, K. (1973). An analysis of the visual component in recognition memory for verbal stimuli. *Memory and Cognition*, 1, 449-453.

Kolers, P.A. & Ostry, D.J. (1974). Time course of loss of verbal information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior*, **13**, 599-612.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language,* **56**, 485-502.

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 & Cognition*, **19**, 309-328.

Sheffert, S. & Fowler, C.A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Learning & Memory*, **34**, 665-685/

Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report in Progress in Speech Perception, Vol. 2.* (pp. 1-12). Belfast, UK: The Queen's University of Belfast, Department of Psychology.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Audio-Visual Speech Perception Without Speech Cues[1]

Helena M. Saldaña,[2] David B. Pisoni, Jennifer M. Fellowes,[3] and Robert E. Remez[3]

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

[1] This research was supported by NIH-NIDCD Research Grant DC-00111 to Indiana University, Bloomington, IN.
[2] Now at House Ear Institute, Los Angeles, CA.
[3] Department of Psychology, Barnard College, New York, NY.

Audio-Visual Speech Perception Without Speech Cues

Abstract. A series of experiments was conducted in which listeners were presented with audio-visual sentences in a transcription task. The visual components of the stimuli consisted of a male talker's face. The acoustic components consisted of: (1) natural speech, (2) envelope-shaped noise which preserved the duration and amplitude of the original speech waveform, and (3) various types of sinewave speech signals that followed the formant frequencies of a natural utterance. Sinewave speech is a skeletonized version of a natural utterance which contains frequency and amplitude variation of the formants, but lacks any fine-grained acoustic structure of speech. Intelligibility of the present set of sinewave sentences was relatively low in contrast to previous findings (Remez, Rubin, Pisoni, & Carrell, 1981). However, intelligibility was greatly increased when visual information from a talker's face was presented along with the auditory stimuli. Further experiments demonstrated that the intelligibility of single tones increased differentially, depending on which formant analog was presented. It was predicted that the increase in intelligibility for the sinewave speech with an added video display would be greater than the gain observed with envelope-shaped noise. This prediction is based on the assumption that the information-bearing phonetic properties of spoken utterances are preserved in the audio + visual sinewave conditions. This prediction was borne out for the tonal analog of the second formant (T2), but not the tonal analog of the first formant (T1) or third formant (T3), suggesting that the information contained in the T2 analog is relevant for audio-visual integration.

Sinewave Speech

Previous research has demonstrated that listeners can perceive linguistic content in non-speech stimuli consisting of three time-varying sinusoidal signals (Remez, Rubin, Pisoni, & Carrell, 1981). The sinusoids in these stimuli tracked the center frequencies of the first three formants of a naturally produced sentence. Although the quality of these signals was very unnatural, listeners were able to correctly transcribe the sentence with a high degree of accuracy (Remez et al., 1981). This seminal study shifted the emphasis of theories of speech perception from the momentary spectral attributes that were believed to underlie phonetic perception (the so-called "speech cues"), to the time-varying or spectro-temporal attributes of speech and its perceptual organization. The sinewave replicas used in this study retained spectral variation in the absence of short-term spectra typical of vocal sound production. Remez et al. argued that listeners were relying on the global time-varying properties of these non-speech patterns to make fine-grained phonetic distinctions.

In the original study, Remez et al. presented listeners with every permutation of the sinewave stimuli (tones 1, 2, and 3 individually and in combinations), and demonstrated that individual tones were unintelligible (at less than 5%). Remez et al. also observed high intelligibility for the combination of tones 1 and 2, but significantly less intelligibility for the combination tones 1 and 3 and combination tones 2 and 3. They argued that information retained in the sinewave replicas specifies the talkers vocal tract transfer function and how these attributes change over an utterance. The differential findings for the various conditions demonstrates that some portions of the speech signal are more informative than others about conveying the dynamic operations of the vocal tract.

In the years following the publication of the report by Remez et al., numerous studies have been carried out replicating the basic finding that speech perception can take place when traditional speech cues are eliminated from the signal (see Remez, Rubin, Berns, Pardo & Lang, 1994; Remez & Rubin, 1984). The technique of sinusoidal speech synthesis has proven to be extremely useful in generating non-speech patterns that preserve important phonetic properties of speech but do not produce speech-like qualities. The fact that listeners are able to perceive the phonetic content of the original utterance from these highly impoverished signals provides an "existence proof" that important phonetic information is available in these dynamic nonspeech patterns that are following the changes in the talker's vocal tract. Recent studies have further demonstrated that these patterns not only retain phonetic information but also reliably preserve detailed information about the identity of the specific talker (Remez, Fellowes, & Rubin, in press).

Multimodal Integration

It is well understood that the primary modality for spoken communication is audition. However, in noisy environments, listeners utilize information from other sensory systems to aid in the recognition and comprehension of speech (Sumby & Pollack, 1954). This finding, in and of itself, is interesting because it suggests to us that the perceptual system takes in all relevant information for a given event and combines it to form one unified percept. As a result of experimental evidence for multimodal integration in speech (McGurk & MacDonald, 1976), numerous theories have been proposed for the process of multi-modal integration (see Summerfield, 1981, for a review). However, a close look at these theories reveals that little attention has been paid to defining the type of information that is integrated by the system (Massaro, 1987; MacDonald & McGurk, 1978; however see Grant, Braida, & Renn, 1994).

In the present study, we utilized sinewave synthesis techniques to ask several questions about the nature of the information involved in multimodal integration speech. We believe that the performance of our subjects provides important new clues about the time-varying attributes of auditory signals that are important in perceiving multimodal displays of speech. Empirically, the project was a straightforward attempt to determine the relative effectiveness of several different kinds of acoustic speech signals when combined with a visual display of an articulating face. The video display in this study was conventional: a live subject was videotaped producing a list of English sentences. Although we presumed that some of the morphological and dynamic attributes of a talking face provide information about the linguistic attributes underlying the articulation, we did not manipulate this source of information in this project; we attempted here simply to control those factors in order to focus exclusively on a related auditory question: What kind of auditory attributes permit the perceiver to make use of the visual attributes? In other words: Does the perception of speech (multi- or unimodally) require specific auditory qualities?

The present investigation examined single-tone sinewave replicas (T1, T2, and T3), the pair-wise combination tones 1, 2, and 3, bit-flipped noise, and natural utterances. Previous research has shown intelligibility of single-tone sinewave sentences in an auditory-alone condition is below 5% (Remez et al., 1981). We expected that visual information would integrate with these single sinewave replicas leading to better intelligibility than video or audio-alone conditions. We also expected to observe different levels of intelligibility depending on which tone was combined with the visual display. This prediction was derived from previous findings which demonstrated different levels of intelligibility for different combinations of tones. If it is the case that some portions of the time-varying signal are better at specifying the dynamic attributes of the vocal tract over an utterance, we would expect to find different levels of integration with a visual display of an articulating face. We also proposed that the increase in intelligibility for the sinewave signals would be higher than any increase found for the bit-flipped noise conditions. This latter hypothesis is based on the assumption that time-varying spectral properties of the acoustic signal are necessary for

audio-visual integration, and that these time-varying properties are obliterated in the bit-flipped noise signals, which only preserve the amplitude and duration of the speech envelope.

Cross-Modal Integration With Sinewave Speech

Method

Participants

Two-hundred and ninety-six normal-hearing listeners, with no prior experience in speechreading served as participants in the experiment. They were each paid $5.00 for their participation. All had normal or corrected vision and reported no history of a speech of hearing impairment at the time of testing.

Stimulus Materials

The speech materials consisted of the following ten sentences :

1. The swan dive was far short of perfect.
2. Where were you a year ago?
3. My dog bingo ran around the wall.
4. A large size in stockings is hard to sell.
5. Kick the ball straight and follow through.
6. The beauty of the view stunned the young boy.
7. Cut the meat into small chunks.
8. Rice is often served in round bowls.
9. The boy was there when the sun rose.
10. My TV has a twelve inch screen.

The sentences were recorded by a male talker who was instructed to speak in a conversational style. The video image consisted of the talker's head and part of his neck. The talker wore a black turtleneck and was recorded against a black background. The taped sentences were then digitized on a Macintosh Quadra 950 at a rate of 30 frames per second. The audio track was sampled at 22 kHz using 16-bit resolution.

Sinewave Synthesis

The auditory portions of the video tape were first low-pass filtered at 4.5 kHz and then sampled at 10 kHz with 12-bit amplitude resolution and stored on a VAX-based computer system. The method of linear predictive coding (LPC) was used to estimate spectra at 5 ms intervals (Markel & Gray, 1976). The output was hand-checked for erroneous values and corrected when necessary. The formant estimates were then used to drive the output of a formant synthesizer (Rubin, 1980), which calculates the waveforms of signals generated by adding multiple independent audio-frequency oscillators (see Remez et al., 1994).

Bit-Flipped Noise

The digital files of the audio tracks of the sentences were also subjected to a random bit-flipping algorithm. The algorithm randomly flipped the sign-bit of 50% of the digital samples in each auditory file.

This manipulation resulted in a signal that retained the amplitude envelope and duration of the original waveform but eliminated the fine-grain structure of the speech signal.

Audio-Visual Sentences

The audio files were then combined with the corresponding video files using a digitally-controlled video editing package. The synchronization was checked by visually comparing the sinewave audio track with the natural audio track.

A final presentation tape was prepared for each condition. Each sentence was presented five times with a 10-second ISI. Following the fifth presentation of the sentence, a prompt on the screen read "Please write your response now." The prompt remained on the screen for 20 seconds. A timer at the bottom of the screen counted down the time remaining. When five seconds remained, two brief tones were sounded.

Experimental Conditions

The study materials were presented under twelve test conditions, which included a Video-Alone control condition as well as the following:

Table 1

Auditory and audio-visual conditions for current investigation

Audio Alone (AA)	Audio Visual (A+V)
Tone1	Tone1
Tone2	Tone2
Tone3	Tone3
BFN	BFN
T123	T123
Natural	

Procedure

Subjects were run in groups of 1 to 9. They were seated in small experimental classroom with a 31-inch color Phillips 31P460-C402 monitor. The acoustic stimuli were presented using a loudspeaker at a comfortable listening level of approximately 75 dB SPL.

Prior to each test session, subjects were given a set of transcription practice sentences to familiarize them with the stimuli and the task (see Remez et al., 1994). The practice portion consisted of 8 sentences made up of T1 + T2 + T3. Each sentence was played five times. The first three sentence transcriptions were given to the subjects. The subjects had to try to transcribe the remaining five sentences without feedback. Following the test session, the practice items were again presented to listeners. The listeners were not told that they would be presented with the same eight sentences again. Listeners were included in the final data analysis only if they could recognize the first three sentences as the first three of the practice session. Two hundred and sixty-two subjects met this selection criterion and were used for the final analysis.

Results

Figure 1 shows the percentage of correct syllables in the transcription task for the auditory-alone combination tone 1 + 2 + 3, audio-visual combination tone 1 + 2 + 3, and the visual-alone condition. An overall analysis of variance for these three conditions revealed a highly significant effect of stimulus presentation, $F(2,81)=182.077$, $p<.001$. Post-hoc contrasts revealed that the performance in the audio-visual condition was significantly greater than either the T 1 + 2 + 3 auditory-alone condition or the video-alone control condition, $F(1,81)=362.45$, $p<.001$.

Insert Figure 1 about here

Figure 2 shows the percentage of correct syllables transcribed for the auditory alone single tones and bit-flipped noise, the audio-visual single tones and bit-flipped noise, and the visual-alone condition. The results from the audio natural condition were excluded from the overall ANOVA because subjects were at ceiling (100%). An overall ANOVA on the audio-alone (A-A) conditions revealed a significant difference in performance, $F(3,77)=16.370$, $p<.001$. Post- hoc comparisons showed that intelligibility performance for tone 2 was significantly greater than tone 1, tone 3, or BFN, $F(1,77)=41.885$, $p<.001$.

Insert Figure 2 about here

An overall analysis on the audio + visual conditions as well as the visual-only condition also revealed an overall effect, $F(4,113)=7.378$, $p<.001$. Post-hoc contrasts showed that the percentage correct for the audio-visual tone 2 condition was significantly higher than the A-V condition for tones 1, 3, BFN, and video-alone, $F(1, 113)=23.178$, $p<.001$. The contrasts also showed a significant benefit for the tone 2 + video condition over the visual-alone condition, $F(1, 113)=13.218$, $p<.001$. There were no differences between the visual-alone scores and the audio + visual scores for Tone 1, 2, or bit-flipped noise.

Discussion

The results from the present investigation suggest that the process of audio-visual integration is super-additive in nature. Intelligibility for the combination tone increased by over 150% when the visual signal was provided to the subject. This is well over the increase we would predict given the participants' performance in the video alone condition. Furthermore, analyses of the single tones and BFN conditions demonstrate that the tonal analog of the second formant (T2) is the most perceptually effective in a

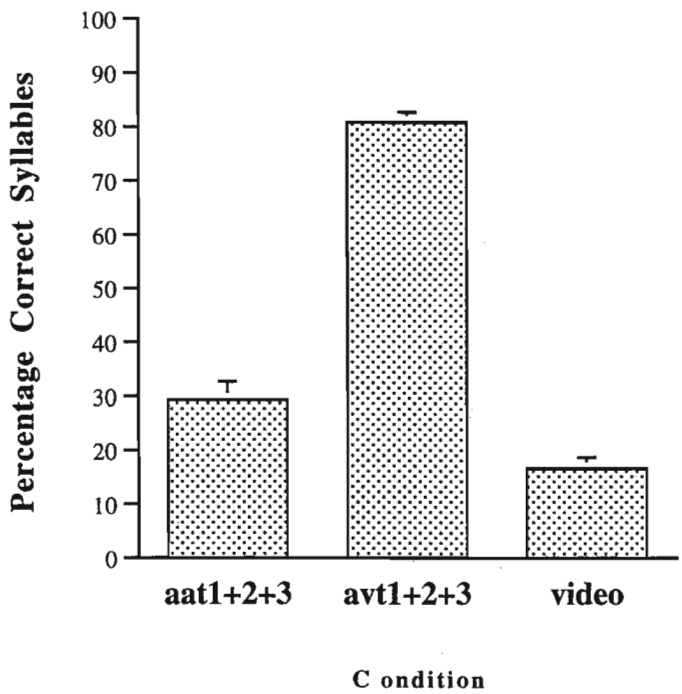


Figure 1. Percentage of correct syllables for the combination tones, auditory alone, and video alone conditions.

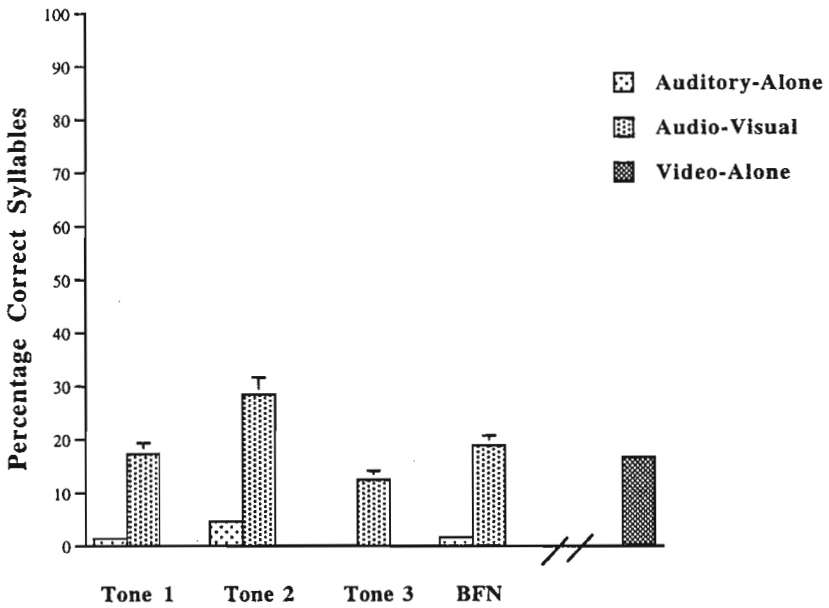


Figure 2. Percentage of correct syllables for individual tones, bit flipped noise, and video alone conditions.

multimodal context. No perceptual benefit was observed in the cross-modal conditions when the tonal analog of the first formant (T1), third formant (T3), or envelope-shaped noise was presented to the listeners. We offer two possible explanations for the observed increase in intelligibility for T2: (1) the variation of the second formant might provide redundant information to the information contained in the dynamic visual display or (2) the tonal analog of the second formant may add nonredundant or complementary information to the information already available in the dynamic visual display. Additional experiments are underway to determine whether the sinewave signal directs the perceiver's attention to information available in the visual display, or whether the two modalities each contribute independent sources of phonetic information conjointly.

The present results indicate that cross-modal integration in speech perception does not depend on speechlike *auditory qualities*, but rather on speechlike *auditory spectral variation*. It is tempting to speculate that cross-modal integration occurs with these unusual acoustic stimuli because we are presenting formant information. However, the tones presented to our listeners are a far cry from the formant structure of natural speech. They contain no fundamental frequency and no harmonic structure. Why are these very simple patterns integrated as if they are formants and why is there such a large gain in performance when this time-varying auditory information is combined with the dynamic information present in the optical display? These are questions that we plan to address in future research.

We believe the present results raise a number of important new questions about speech perception and spoken language processing. It is clear that coherent variation within and across auditory and visual modalities provides the perceiver with reliable information about a unitary perceptual event that is distributed in both time and space. This unitary multi-modal speech event deserves our continued attention in both the mature perceiver and in development. Any complete theory of speech perception and spoken language processing must begin to take these fundamental facts about perception and perceptual systems into account and must offer a coherent framework for explaining multi-modal perception of linguistic events.

References

Grant, K.W., Braida, L.D., & Renn, R.J. (1994). Auditory supplements to speechreading: Combining amplitude envelope cues from different spectral regions of speech. *Journal of the Acoustical Society of America,* **95**, 1065-1073.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception process. *Perception & Psychophysics*, **24**, 253-257.

Markel, J. Gray, A., Jr. (1976). *Linear Prediciton of Speech*. Springer-Verlag: New York.

Massaro, D. W. (1987). *Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry.* Erlbaum: Hillsdale, NJ.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature,* **264**, 46-748.

Remez R.E. , Fellowes J.M., & Rubin P.E. (in press).Voice identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*.

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

Remez, R.E. & Rubin, P.E. (1984). Perception of intonation in sinusoidal sentences. *Perception & Psychophysics*, **35**, 429-440.

Remez, R.E., Rubin, P.E., Berns, S.E., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review,* **101**, 129-136.

Rubin, P.E. (1980) *Sinewave Synthesis.* Internal memorandum, Haskins Laboratories, New Haven, CT.

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. (1981). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 3-51). Erlbaum: London.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Perceptual Learning of Natural and Sinewave Voices[1]

Sonya M. Sheffert, David B. Pisoni, Jennifer M. Fellowes[2] and Robert E. Remez[2]

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

[1] This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University-Bloomington, and NIH-NIDCD Research Grant DC00308 to Barnard College, Columbia University.

[2] Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027.

Perceptual Learning of Natural and Sinewave Voices

Abstract. This report describes the results of a perceptual training study that was designed to explore how listeners learn to categorize novel voices and how knowledge of a familiar voice generalizes to novel utterances. The speech samples from which the listeners learned to identify individuals were of two kinds: Naturally produced English sentences and sinewave replicas of these sentences. The sinewave items were nonspeech tonal patterns that preserved coarse-grained properties of the talker's vocal tract transfer function while eliminating traditional cues to voice quality. Listeners were trained over several days to identify by name ten talkers from sentence length sinewave or natural speech utterances. Knowledge about the talker's voice was then assessed using two generalization tests in which listeners heard a novel set of sentences and were required to identify the speaker. In one generalization test, the sentences were sinewave replicas whereas in the other generalization test, the sentences were naturally produced. The results showed that perceptual learning of a talker's voice can occur even when specific acoustic products of vocal articulation are eliminated from the signal. The data also showed that speaker-specific knowledge acquired during this perceptual training task generalized to novel natural and novel sinewave sentences. Variability in the degree of perceptual learning affected generalization of speaker knowledge. The results of this study show that listeners can learn about a talker's voice from highly impoverished acoustic signals when the products of vocal articulation are eliminated, and that this knowledge generalizes to novel utterances produced by these same talkers.

Introduction

When a speaker produces an utterance, the listener recovers from the acoustic signal not only information about the consonants and vowels that compose the message, but information about the speaker's vocal tract morphology, affect and pronunciation habits. Historically, the perception and representation of the linguistic content of an utterance has been thought to be separate and independent from the processes needed for the encoding of speaker information (Halle, 1985; Laver & Trudgill, 1979). Recent findings, however, suggest a different perspective on the relation between linguistic and "indexical" information in speech, namely, one in which these two sets of attributes interact. The line of evidence most relevant to the current investigation comes from a recent series of perceptual learning studies conducted by Nygaard and colleagues (Nygaard & Pisoni, 1995; see also Nygaard, Sommers & Pisoni, 1994). In their procedure, listeners are trained over several days to identify a set of talkers from sentence length utterances. Subjects then are given a speech intelligibility test in which they are asked to transcribe new sentences presented in white noise. Nygaard and Pisoni (1995) found that subjects who had become familiar with the talkers were able to transcribe sentences more accurately than subjects who were unfamiliar with the speakers. This finding demonstrates that familiarity with the speaker who produced a sentence facilitates perceptual analysis of novel sentences, and supports the inference that linguistic and nonlinguistic information are not processed independently (see also Mullennix & Pisoni, 1990). This interaction between speaker and phonetic information also suggests similarities in the nature of the perceptual operations and neural representations that underlie voice recognition and speech perception.

Although the perceptual dimensions in which a voice is represented in memory are largely unspecified, several acoustic-phonetic cues have been traditionally assumed to underlie speaker recognition

(Bricker & Pruzansky, 1976; Laver & Trudgill, 1979). These include fundamental frequency, vocal tract resonances, glottal source, harmonic structure and the fine-grained power spectra of nasals and vowels. Recent findings reported by Remez, Fellowes and Rubin (in press) call into question the necessity of these traditional cues for speaker recognition. They demonstrated that listeners can identify familiar talkers in the absence of such variables using the technique of sinusoidal speech synthesis, which generates a nonspeech pattern that specifies the dynamics of a talker's vocal tract transfer function. Sinewave utterances are time-varying sinusoidal patterns that track the changing center frequencies of the naturally produced utterance from which they are modeled. These nonspeech tonal stimuli can be thought of as a highly simplified representation of the frequency and amplitude changes present in speech, an "acoustic caricature" of the original utterance. Although the signal is highly impoverished, most listeners are nevertheless able to perceive the linguistic content of the utterance (Remez, Rubin, Pisoni & Carroll, 1981), indicating that the dynamic properties of the sinewaves are sufficient to support phonetic perception.

The recent findings of Remez et al. (in press) further show that the global time-varying properties of the sinewaves also preserve speaker-dependent aspects of speech. In Remez et al. (in press), sinewave utterances modeled from the natural productions of ten speakers were presented to listeners in a voice recognition task. The listeners were members of the staff at Haskins Laboratories who had become highly familiar with the speakers over many years of social contact. In this task, they were required to identify the speaker from which the replica originated. To perform this voice recognition task, listeners had to draw on their long-term knowledge of a talker's voice and speaking style and compare this to the information preserved in the sinewave signals. The results showed that listeners were generally successful at identifying the familiar voices of their colleagues from the sinewave utterances. Identification accuracy exceeded chance for 6 of the 10 speakers. This finding indicates that information in the sinewaves specifying changes in the talker's vocal tract is sufficient to support talker identification among listeners who were highly familiar with the speakers, and that speaker identification can take place even when traditional voice recognition cues are eliminated from the signal.

The results also showed considerable variability in the recognizability of different speakers, as evidenced by the fact that recognition accuracy was below chance for some speakers in the set. Because Remez et al. (in press) did not experimentally manipulate or control familiarity, it is impossible to know whether the observed variability was due to differences in the degree of familiarity, or to other factors present in the speaker ensemble, such as perceptual distinctiveness or discriminability of the stimuli. The present study uses the same test items as Remez et al. (in press), but controls for the amount of speaker familiarity by using a laboratory-based training procedure (cf. Nygaard, et al., 1994) in order to examine in detail how variability in the rate and degree of perceptual learning affects the identifiability of different speakers.

The present investigation has several objectives. First, in Experiment 1, we wanted to establish whether perceptual learning of voices can take place in the absence of the acoustic cues typically assumed to underlie talker identification. Second, we wanted to assess the effects of perceptual learning of sinewave speakers on the generalization to novel utterances. In particular, will perceptual learning of speaker information be context-specific, generalizing only to other sinewave replicas, or will training also generalize to novel natural speech utterances? A final objective was to assess the effect of the degree of perceptual learning on the generalization of speaker-specific knowledge.

Experiment 1

Familiarity of the sinewave speakers was experimentally manipulated by training listeners to identify by name the ten talkers producing the original sentences. In Experiment 1, the sentences were sinewave replicas of the natural utterances. Subjects were trained until they were able to identify the ten talkers with at least 70% accuracy. Speaker knowledge was then assessed using two generalization tasks in which listeners heard a novel set of sentences and were required to identify the speaker. In one generalization test, the sentences were sinewave replicas whereas in the other generalization test, the sentences were naturally produced utterances. In both cases, the generalization tests used utterances that the subjects had not heard before during training.

Based on the findings of Remez et al. (in press), which demonstrated talker identification from sinewave utterances, we predicted that with appropriate training and feedback, the higher-order relational information preserved in the sinewave replicas would support the perceptual learning of speakers. We also expected that the speaker-specific knowledge acquired during sinewave training would not be dependent on the specific training items but would instead generalize to novel natural and novel sinewave sentences. We did, however, expect that generalization would be based on the similarity to known examples, and consequently, would be greatest in the condition in which the perceptual form of the sentences was the same across training and test. Specifically, talker identification would be higher on the sinewave generalization test as compared to the natural speech generalization test.

Method

Subjects

Nineteen adult subjects were recruited from the Bloomington, Indiana, community. Of these, five subjects failed to complete the study due to work or school commitments, and six were excused because of extremely slow progress during the initial training sessions. The remaining eight subjects completed the sinewave training phase and the two generalization tests. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. Subjects were paid for their participation.

Test Materials

The natural and sinewave sentences used in the present experiments were the same materials developed by Remez et al. (in press). The stimulus materials consisted of two sets of sentences. The first set contained nine natural utterances produced by five male and five female talkers. Each talker produced all nine sentences, for a total of 90 items. Audio recordings were obtained by asking speakers to read the sentences aloud in their natural speaking style. The sentences were then recorded on audiotape in a sound-proof booth and were low-pass filtered at 4.5 kHz, digitally sampled at 10 kHz, equated for root mean squared (RMS) amplitude and stored as sampled data with 12-bit resolution.

The second set of sentences were sinewave replicas of the original natural speech tokens. To create these items, the frequencies and amplitudes of the first three formants were derived at 5 msec intervals interactively relying on two representations of the spectrum: 1) linear predictive coding (LPC), and 2) discrete fourier transforms (DFT). Three time-varying sinusoids were then synthesized based on the center frequencies and amplitudes of the formants (Rubin, 1980). The synthesis algorithm preserved higher-order patterns of spectro-temporal change of the vocal tract transfer function, while eliminating the fundamental

frequency, harmonic relations and fine-grained spectral information. Subjectively, the sentences were difficult to understand and sounded very unnatural.

Three sentences were randomly selected (without replacement) for each of the three phases of the experiment (training, natural speech generalization and sinewave speech generalization). All sentences were rotated through all conditions for each listener to ensure that the observed effects were not due to any specific subset of the sentences or any order effects.

Procedure

Training Phase

Listeners were trained over several days to learn to identify the names of the 10 speakers using the sinewave utterances. Subjects were tested in groups of three or fewer in a quiet listening room. During each training session, subjects heard a random ordering of five repetitions of three sentences from each talker (150 items total). The same three sentences were used for each talker in each training session, and subjects were told before hand which three sentences they would be hearing. The sinewave training sentences were presented binaurally to subjects at 75 dB SPL over matched and calibrated stereophonic headphones (Beyerdynamic DT100). Subjects were asked to listen carefully to each sentence and to pay close attention to the talkers' voices. Each time a sentence was presented, the subject was required to press one of ten keyboard buttons labeled with each speaker's name. Keys 1-5 were labeled with female names and keys 6-10 with male names. Each time a subject made a response, the accuracy of that response and the name of the correct talker was displayed on the computer screen in front of the subject and recorded in the computer. Each training session lasted approximately 30 minutes. Training was continued until subjects achieved an average of 70% correct speaker recognition performance.

Familiarization Phase

Before beginning each of the generalization tests, subjects completed a brief familiarization task to remind them of the correspondence between the sinewave tokens and the speakers. The familiarization task was simply an abbreviated version of a training session in which subjects listened and responded to one instance of each sentence in each talker (30 items total). The items were presented in a random order and subjects received feedback after each response. The familiarization task took approximately 8 minutes.

Generalization Tests

After reaching a 70% correct criterion in the sinewave training phase, subjects completed two generalization tests. One generalization test presented three novel sinewave sentences, whereas the second test presented three novel naturally produced sentences. All of the sentences presented during the generalization tests were new to the subject. Half the subjects received the natural generalization test before the sinewave generalization test, while the other half received the tests in the opposite order. Each generalization test presented five repetitions of each of the three sentences in a random order (150 items total). Subjects were informed of the sentences they would be hearing before the start of each test. Subjects were asked to attend specifically to the talker's voice and to identify the speaker by pressing one of the ten buttons on the keyboard as they had done in the previous training phase. Subjects did not receive feedback during either of the two generalization tests.

Results and Discussion

Training Performance

Analysis of the training data revealed that listeners were, in fact, able to learn to identify the ten speakers from these highly impoverished sinewave signals. Their performance showed continuous improvement during the training phase. After the first training session, talker identification performance was above chance and steadily increased by an average of 5% each day. By the last day of training, listeners were able to identify the speakers with a mean accuracy of 76%. Figure 1 displays the mean identification performance as a function of training days and speaker sex. Figures 2a and 2b display each subject's talker identification performance as a function of training days and speaker sex. The latter figures illustrate that learning progressed at different rates for different subjects. The number of days needed to reach the 70% criterion varied among the subjects from 9 days to 16 days. Because the number of training days differed across subjects, speaker recognition improvement was assessed using a sample of four training days. Specifically, we used the first day of training, two evenly separated days in the middle, and the last day of training for the statistical analysis. For example, 16 days of training were reduced to Days 1, 6, 11 and 16.

Insert Figure 1 about here

Insert Figures 2a and 2b about here

A repeated measures analysis of variance with the factors of days of training and speaker sex was conducted on the accuracy data (shown in Figure 1). The analysis revealed significant effects for days of training, $F(3, 42) = 221.49$, $p < .0001$, and speaker sex, $F(1, 14) = 34.02$, $p < .0001$. The latter reflected the advantage in the identifiability of the female talkers during training.

The data from the last day of training also showed variability in the identifiability of different speakers within the training set. Figure 3 shows speaker recognition performance on the last day of training as a function of speaker. Examination of the graph shows the two most accurately identified talkers were female ("F1" and "F2") whereas the two least accurately identified talkers were male ("M2" and "M5"). An ANOVA comparing talker recognition performance on the last day of training revealed a significant effect of speaker sex, $F(1, 78) = 18.69$, $p < .0001$, confirming that the female speakers were more accurately identified than the male speakers.

Insert Figure 3 about here

We also observed considerable variability among the speakers within each sex. An ANOVA with speaker as a factor was conducted on the training scores for each sex. Differences were found among the female speakers, $F(4, 28) = 6.01$, $p < .001$, and the male speakers, $F(4, 28) = 8.35$, $p < .0001$. Taken together, the training data demonstrate that listeners can learn to identify individuals from sinewave replicas

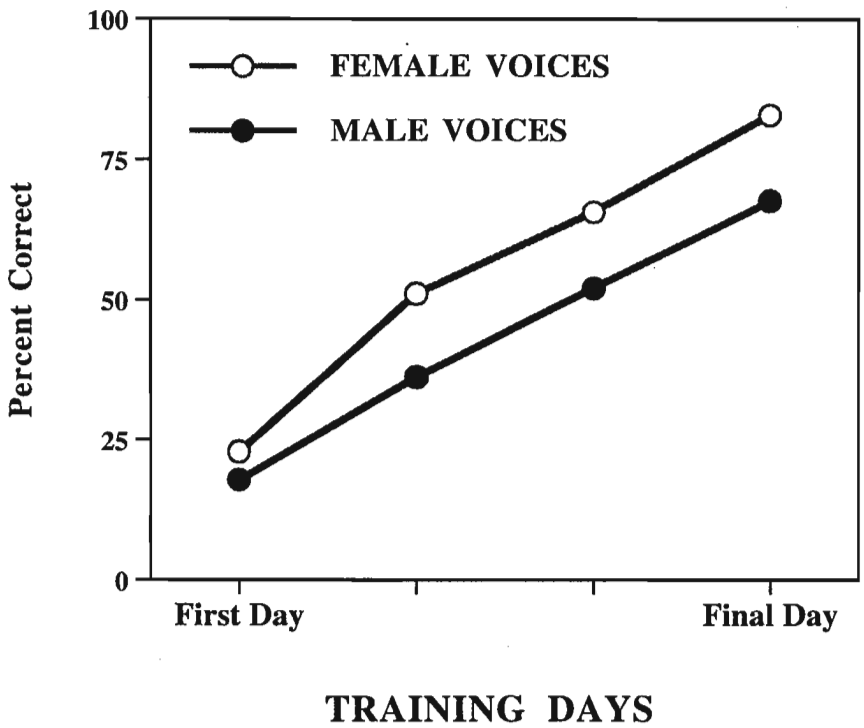
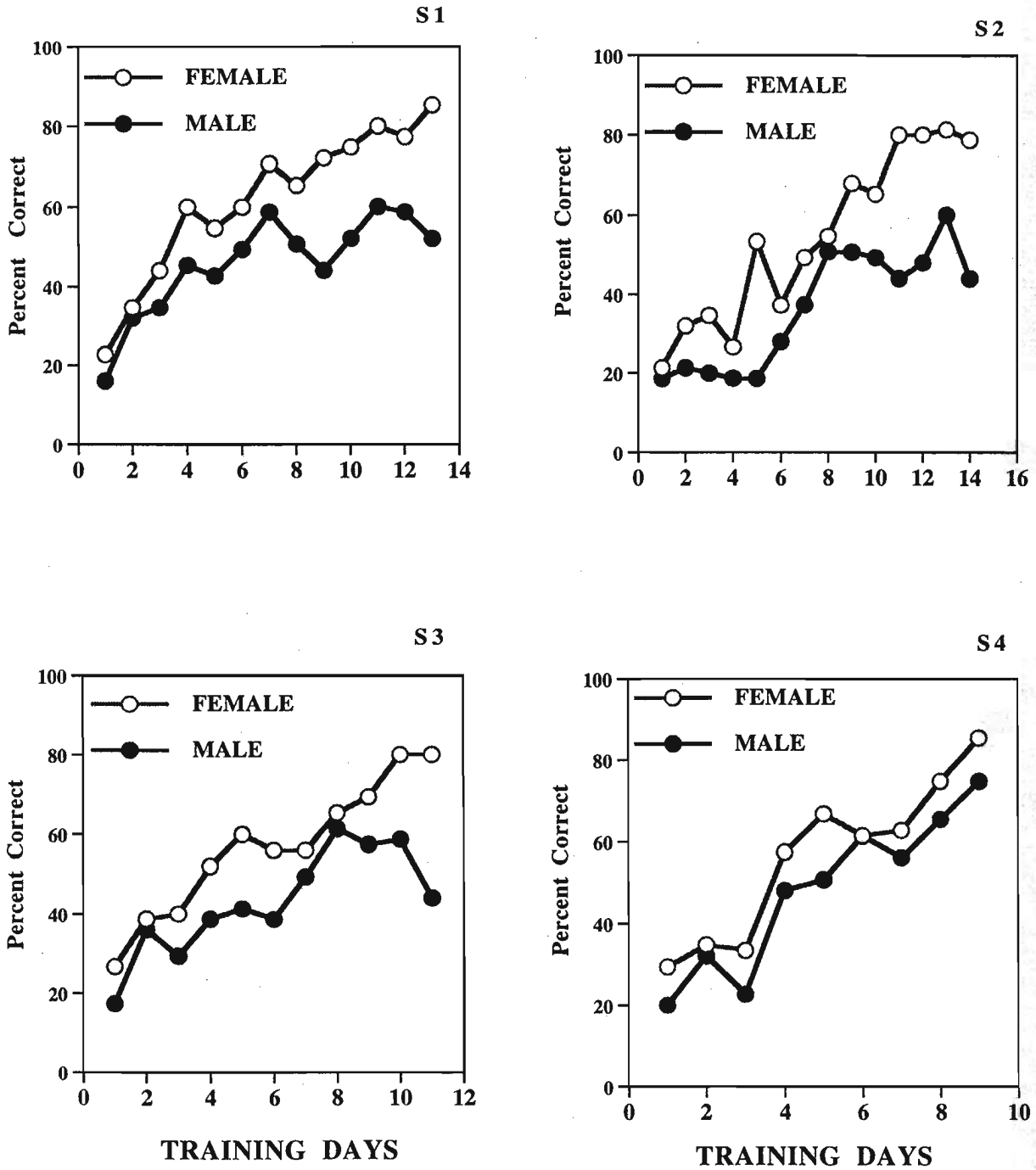
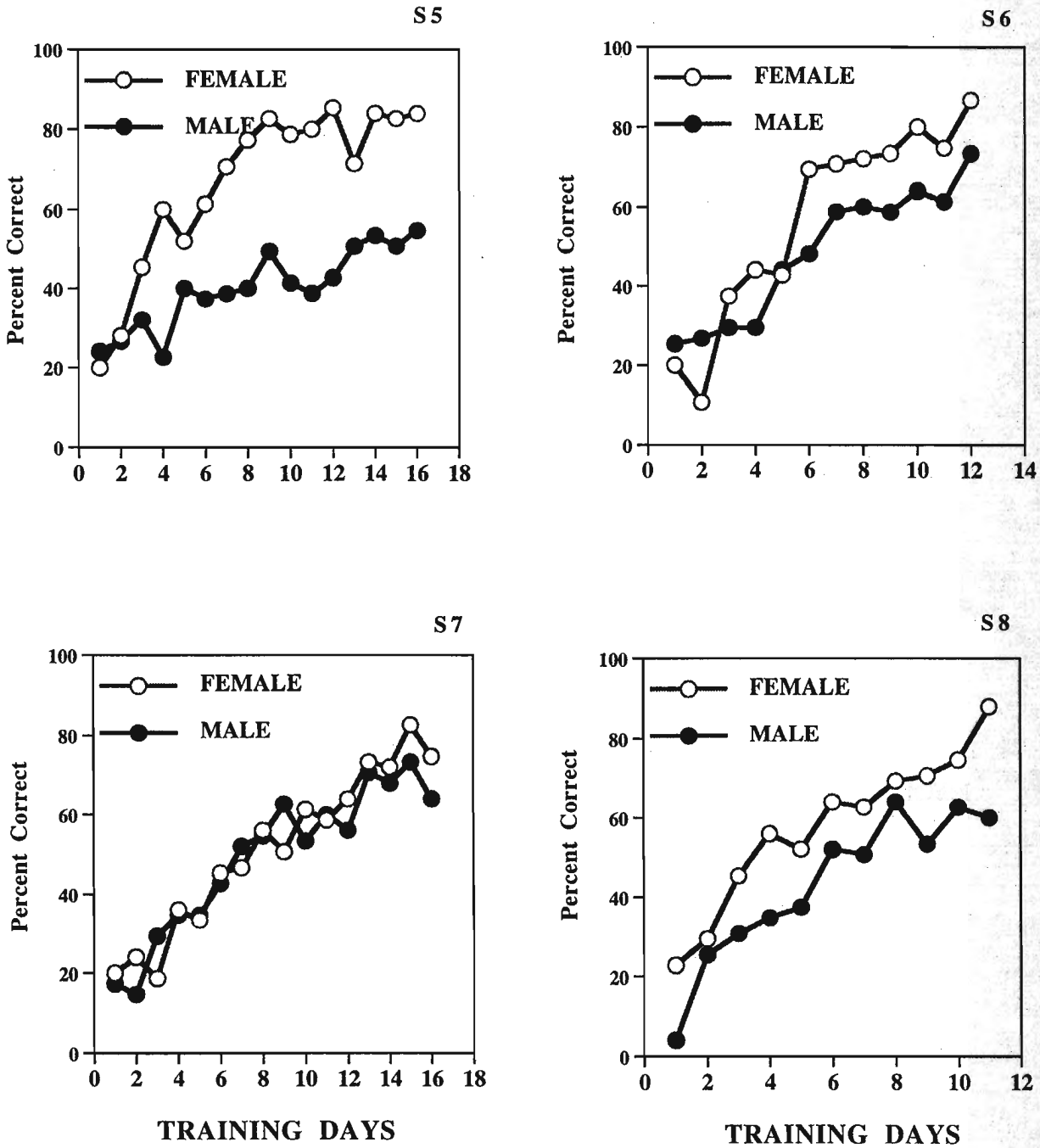


Figure 1. Mean speaker identification performance on sinewave replicas as a function of training days and speaker sex.



Figures 2a. Mean speaker identification performance on the sinewave training for subjects 1-4 as a function of training days and speaker sex.



Figures 2b. Mean speaker identification performance on the sinewave speech training for subjects 5-8 as a function of training days and speaker sex.

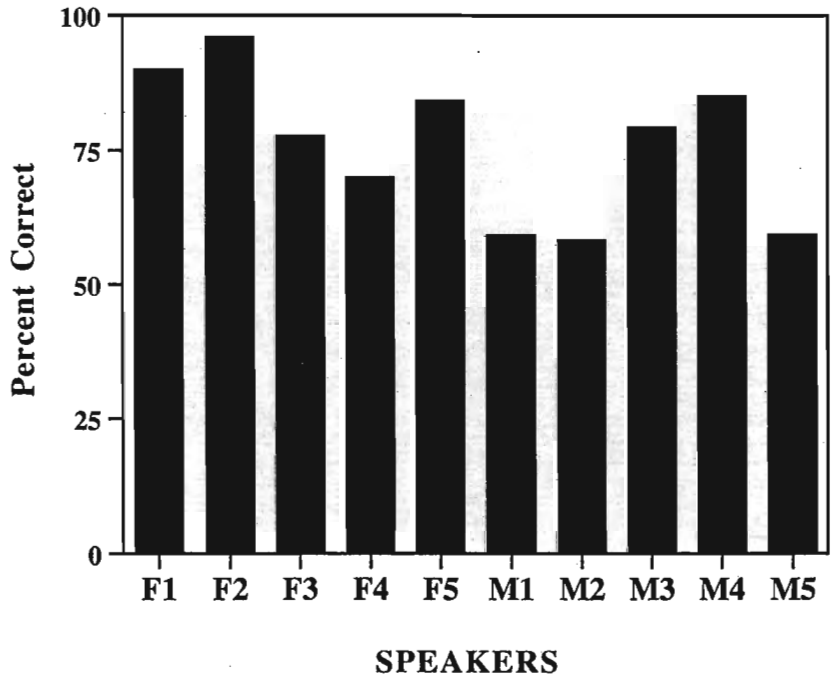


Figure 3. Speaker identification performance on sinewave replicas for the last day of training as a function of speaker. F1 though F5 refer to the female speakers; M1 through M5 refer to the male speakers.

of natural speech, although the rate and degree of learning varies as a function of the identifiability of different speakers within the training set.

Generalization Performance

Because there were large differences in the identifiability of different speakers within the training set, the statistical analysis of the generalization tests was conducted using generalization scores to normalize for different levels of performance. These scores were obtained by dividing the talker identification accuracy on the generalization test performance by talker identification accuracy on the training task.

At the time of the generalization testing, half of the subjects received the natural speech generalization test before the sinewave generalization test, whereas the other half completed the tests in the opposite order. To assess whether the order of the tests affected speaker recognition, an ANOVA on test order was conducted on the generalization scores from each generalization test. In both cases, the effect of test order was not significant. Consequently, the data from the two groups were pooled and all subsequent analyses ignore this factor.

Figure 4 displays the generalization scores for the natural speech and sinewave generalization tests for each speaker. The generalization data from both tests show that speaker-specific knowledge acquired during perceptual learning of sinewaves was not dependent on the specific sample used during training, but generalized to novel sentences including both natural and sinewave materials. Moreover, the same level of generalization occurred in both tests, despite the fact that in the natural speech condition, both the content and the acoustic form of the sentences differed from the items used during training. Specifically, the data showed that listeners' ability to recognize speakers decreased from 76% correct at the end of training to 46% correct for the natural test and 44% correct for the sinewave test. An ANOVA comparing the overall means from each of the three conditions revealed a significant effect, $F(7, 2) = 5.23$, $p < .0001$. Recognition was significantly different from training for the natural speech trials [$t(7) = 7.34$, $p < .001$] and for the sinewave replica trials [$t(7) = 11.18$, $p < .0001$]. However, the difference in generalization between the two generalization tests was not significant.

Insert Figure 4 about here

One question we were interested in concerns whether the differences listeners exhibit in their ability to identify speakers during training generalizes to novel utterances. During training, female speakers were recognized more accurately than the male speakers. In contrast, although there was a numerical trend for the female speakers to be recognized more accurately than the male speakers in both generalization tests, these differences were not reliable in either condition, as shown in Figure 5.

Insert Figure 5 about here

The training data also showed within-sex variability in speaker identification. However, smaller within-sex differences were found for the generalization test data. An ANOVA with the factor of speaker

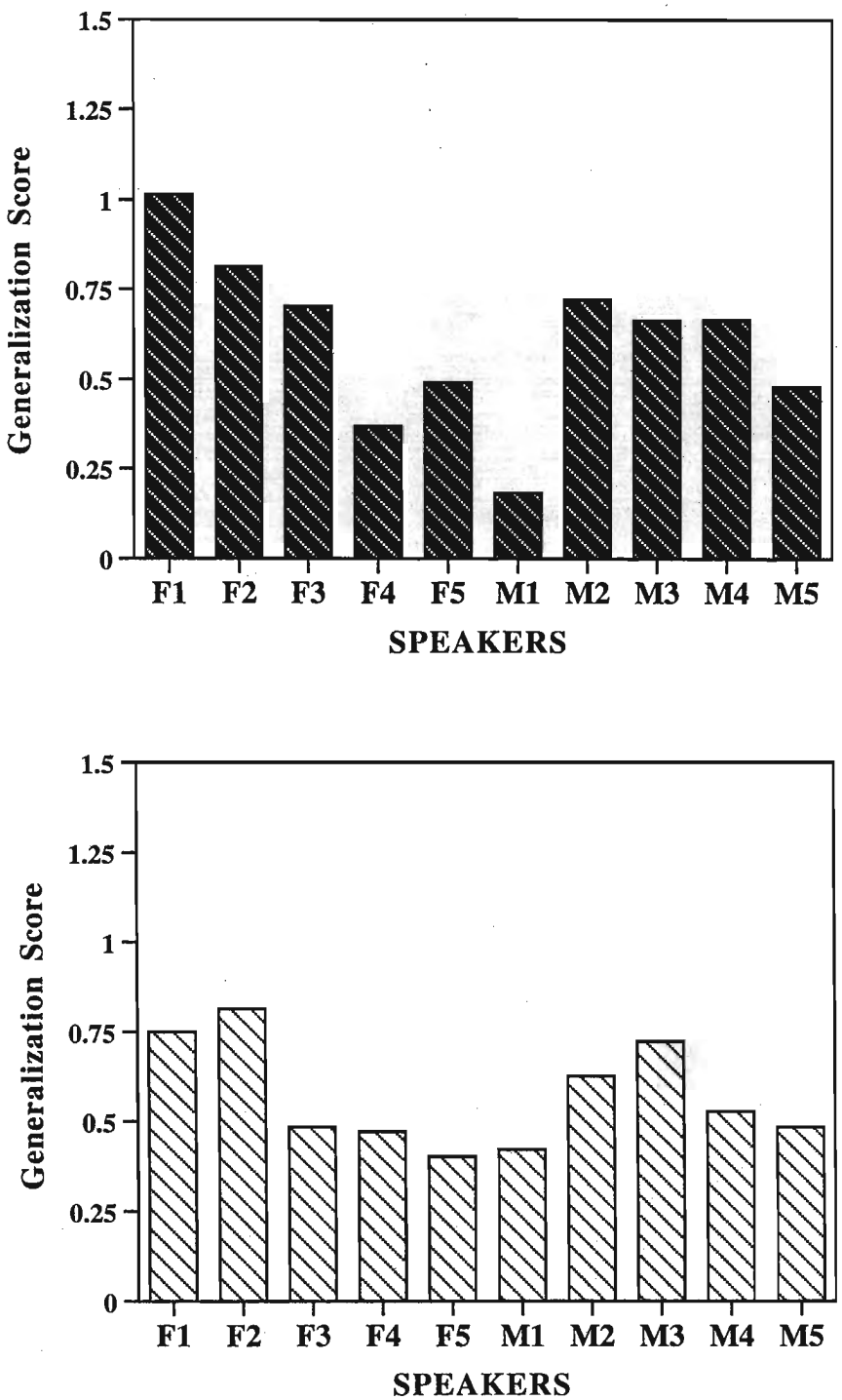


Figure 4. Mean speaker identification performance on the natural speech generalization (top panel) and sinewave replica generalization (bottom panel) as a function of training days and speaker sex. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

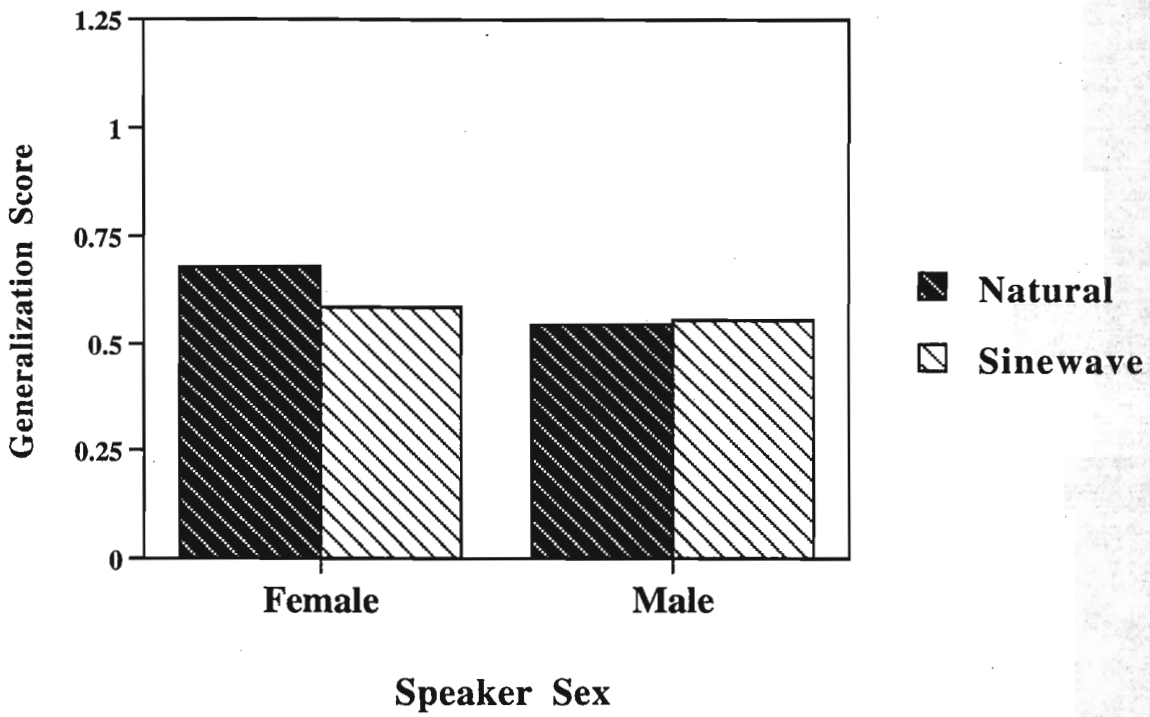


Figure 5. Mean generalization scores on the natural speech and sinewave replica generalization tests (following sinewave training) as a function of speaker sex.

was conducted on the natural speech generalization scores for each sex. The effect of speaker was marginally significant for the female speakers, $F(4, 28) = 2.63$, $p < .06$, but statistically significant for the male speakers, $F(4, 28) = 2.38$, $p < .05$. For the sinewave generalization task, reliable differences were found only among the female speakers, $F(4, 28) = 7.68$, $p < .001$. The effect of speaker did not approach significance for the male speakers.

Another way to examine the relationship between perceptual learning and speaker generalization is to examine the relative ranking of the speakers across each condition. The data show a modest overall rank order correlation for the ten speakers' identifiability at training and at test (Spearman's rho = .542 for natural speech; .492 for sinewave stimuli), suggesting that the speakers that were most easily identified at training were also likely to be well identified at test. Curiously, the correlation between each of the generalization tests was quite high (Spearman's rho = .830), indicating that the identifiability of a particular talker was more similar across the two generalization tests than across the sinewave training and generalization test.

In summary, we controlled speaker familiarity in this experiment by training all our listeners to a specific talker identification accuracy level. The motivation for this was to determine if the variability in speaker identification found by Remez et al. (in press) arose from differences in the discriminability of speakers within the speaker ensemble, or, instead, from differences in a priori speaker familiarity. In the present study, as in Remez et al., we observed striking differences in the identifiability of speakers within our training set. This suggests that perceptual distinctiveness or discriminability of the speakers in the set is the source of the discrepancies in speaker identification performance. In addition, we found that listeners can be trained to identify different speakers from nonspeech tonal analogs of speakers' utterances. This shows that talker identification can be accomplished solely from phonetic attributes present in these sinewave replicas, in the absence of voice quality. We also found evidence that perceptual learning from the sinewave training task generalized to novel natural and novel sinewave sentences. The fact that generalization was equivalent across the two generalization tests shows that the perceptual learning was not context-specific, and suggests the possibility that the same acoustic correlates of voice identity are being utilized in both the sinewave and natural speech generalization tests It would be useful to determine whether the symmetry in generalization performance that occurred following training on the sinewave utterances is particular to that perceptual learning task, or if the pattern of generalization can be found following perceptual learning of natural voices. Experiment 2 provides this comparison.

Experiment 2

The design and method of Experiment 2 was identical to Experiment 1, except that the training sentences were natural speech utterances, rather than sinewave replicas. The natural sentences were used to train listeners until they were able to identify the 10 speakers with at least 70% accuracy. Generalization of speaker knowledge was then assessed using natural speech and sinewave generalization tasks in which listeners heard new sentences and identified the speaker. Based on previous research (Nygaard & Pisoni, 1995), we predicted that the perceptual learning of talkers from natural speech sentences would proceed rapidly and would readily generalize to novel natural speech sentences. Furthermore, the findings from Remez et al. (in press) and from Experiment 1 lead us to expect that training on the natural speech samples would facilitate speaker identification from sinewave replicas.

Method

Subjects

Eight new subjects were recruited from the Bloomington, Indiana, community. All subjects were native speakers of American English and reported no history of a speech or hearing disorder. Subjects were paid for their participation.

Test Materials

The 90 natural and 90 sinewave sentences used in Experiment 2 were identical to the sentences used in Experiment 1. Three sentences were randomly selected (without replacement) for the training, natural speech generalization and sinewave speech generalization tasks. All sentences were rotated through all conditions for each subject to eliminate any effects due to specific stimulus items or order effects.

Procedure

Training Phase

Listeners were trained to learn to explicitly identify the names of 10 speakers from the natural speech sentences utteranced under the same training conditions used in Experiment 1. Again, training was continued until subjects achieved an average of 70% correct talker identification performance.

Familiarization Phase

The familiarization task preceded the natural speech and sinewave replica generalization tests. A random ordering of 30 natural speech training items was presented for speaker identification. Subjects received feedback after each response using the same methods as in Experiment 1.

Generalization Tests

The two generalization tests (natural speech and sinewave replica) were identical to those used in Experiment 1. Each generalization test presented five repetitions of three novel sentences in a random order (150 items total) for speaker identification. Subjects were informed about the sentences they would be hearing before the start of each test. They were also were instructed to attend to the talkers voice and to identify the speaker by pressing the appropriate button on the keyboard. There was no feedback. The test order was counterbalanced across the subjects.

Results and Discussion

Training Performance

As expected, listeners learned to identify the ten different speakers from the naturally produced sentences very rapidly. Identification performance reached criterion for five of the eight subjects after only one training session. The remaining three listeners reached criteria by the end of the second session. Speaker identification performance averaged 78%.

The training data also revealed differences in the identifiability of female and male speakers. Figure 6 displays identification performance on the last day of training as a function of speaker. The graph shows that overall, female talkers were identified better than male talkers (87% vs. 71% correct for female and male speakers, respectively). In fact, only one male speaker ("M4") exceeded the identification accuracy of the female speakers. An ANOVA comparing speaker recognition performance on the last day of training revealed a significant effect of speaker sex, $F(1, 78) = 17.5$, $p < .0001$.

Insert Figure 6 about here

Variability in talkers' identifiability was also observed within each sex. An ANOVA with the factor speaker was conducted on the training scores for each sex. Reliable differences were found among the female speakers, $F(4, 28) = 2.90$, $p < .05$, and the male speakers, $F(4, 28) = 3.14$, $p < .05$. These findings with natural speech are similar to the observed in Experiment 1 using sinewave replicas. In sum, the natural speech training data reveal rapid, although somewhat variable, perceptual learning of the individual talkers within our training ensemble.

Generalization Performance

As in Experiment 1, because we found no differences in the mean test performance as a function of test order, the two test order groups were pooled to form a single composite test group. The statistical analysis of the generalization tests following training on natural speech was conducted on generalization scores.

Figure 7 displays the generalization scores for the natural speech and sinewave replica generalization tests for each speaker. The data for the two generalization tests differed markedly. Listeners' ability to recognize individuals was 88% for the natural speech generalization test but only 27% for the sinewave generalization test. An ANOVA comparing the overall means from each of the three conditions (training, natural and sinewave) revealed a significant effect, $F(7, 2) = 5.23$, $p < .0001$. Surprisingly, performance on the training task (78% correct) was reliably *lower* than performance on the natural speech test [$t(7) = 3.17$, $p < .05$]. This effect may be the result of the familiarization task preceding the generalization tests. The purpose of the familiarization task was to remind subjects of the correspondence between a particular name and talker, and is itself an abbreviated training task. Although the familiarization task only presents 30 items, it nevertheless increased listeners talker knowledge, allowing them to perform better on the subsequent generalization test than on training task. Training performance was, however, significantly higher than performance on the sinewave replica generalization test [$t(7) = 21.5$, $p < .0001$]. In addition, the two generalization tests differed reliably from each other [$t(7) = 22.89$, $p < .0001$].

Insert Figure 7 about here

Female and male voices were recognized with equal accuracy in both generalization tests, as shown in Figure 8. An ANOVA with the factor of speaker sex was conducted separately on the natural speech generalization scores and the sinewave replica generalization scores. The effect of speaker was not significant in either of the generalization tests.

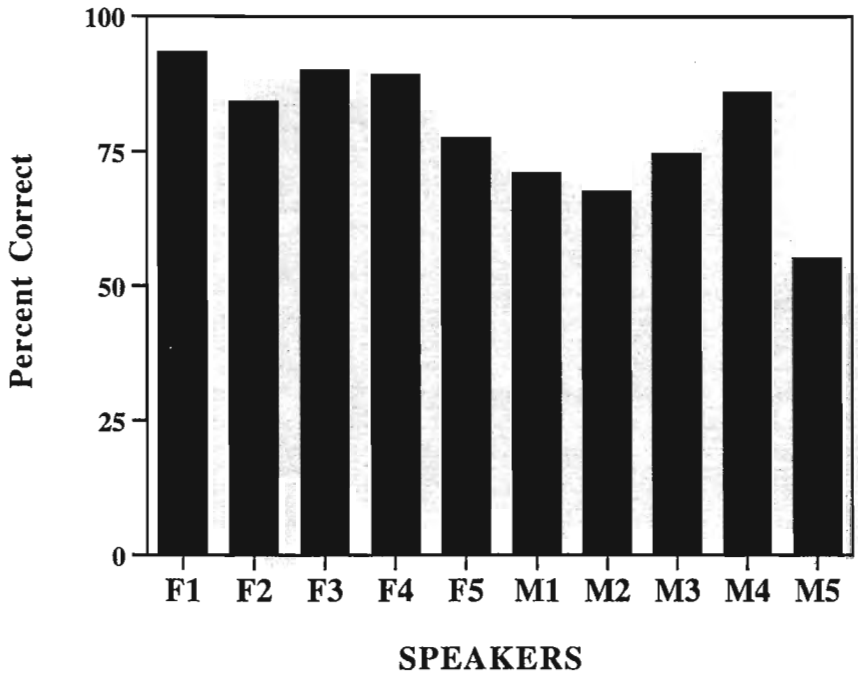


Figure 6. Mean speaker identification performance on natural speech for the last day of training as a function of speaker. F1 thought F5 refer to the female speakers; M1 through M5 refer to the male speakers.

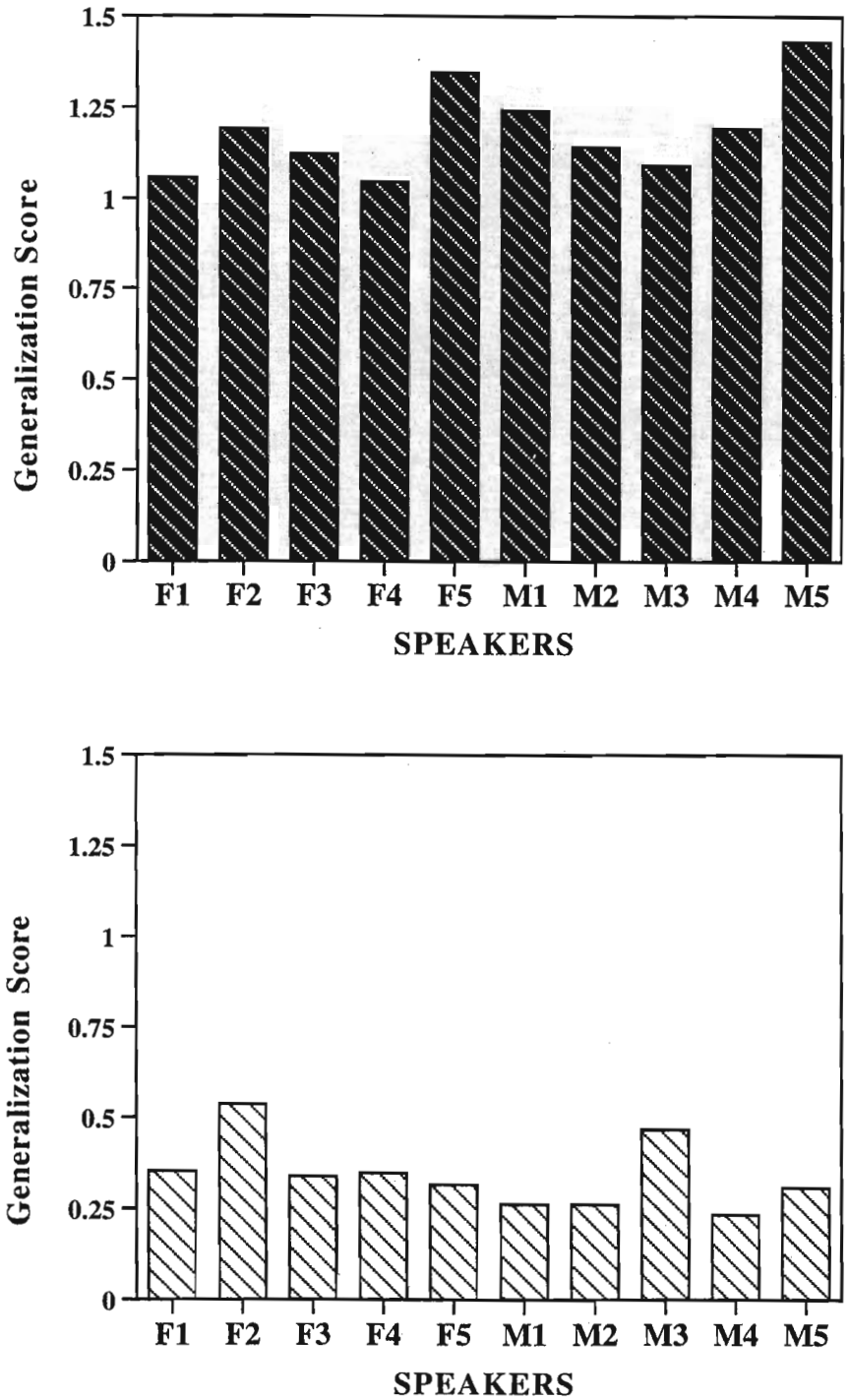


Figure 7. Mean speaker identification performance on the natural speech generalization (top panel) and sinewave replica generalization (bottom panel) as a function of training days and speaker sex. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

Insert Figure 8 about here

There were also no differences in the identifiability of voices among the female and male speakers in either generalization test condition, as shown in Figure 8. An ANOVA with the factor speaker was conducted on the natural speech generalization scores for each sex. The effect of speaker was marginal for the female speakers $F(4, 28) = 2.58$, $p < .06$. No significant differences were found among the male speakers. For the sinewave replica generalization task, the effect of speaker did not approach significance for either the female or the male speakers.

A Spearman's rho correlation was conducted to assess the relationship between perceptual learning and generalization performance. The analysis indicated that the identifiability of a particular talker was most similar across the natural training and natural test conditions (Spearman's rho = .752), less similar between training and the sinewave test (Spearman's rho = .515), and least similar across the two generalization tests (Spearman's rho = .424).

In summary, Experiment 2 shows that subjects easily learned to identify a talker from naturally produced sentences, and that this knowledge generalized readily to novel natural speech utterances. In contrast, subjects had difficulty recognizing speakers from novel sinewave utterances, although overall performance was marginally above chance. The implications of these findings will be considered in detail in the general discussion section.

General Discussion

The findings from these two experiments demonstrate that perceptual learning of a talker's voice can take place in the absence of traditional acoustic correlates of voice quality. In Experiment 1, we found that listeners were gradually able to learn to identify and label speakers from sinewave replicas of naturally produced sentences. This result, in conjunction with the recent findings of Remez et al. (in press), demonstrate that information about the dynamics of a talker's vocal tract transfer function include talker-specific information, and that this information can be exploited during perceptual learning. The perceptual training data from both experiments also revealed considerable variability in the identifiability of different speakers. In particular, female speakers from this set of voices were identified more accurately than male speakers, and this advantage for female speakers was equal in both experiments (16%). However, variability in the identifiability of different speakers was less evident in the generalization tests. In both experiments, differences in the overall identifiability of female and male speakers in either test condition were either small or absent.

We also found evidence that the knowledge obtained from the perceptual learning task can be readily generalized to natural and sinewave speech. In particular, subjects who learned sinewave samples were able to identify a talker from novel sinewave replicas and from novel natural speech samples with equal accuracy. Moreover, variability in the learning of different speakers also generalized well to both tasks. In fact, the ranking of talkers based on identification accuracy was more similar between the sinewave training and natural speech test than between the sinewave training and the sinewave generalization test .

We found exactly the opposite pattern of results in Experiment 2. For subjects who were trained on the natural speech sentences, stimulus generalization was greatest in the condition in which the perceptual

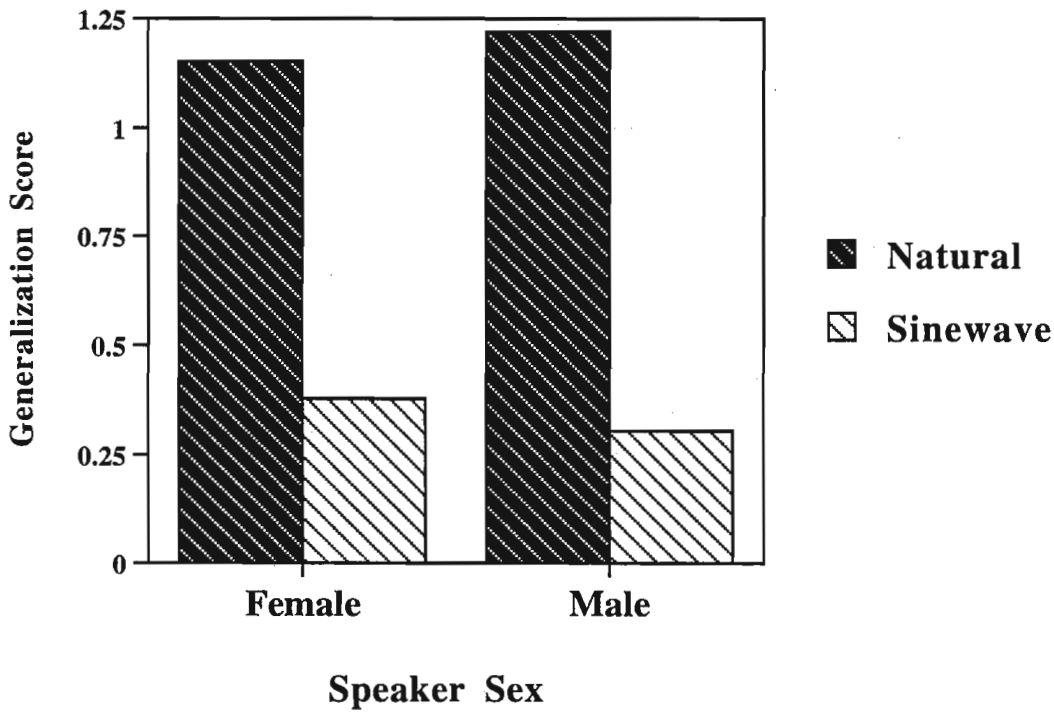


Figure 8. Mean generalization scores on the natural speech and sinewave replica generalization tests as a function of speaker sex.

form of the sentences was the same across training and test. The data show that talker identification at training was similar in magnitude and in patterning to the natural speech test, but quite different from the sinewave speech generalization test. Specifically, overall speaker identification performance and the identifiability of a particular talker was more similar across the natural training and natural test conditions than across the natural training and sinewave test conditions. However, accuracy identifying a speaker on the sinewave generalization test was very poor, although above chance.

A question to ask, then, is why the performance of our subjects was considerably worse than the performance of subjects in Remez et al. (in press). In the present study, listeners' ability to recognize individuals was only 27% for the sinewave generalization test, but approximately 55% in the talker identification test used by Remez et al (see their Figure 4). Four methodological differences in our experiment and theirs may account for the difference in sinewave speaker identification performance across these experiments. First, in our task, listeners heard three novel sentences three times during the generalization test, whereas in Remez et al. (in press) the same sinewave sentence from each of the ten talkers was presented six times. In this case, the fact that the linguistic content of the utterance was the same from trial to trial may have improved listeners ability to discriminate among the ten individuals. Second, our listeners did not have any knowledge or familiarity with sinewave speech before the generalization test task, whereas several of the listeners used by Remez et al. (in press) were familiar with what sinewave speech was and how it sounded. This may have increased subjects' ability to quickly focus their attention on the talker-specific phonetic information present in the sinewave replicas. Finally, in our experiment, familiarity was acquired through perceptual training with a small number of sentences over a few days. In contrast, the listeners in Remez et al. (in press) had acquired their speaker knowledge naturally from many hundreds or even thousands of utterances over the course of many years. We would expect, therefore, that they would have qualitatively richer speaker representations than the listeners in the present study.

One implication of these observations is that talker-specific generalization depends not only on familiarization with the specific acoustic attributes of a talker's voice, but also familiarization with the particular acoustic format in which a speaker is presented. For example, in Experiment 1, subjects became familiar with the sinewave materials during the course of training and were already highly familiar with natural speech. Consequently, they may have been better able to focus attention on the acoustic-phonetic information specific to each speaker during the generalization test phase. However, subjects in Experiment 2 had no prior experience with sinewave utterances. The novelty of the sinewave test items may have diverted their attention away from the voice-specific phonetic information in these highly unnatural sound patterns. In addition, because sinewaves are only an abstract representation of familiar natural speech properties, subjects may have had difficulty perceiving commonalties across the two different perceptual formats. A particular sinewave speaker may have failed to remind listeners of a familiar natural voice simply because the former was not perceived to be perceptually similar to a known speaker category. Experiments designed to determine the role of attention on the acquisition and generalization of speaker knowledge are currently underway in our laboratory.

In summary, the present experiments reveal an asymmetry in the generalization of speaker knowledge from natural and sinewave utterances. Speaker-specific knowledge acquired during training on sinewaves shows generalization to novel sinewave sentences as well as naturally produced utterances, whereas speaker-specific knowledge acquired during training on natural speech does not show generalization to sinewave utterances. The results also showed that variability in the degree of perceptual learning affected generalization of speaker knowledge to novel natural and sinewave sentences. Finally, the experiments showed that perceptual learning of a talker's voice can occur even when specific acoustic

products of vocal articulation are eliminated from the signal, and suggest that attention plays an important role in learning and generalization of speaker-specific knowledge.

References

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

Halle, M. (1985). Speculations about the representation of words in memory. In V.A. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 101-114). New York: Academic Press.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K.R. Scherer & H. Giles (Eds.), *Social Markers in Speech* (pp. 1-32). Cambridge, UK: Cambridge University Press.

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

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. (1995). Talker and task-specific perceptual learning in speech perception. *Proceedings of the XIIIth International Congress of Phonetic Sciences*. Stockholm: Stockholm University, **1**, 194-197.

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

Remez, R.E., Fellowes, J.M., & Rubin, P.E. (In Press). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*.

Rubin, P.E. (1980). *Sinewave Synthesis.* Internal Memorandum, Haskins Laboratories, New Haven CT.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Multimodal Encoding of Speech in Memory: A First Report[1]

David B. Pisoni, Helena M. Saldaña,[2] and Sonya M. Sheffert

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

[1] This work was supported in part by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University–Bloomington.
[2] Now at House Ear Institute, Los Angeles, California.

Multimodal Encoding of Speech in Memory: A First Report

Abstract. Why do people like to watch videos on TV? Why is there now increased interest in video telephones and multi-media technologies that were developed back in the 1960s? Obviously, the availability of new digital technology has played an enormous role in this transition. But, we also believe this is in part due to the same operating principle that encourages listeners in noisy environments to orient toward a talker's face. A multimodal speech signal is extremely robust and informative and provides information that perceivers are able to exploit during perceptual analysis. In this paper, we present results from two experiments that examined performance in immediate memory and serial recall tasks with normal-hearing listeners using unimodal (auditory-only) and multimodal (auditory + visual) presentation. Our findings suggest that the addition of visual information in the stimulus display about the speakers' articulation affects the efficiency of initial encoding operations at the time of perception and also results in more detailed and robust representations of the stimulus events in memory. These results have implications for current theories of speech perception and spoken language processing.

Introduction

Stimulus Variability in Speech Perception

In recent years, we have been studying the problem of stimulus variability in speech perception and spoken word recognition. Our research has shown that the neural mechanisms used in speech perception encode and store fine details of the stimulus input and that many sources of information are not lost or discarded during the perceptual process. These findings have encouraged us to reassess our views about several long-standing theoretical issues in speech perception such as acoustic-phonetic invariance and perceptual normalization, as well as traditional assumptions about the architecture of the mental lexicon (see Pisoni, 1996, for a review).

Mullennix, Pisoni, and Martin (1989) reported that intelligibility for isolated words in noise is influenced by the number of talkers that are presented in the experimental condition. Listeners were presented with lists spoken by a single talker or lists spoken by 15 different talkers. Results showed that identification performance was always better for words presented in the single-talker lists than the multiple-talker lists. These findings suggest that under high talker-variability conditions, listeners engage in some form of on-line "recalibration" each time a new voice is encountered in a set of test trials. These initial results along with other follow-up studies using naming (Mullennix et al., 1989) and speeded classification (Mullennix & Pisoni, 1990) suggest that spoken word recognition is related to the processing of the talker's voice.

Perceptual Learning

Subsequent perceptual learning experiments demonstrated that detailed properties of a talker's voice are encoded into memory and can be used to facilitate word recognition in noise (Nygaard, Sommers, & Pisoni, 1994). These experiments suggest that listeners incidentally encode information about the vocal source attributes when listening to different speakers. It was argued that listeners retained a "procedural memory" for a talker's voice, in addition to specific details about a linguistic event. Thus, the neural

representation of spoken words may encompass both an abstract phonetic description of the utterance as well as detailed information about the structural description of the talker's vocal tract.

Multimodal Speech

Taken together, the results of these lines of research suggest that listeners are tracking and encoding many detailed changes in their perceptual environment. In the present report, we describe the results of two studies that have extended our research on stimulus variability in speech perception to the case of multimodal speech perception. The first study reports results on immediate memory span; the second describes findings on serial recall of lists of isolated words. Both studies compared unimodal and multimodal presentation formats to assess how these sources of information interact and influence the representation of spoken words in memory.

It has been known since the early 1950s that listeners show substantial increases in intelligibility of speech when they attend to the talker's face (Sumby & Pollack, 1954). Other studies have demonstrated that visual articulatory information can override or fuse with an auditory speech signal causing a listener to report hearing a combination of the auditory and visual signal (McGurk & McDonald 1976). Some theorists have proposed that articulatory events can be conveyed to the listener through several sensory modalities and that multiple sources of information are used by the perceptual system to recognize speech (Fowler & Rosenblum, 1991). Based on several findings in the area of multimodal speech perception, Summerfield (1981) argued for a theory of speech processing which takes into account the integration of various sources of information (auditory, visual, tactile). We believe that methodologies utilized by our lab to look at the issue of stimulus variability might prove enlightening in this area. In particular, we are interested in the issue of whether visual information about a speech event is encoded in memory, and, if so, whether this process requires additional resources from the perceptual system.

Previous research has demonstrated an effect of talker variability on immediate memory span (Saldaña, 1995). This study revealed that listeners' resources are taxed by the use of multiple voices in a traditional memory span experiment. These findings have important implications for current conceptions of working memory. Perhaps one of the most influential theories in working memory to date has been proposed by Baddeley and Hitch (1974), who posit an "articulatory loop" as a mechanism for working memory span. According to Baddeley, memory span is constrained by how many items a subject can repeat or that can be "refreshed" by a set of articulatory control processes in approximately a two second duration. Our previous results showed that working memory span is also affected by the amount of information that is contained in each representation. Research on the nature of working memory span is particularly important in light of the role that the concept of working memory plays (either implicitly or explicitly) in almost all current theories of language processing. One goal of our current research is to determine the effect that visual speech information has on working memory.

We are also interested in the effect that visual speech information has in secondary (long-term) memory. Although previous research from our lab has consistently shown that stimulus variability influences perceptual processing, we have also found a benefit in recall due to stimulus variability (Goldinger, Pisoni, & Logan, 1991). This research demonstrated that listeners are actually better at recalling items from multiple-talker lists than single-talker lists, however, this effect is only found in the primacy portion of the list and is only observed when listeners are given sufficient time for rehearsing the word lists. Goldinger et al. (1991) proposed that the addition of multiple voices aids in the elaboration and transfer of information into long term memory. Following from this result, we were interested in the effect of visual speaker information on serial recall.

Immediate Memory Span

Method

Subjects. Twenty-one subjects participated in the experiment as partial fulfillment of a class requirement in Introductory Psychology. Four subjects were discarded from the final analysis for failing to follow instructions. All subjects were native speakers of English with normal hearing and normal or corrected vision. The experiment lasted approximately 45 minutes.

Stimulus Materials. The stimuli consisted of a list of letters spoken by a female actor. The actor was videotaped with a camcorder which was patched into a professional video recorder. A microphone was positioned near the actor's mouth and out of view of the camera. The actor produced 26 letters in a sound-attenuated recording studio. Each item was presented to the actor in random order on a teleprompter and the actor was instructed to look directly into the camera and say the item clearly. The video clip was digitized using a Macintosh Quadra 950 at 30 frames per second. The audio signal was sampled at 22 kHz with 16-bit resolution. Individual clips of each utterance were made by cutting the clip when the mouth was in a neutral position prior to an utterance and then in a neutral position after the utterance.

A preliminary intelligibility test demonstrated that all of the stimuli were intelligible at 100%. A subset of the stimulus items were then chosen for the memory span experiment. The items selected were: B, D, F, H, J, K, L, M, Q, R, Z. The lists were presented in a staircase fashion, with the list increasing or decreasing by one item on each trial. Each subject started with an easy list of four items. The list then increased in length by one item on each trial until subjects were presented with a nine item list. Then the length of the list decreased by one item on each trial until subjects were presented with a four item list. For each half of the experiment, the subject was presented with each list length six times. The first six trials of each part of the experiment served as practice trials and were not included in the final analysis. Two presentation conditions were used: unimodal and multimodal. The presentation of the unimodal and multimodal conditions were blocked and counterbalanced.

Procedure. Subjects were presented auditory items over the headphones. The first item was always a 1000 Hz 500 ms tone accompanied by a visual display that said "get ready." The ready signal was followed by a list of test items, which was then followed by a second tone. The subjects were instructed to wait for the second tone and then repeat the letters that they heard in the exact order that they were presented. The responses were recorded by hand by an experimenter in the next booth. Subjects were told to keep their gaze focused on the monitor throughout the entire experiment.

Results

A list was scored as correct if, and only if, all items were recalled in the proper serial order (Saldaña, 1995). An overall analysis of variance was conducted with Mode (2), and List Length (6), as within-subject variables and order (2), as a between-subjects variable. The analyses revealed an overall effect of Mode $F(1,15)=16.17$, $p<.01$. Subjects' memory span was longer for unimodal presentations compared to multimodal presentations. There was also an overall effect of List Length $F(5,75)=160.23$, $p<.01$. Subjects were better at shorter lists than longer lists. There was no significant effect of order. Post-hoc comparisons revealed that the presentation modality was significant at List Length 6 and 7, $F(1,75)=5.53$, $p<.05$; $F(1,75)=13.881$, $p<.05$, respectively. Unimodal memory span was higher than multimodal span.

Insert Figure 1 about here

Discussion

The results show that the addition of visual information to auditory speech results in a shorter working memory span. This finding is consistent with previous talker variability results which demonstrate that working memory is constrained by the amount and quality of information that is being processed by the system, not just the absolute duration of items. However, in contrast to previous talker variability results, this result can be accounted for within Baddeley's framework of working memory. The theory proposes three mechanisms for working memory: the articulatory loop, the visual spatial sketch pad, and the central executive system. The articulatory loop is responsible for maintaining phonological codes, while the visual spatial sketch pad is responsible for rehearsing spatial location as well as spatial movement. It is assumed that working memory is a limited capacity system, therefore, the ability to process information in the articulatory loop is affected by any simultaneous processing of information in the visual spatial sketch pad. According to this view, it might be expected that the auditory and the visual information are being processed separately at the level of working memory, and are making demands on a common set of resources.

The present experiment was conducted under very favorable signal-to-noise ratios and as a consequence it is an unusual example of audio-visual speech perception. For normal listeners under good listening conditions, visual information is not necessary for recognition. It is possible that our results were not due to the additional processing of visual information but rather a consequence of the distracting quality of the audio-visual condition. One way to investigate this issue is to determine whether visual information has been transferred to long-term memory store. Our next experiment addressed this issue using a serial recall procedure.

Serial Memory

Method

Subjects. Forty subjects participated in the experiment as partial fulfillment of a class requirement in Introductory Psychology. All subjects were native speakers of English with normal hearing and normal or corrected vision.

Stimulus Materials. The stimuli consisted of thirty lists of 10 monosyllabic English words that were taken from the Johns Hopkins Laser Disk Corpus (Bernstein & Eberhardt, 1986). All of the words were produced by a male talker.

Procedure. The subjects were tested individually. The subjects were randomly selected to participate in either the multimodal or the unimodal condition. On each trial, subjects were presented with a list of 10 words with an SOA of 2 seconds. A 1000 Hz 500 ms tone was sounded prior to and following the presentation of each word list. After the second tone, the subjects were instructed to write down the items presented to them. They were allowed to output their responses in any order. However, they were instructed to write each item in the space on their answer sheet which corresponded to the order in which the words were presented.

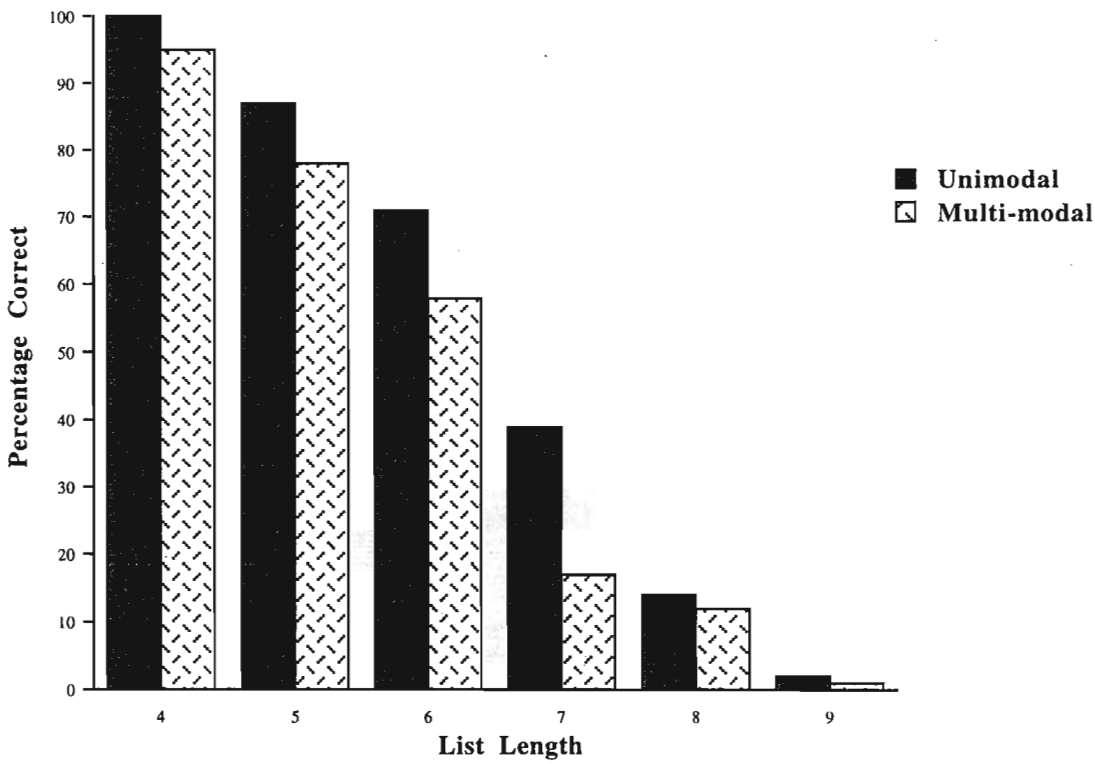


Figure 1. Percentage of correct lists for each list length in Experiment 1.

Results

The items were scored as correct only if they were written down in the space corresponding to the order of presentation. An analysis of variance revealed a main effect of position $F(9,342)=77.71$, $p<.01$, as well as a marginally significant interaction of position and condition $F(9,242)=1.78$, $p=.07$. Post-hoc analyses showed significant differences between the audio and the audio-visual conditions for serial positions 1 and 2 $F(1,38)$ 14.05, $p<.001$, $F(1,38)$ 357.72, $p<.001$, respectively.

Insert Figure 2 about here

Discussion

The present findings are consistent with previous findings from our lab which demonstrated a benefit in recall for multiple-talker lists in the primacy portion of the serial position curve. In earlier papers, we argued that this effect was due to elaborative encoding and transfer of information into a long-term memory store (Goldinger et al., 1991). We believe that the same explanation is appropriate for the present set of results. This conclusion is supported by the finding that the benefit of visual information is only evident in the primacy portion of the list which is usually believed to reflect recall of items from long-term memory. This account is also supported by our previous immediate memory span results, which indicate that the limited capacity working memory system is taxed by the perceptually rich multimodal presentations.

General Discussion

The present set of findings on immediate memory span and serial recall for multimodal presentations add to our earlier results which demonstrate that specific details about the form of the speech signal are processed and encoded by the perceptual system. We suggest that attributes of the talkers face are encoded in working memory and transferred to representations in long-term memory. In addition, the present findings on immediate memory span raise several important questions about the "articulatory loop" hypothesis of working memory, specifically, questions about the nature of information that working memory has access to, as well as the properties and operations of the rehearsal mechanism in working memory. It is now clear that immediate memory span is affected by attributes of both the talker's voice (Saldaña, 1995), and the talker's face. The present results suggest that this information may not be perceptually integrated at the time of initial encoding. Instead the visual information in the talker's face may use resources from the visual-spatial sketch pad which places additional demands on a limited capacity working memory system.

The results obtained in the serial recall experiment under multimodal presentations suggest that the additional visual information about the talkers face is retained and used in subsequent recall. We propose that the additional information facilitates the rehearsal process and/ or the transfer of items to long-term memory. The findings from this experiment are similar to the results obtained several years ago using multiple voices (Goldinger et al., 1991). The presence of additional information about an item, such as the talkers voice, appears to provide the perceptual system with the ability to build a more detailed or robust representation. The presence of additional stimulus dimensions about a talkers voice may aid retrieval mechanisms which use discriminability and distinctiveness to recover items from memory. Apparently, multimodal presentation of speech also helps this elaboration process to work more efficiently. However,

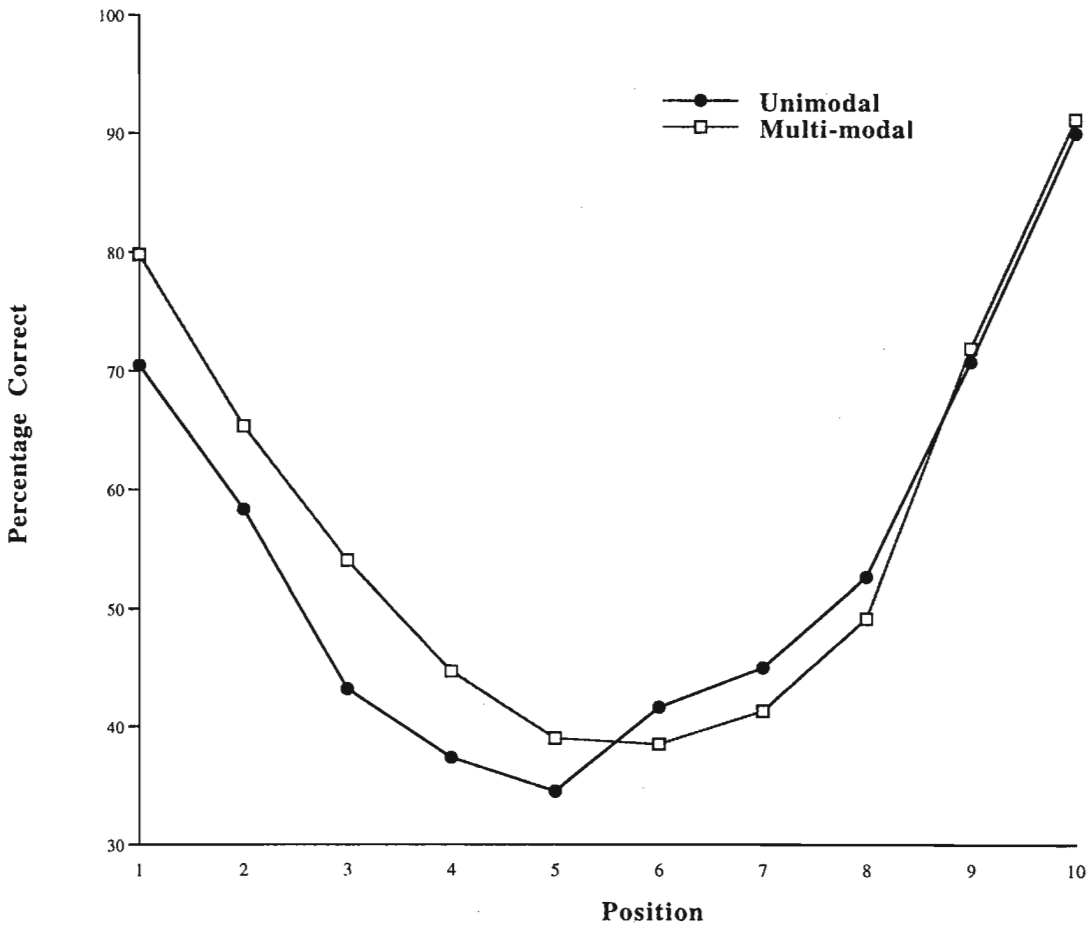


Figure 2. Serial position curve for auditory alone and multimodal conditions in Experiment 2.

this elaboration is very selective in nature, showing up, as in our earlier recall experiments, only in the primary portion of the serial position curve.

References

Baddeley, A.D. & Hitch, G. (1974). Working memory . In G. A. Bower (Ed.), *Recent Advances in Learning and Motivation, Volume 8*. New York: Academic Press.

Bernstein, L.E. & Eberhardt, S.P. (1986). *Johns Hopkins Lipreading Corpus I-II: Disc 1*. Baltimore, MD: Johns Hopkins University.

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

Fowler, C.A. & Rosenblum, L.D. (1991). Perception of the phonetic gesture. In I. G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the Motor Theory* (pp. 33-59). Hillsdale, NJ: Erlbaum.

Johnson, K. & Mullenix, J. (1996). *Talker Variability in Speech Processing*. New York: Academic Press.

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

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.

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

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

Pisoni, D.B. (1996). Some Thoughts on "Normalization" in Speech Perception. In K. Johnson & J.W. Mullennix (Eds.), *Talker Variabiltiy in Speech Processing*. Academic Press: San Diego.

Saldaña, H.M. (1995). The effects of talker-specific information on immediate memory span. Paper presented at the 129th meeting of the Acoustical Society of America, Washington DC, May 30-June 3.

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. (1981). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip Reading* (pp. 3-51). London: Erlbaum.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

The Relationship Between Stimulus Variability, Auditory Memory, and Spoken Word Recognition in Listeners with Hearing Impairment[1]

Karen I. Kirk,[2] David B. Pisoni, David Crotzer,[2] Donald L. Schilson,[2] and Ann E. Kalberer[2]

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

[1]This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University, Bloomington, IN. We thank Michelle Wagner-Escobar, M.A., and Helena Saldaña, Ph.D., for their assistance.
[2]Also Department of Otolaryngology, Indiana University School of Medicine, Indianapolis, IN.

The Relationship Between Stimulus Variability, Auditory Memory, and Spoken Word Recognition in Listeners with Hearing Impairment

Abstract. This study examines the effects of stimulus variability on word recognition and immediate memory performance in normal-hearing and hearing-impaired listeners. Experimental subjects were adult hearing-aid users, while normal-hearing adults served as controls. Word recognition tests included a talker variability test and a speaking rate test. Immediate memory tests included an auditory digit span test and an auditory letter span test. Preliminary results indicate that normal-hearing listeners showed better word recognition and better recall of auditorily-presented signals than did hearing-impaired subjects, indicating a close relationship between phonetic coding and immediate memory.

Introduction

Speech perception is a complex process involving perceptual analysis and encoding of sensory information, retrieval of previously stored information from memory, and the interpretation, integration and assimilation of various knowledge sources. In listeners with normal hearing, both perceptual analysis and word retrieval are influenced by the presence of stimulus variability. For example, varying the talker's voice from trial-to-trial produces a decrease in both spoken word recognition scores and serial word recall performance when compared to a single-talker condition, presumably because the perceptual normalization process consumes common processing resources. This study examined the effects of stimulus variability on word recognition and immediate memory performance in normal-hearing and hearing-impaired listeners.

Purpose

The goals of this project were:

1) To compare the immediate memory spans of listeners with normal and impaired hearing.
2) To examine the effects of talker variability on immediate memory in listeners with normal or impaired hearing, and
3) To examine the relationship between immediate memory and word recognition performance in listeners with normal or impaired hearing.

Methods

Subjects

All subjects were between the ages of 18-65 years. Experimental subjects were selected from a group of hearing aid users seen at Indiana University Medical Center. Selection criteria included a bilateral mild-to-moderate hearing loss with speech discrimination scores of 75% or greater on the NU-6 (Tillman & Carhart, 1966), and a minimum of three months experience with their current hearing aid. Four males and three females participated as experimental subjects. Six adults with normal-hearing (i.e., $\leq$20 dB HL at octave frequencies from 250-4000 Hz) served as control subjects. They were recruited from students and staff at the Indiana University Medical Center.

Table 1

Subject Characteristics

		Age (yrs)	Auditory Thresholds (dB HL)				NU-6
			500 Hz	1000 Hz	2000 Hz	4000 Hz	
Hearing-Impaired Listeners	Mean	55.3	32.1	36.4	40	46.4	96.9%
	Range	40-64	25-45	20-60	30-55	35-65	88-100%
	SD	8.3	7.0	16.3	10.0	11.1	4.9%
Normal Hearing Listeners	Mean	39.7	5	-1.3	2.5	5	100%
	Range	29-51	-10-15	-10-5	0-10	0-10	-
	SD	11.8	12.3	6.3	5.0	4.1	-

Stimulus Materials

Word Recognition Tests. All stimulus materials were selected from computerized databases maintained at Indiana University Speech Research Laboratory. These databases contain tokens of spoken words, letters, and digits produced by multiple talkers, both male and female, that have been digitized and stored as individual computer files.

Talker Variability Test - From a digital database containing 6000 words (300 words from the Modified Rhyme Test [House et al., 1965] recorded by 10 male and 10 female talkers), 100 words were chosen based on computational analyses of their lexical properties. Half were "easy" (i.e., they occurred often in English, and contained few phonetically-similar words with which they could be confused), and half were "hard" (i.e., they occurred infrequently and were phonetically similar to many other words). The single-talker condition had 25 "easy" and 25 "hard" words produced by one male talker. The multiple-talker condition was similar except that the talker varied randomly from trial-to-trial. These materials were based on a previous study by Mullennix et al. (1989), using normal-hearing listeners.

Speaking Rate Test - Two hundred words were selected from a recorded database of 3000 PB words. The single-speaking-rate condition contained 50 words spoken by a single male talker at a medium speaking rate (mean word duration = 533 ms). The mixed-speaking-rate condition contained 150 words produced by the same male talker, but speaking rate varied from trial-to-trial, and included a total of 50 fast, 50 medium, and 50 slow tokens. The average duration for these tokens was 375 ms, 533 ms, and 905 ms, respectively. These stimuli were developed by Sommers et al. (1994).

Immediate Memory Tests. Two immediate memory tasks, an Auditory Digit Span test and an Auditory Letter Span test, were presented under two conditions: a single-talker condition in which one male talker produced all the items in a list, and a multiple-talker condition where a novel talker produced each item in a list. Table 2 summarizes the Digit Span and Letter Span tests. Nine different list lengths were constructed for each test. The lists ranged from two-to-10 items each and were drawn successively without replacement.

Table 2

Immediate Memory Span Stimulus Materials

Talker Condition	Letters (H B J K M R Q D F Z L)	Digits (0 1 2 3 4 5 6 7 8 9 10)
Single	45 lists	45 lists
Multiple	45 lists	45 lists

Procedures

All subjects were tested inside a sound-attenuated booth while seated at a table facing a loudspeaker. Stimuli were presented via free field at approximately 72 dB SPL. Hearing-impaired listeners used their hearing aids during testing. Stimulus presentation was blocked by test, but test order and the order of conditions within a test were randomized.

Word Recognition Tests

Subjects responded by repeating the word they heard and their responses were transcribed by the examiner. Only responses that exactly matched the target item were counted as correct (e.g., a plural response to a singular target was scored as an incorrect response).

Immediate Memory Span Tests

Prior to each memory test, subjects completed an identification task for the stimuli within that set. They were first given a written sheet containing the set of letters or digits from which the items could be drawn, and then asked to identify each item presented in isolation. In the single-talker conditions, each letter or digit was randomly presented three times. Because there were 121 stimuli to identify in the multiple-talker conditions, (a maximum of 11 list items X 11 different talkers), subjects heard each token only once. During each memory test, 45 lists were presented in a staircase fashion, increasing from two items to 10 items and then decreasing back from 10 items to two items. Following the presentation of each list, subjects responded by writing their responses on sheets of paper containing a blank space for each list item. Lists were scored as correct if and only if, all the items were recalled in their correct temporal order.

Results

Word Recognition Tests

A summary of the word recognition performance for the two groups of listeners is shown in Table 3 broken down by the four conditions in each listening test. Across the four talker-variability conditions, word recognition scores were consistently higher for the normal-hearing listeners than the hearing-impaired listeners ($p \leq .05$). This was true for both the lexically "easy" and "hard" words, and the single-talker and multiple-talker conditions. As expected, both groups had better word recognition performance in the single-talker than the multiple-talker conditions ($p \leq .04$). Easy words were identified with greater accuracy than the "hard" words, although the differences were significant only in the multiple-talker conditions ($p \leq .02$).

Normal-hearing listeners always identified words from the single-talker conditions better than the hearing-impaired listeners. These differences in performance between normal-hearing and hearing-impaired listeners increased even more in the multiple-talker conditions, which had increased stimulus variability.

Finally, the same pattern of results was observed for the speaking-rate tests. The normal-hearing listeners' word recognition performance was consistently better than that of the hearing impaired in all conditions ($p \leq .03$), except for slow words in the mixed-speaking-rate condition. The difference in performance between the two groups was largest at the fast speaking rate ($p \leq .008$).

Table 3

Mean percent of words correctly identified on the Talker test and the Speaking Rate tests for both subject groups.

		Talker Test				Speaking Rate Test			
		Single-Talker		Multiple-Talker		Multiple-Rate			
		Easy	Hard	Easy	Hard	Single	Fast	Med.	Slow
Normal-Hearing	Mean	85.6	86.0	79.8	70.0	82.7	71.2	83.7	85.8
	(SD)	(7.7)	(7.4)	(10.6)	(9.7)	(9.1)	(4.6)	(7.0)	(8.3)
Hearing-Impaired	Mean	65.2	60.1	50.9	37.1	64.5	50.0	63.7	70.5
	(SD)	(13.9)	(16.5)	(10.2)	(14.4)	(13.5)	(16.9)	(17.4)	(17.4)

Immediate Memory Span Tests

Figures 1 and 2 illustrate the percent of lists correctly recalled on the Digit Span Test by the two subject groups for the single-talker and multiple-talker conditions, respectively. Figures 3 and 4 present the results for the Letter Span test. A comparison of average memory spans (i.e., the maximum list length at which all five lists were correctly recalled) for the two groups is shown in Table 4.

The results of the Immediate Memory Span tests showed that both groups of listeners demonstrated reduced digit and letter recall as list length increased. In addition, the average memory span consistently was longer for the normal-hearing subjects than for the hearing-impaired subjects for all conditions except the multiple-talker condition on the Letter Span test, but these differences were significant only for letter recall in the single-talker condition ($p \leq .02$). Generally, introducing multiple talkers had little effect on either digit or letter recall. However, the hearing-impaired subjects had significantly longer memory spans for letter recall in the multiple-talker condition than in the single-talker condition ($p \leq .03$).

Insert Figures 1, 2, 3 and 4 about here.

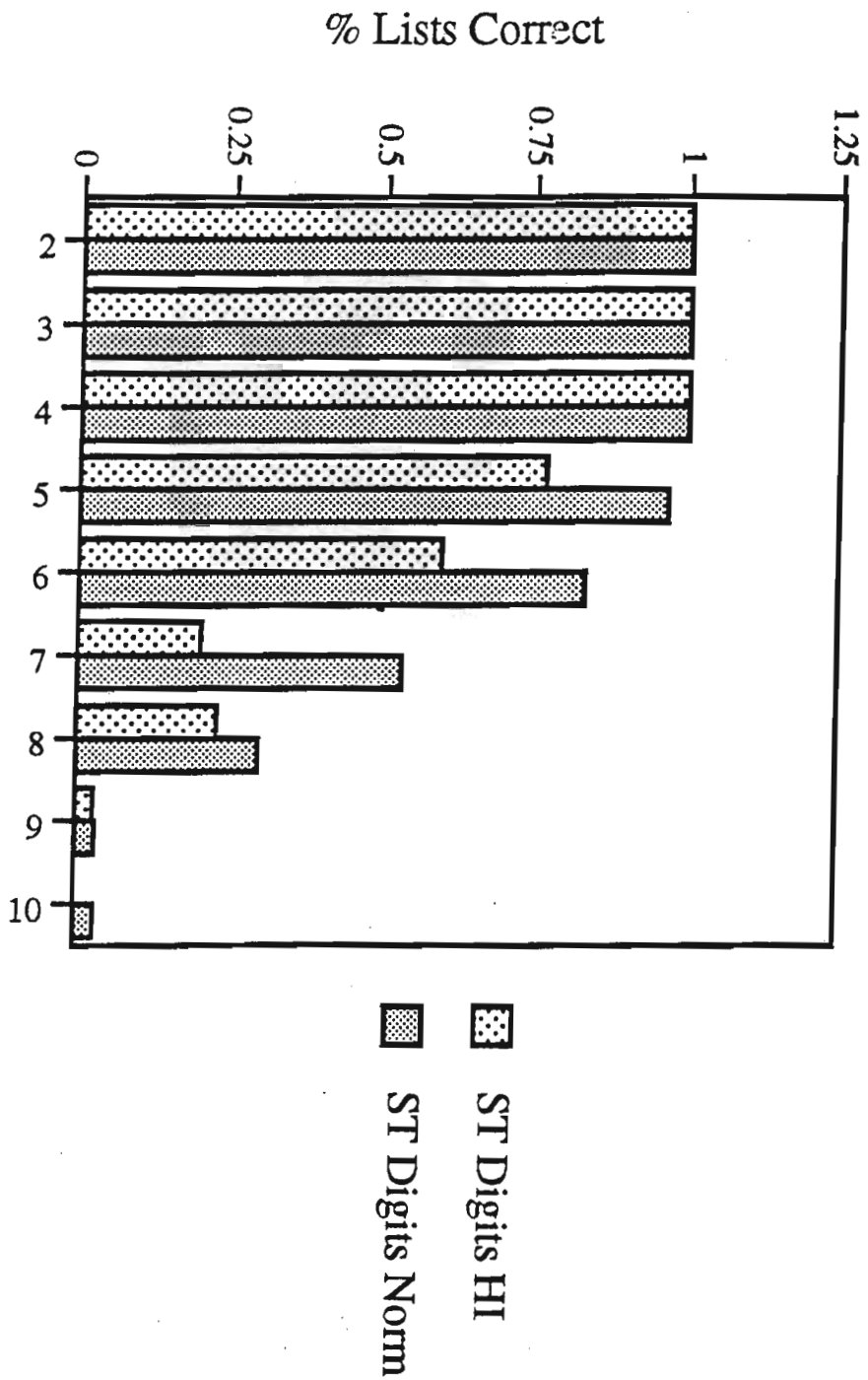


Figure 1. Performance of normal-hearing and hearing-impaired listeners in the single-talker digits condition.

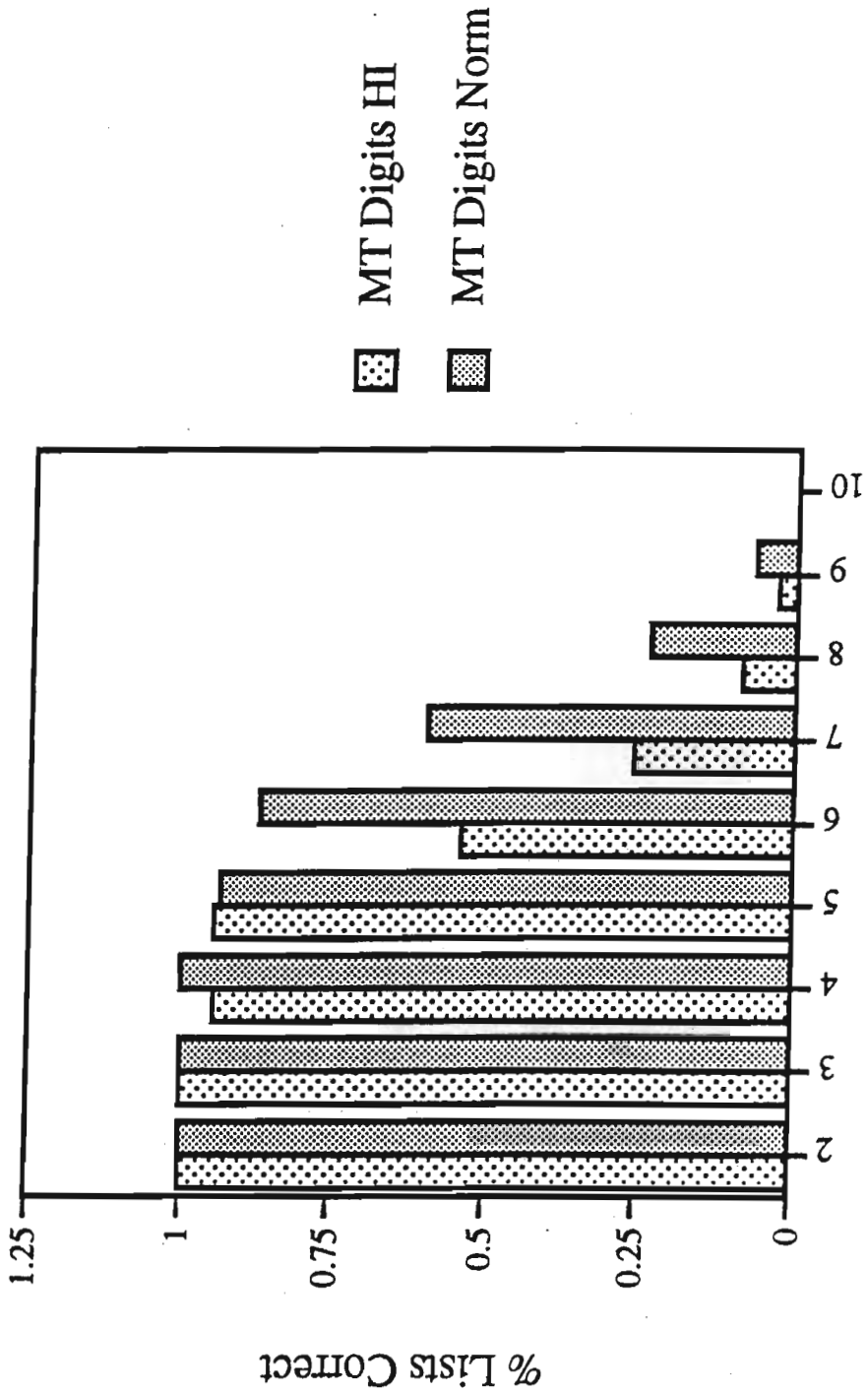


Figure 2. Performance of normal-hearing and hearing-impaired listeners in the multiple-talker digits condition.

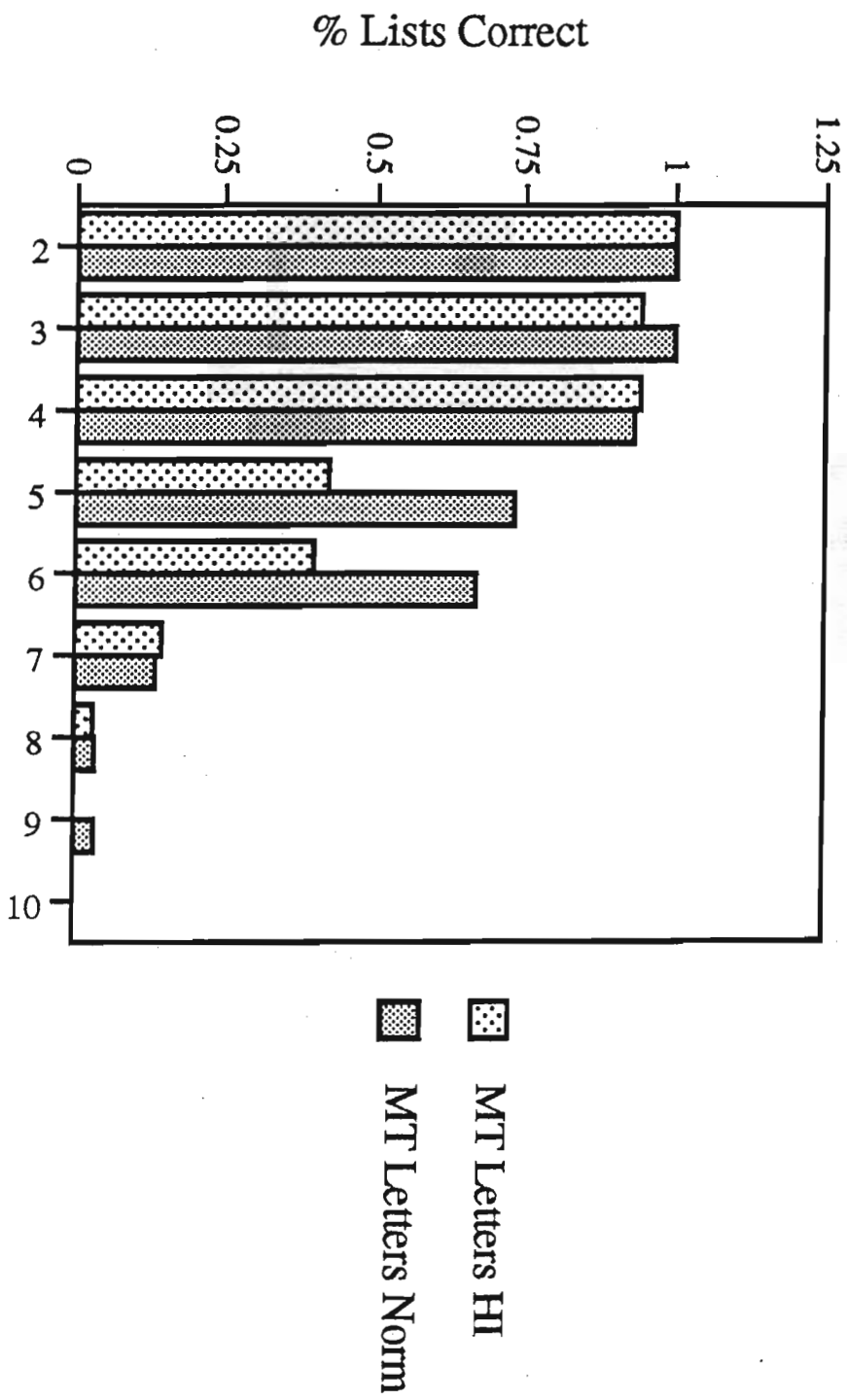


Figure 3. Performance of normal-hearing and hearing-impaired listeners in the single-talker letters condition.

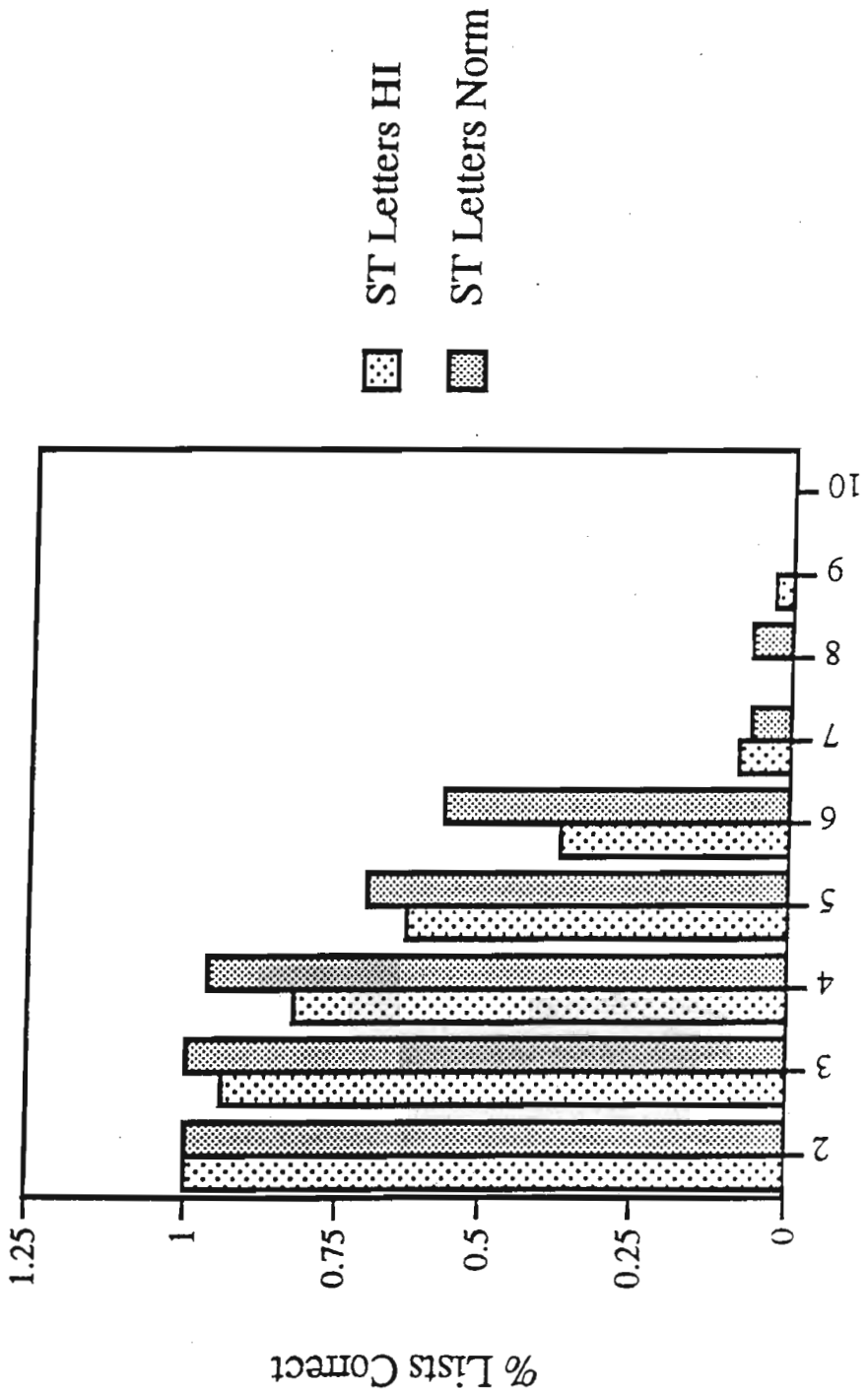


Figure 4. Performance of normal-hearing and hearing-impaired listeners in the multiple-talker letters condition.

Table 4

Average immediate memory span.

		Digit Span Test		Letter Span Test	
		Single-Talker	Multiple-Talker	Single-Talker	Multiple-Talker
Normal-Hearing	Mean	5.5	5.7	4.5*	3.7
	(SD)	(1.1)	(1.0)	(1.1)	(0.5)
Hearing-Impaired	Mean	4.4	4.9	3.0*	3.7
	(SD)	(0.8)	(1.1)	(0.6)	(0.5)

*$p \leq .02$

Correlations Between Word Recognition Scores and Immediate Memory Span

Tables 5 and 6 present the correlations for the normal-hearing and hearing-impaired listeners, respectively. Word recognition performance was significantly correlated ($p \leq .05$) only with memory span for letters in the multiple-talker condition, but the direction of the correlation differed for the two subject groups.

Normal-Hearing Subjects - Significant negative correlations were found between memory span for letters in the multiple-talker condition and performance on at least one condition in the Talker test and the Speaking-Rate test. Significant correlations were found with performance on the "easy" words in the single-talker condition ($r = -.88$), the single-speaking-rate condition ($r = -.85$), and the slow words in the multiple-speaking-rate condition ($r = -.90$).

Hearing-Impaired Subjects - Significant positive correlations were found between memory span for letters in the multiple-talker condition and performance on several conditions in the Talker test and the Speaking-Rate test. Significant correlations were found with performance on the "hard" words in the single-talker condition ($r = +.83$) and the multiple-talker condition ($r = +.76$). Significant correlations were also found with performance on the single-speaking-rate condition ($r = +.96$), the slow words in the mixed-speaking-rate condition ($r = +.90$), and the fast words in the mixed-speaking-rate condition ($r = +.91$).

Table 5

Correlations between word recognition scores and performance on the immediate memory span tests for normal-hearing listeners (N=6).

		Normal-Hearing Listeners			
		Digits		Letters	
		Single-Talker	Multiple-Talker	Single-Talker	Multiple-Talker
NU-6		--	--	--	--
Single-Talker	Easy	-.02	-.04	-.66	-.88*
	Hard	.23	.31	-.42	-.46
Multiple-Talker	Easy	.20	-.02	-.41	-.57
	Hard	.27	.24	-.04	-.10
Single Rate		.15	.15	-.64	-.85*
Mixed Speaking Rate	Slow	.18	-.09	-/38	-.90*
	Med.	.55	.30	-.09	-.70
	Fast	.09	-.10	-.19	-.60

*$p \leq .05$

Table 6

Correlations between word recognition scores and performance on the immediate memory span tests for hearing-impaired listeners (N=7).

		Hearing-Impaired Listeners			
		Digits		Letters	
		Single-Talker	Multiple-Talker	Single-Talker	Multiple-Talker
NU-6		.52	.34	-.18	.21
Single-Talker	Easy	.67	.59	-.04	.55
	Hard	.35	.27	.27	.83*
Multiple-Talker	Easy	.53	.47	.27	.65
	Hard	.46	.31	.16	.76*
Single Rate		.33	.39	.13	.96**
Mixed Speaking Rate	Slow	.05	.04	.35	.90**
	Medium	.06	-.13	.53	.74
	Fast	.29	.20	.20	.91**

*$p \leq .05$
**$p \leq .01$

Discussion

The present results should be considered preliminary and suggestive because of the small sample size. However, several interesting trends emerged. First, it appears that normal-hearing listeners displayed better word recognition *and* better recall of auditorily-presented signals than do hearing-impaired subjects. One possibility for this difference is that additional processing resources are required to analyze the degraded speech signal received by listeners with hearing impairment, and to match these signals to sound patterns stored in their mental lexicons. Secondly, the results showed that introducing stimulus variability yielded decreases in word recognition performance, but not in immediate memory span for letters or digits. The word recognition data replicate the earlier results of Kirk, Pisoni, & Miyamoto (1995). The only significant effect of introducing stimulus variability on immediate memory was an unexpected *increase* in memory span for letters for the hearing-impaired subjects. This finding was somewhat surprising. The reasons for this outcome are not clear and the findings need to be replicated in another group of hearing-impaired listeners. It is very possible that the items on the multiple-talker, letter-span lists were more perceptually discriminable than those in the single-talker, letter-span lists, and therefore could be encoded in such a way as to preserve their distinctiveness in memory. This would reduce the confusability among perceptually similar pairs of letters. This could be tested by selecting confusable and non-confusable letters in subsequent tests.

Only immediate memory span for letters was significantly correlated with word recognition performance. The letter recall task was more difficult than the digit recall, because the former requires the listener to make finer phonetic distinctions among confusable items in the list than the latter. It was not uncommon for hearing-impaired listeners to correctly identify phonetically-similar letters (e.g., /b/ and /d/) in isolation but then to confuse them once they appeared in a list in the memory span for letters task. Thus, hearing-impaired subjects who were best at identifying lexically "hard" words (i.e., those that have many phonetically similar words with which they can be confused) were also the same listeners who were best at letter recall, and showed increased memory spans.

Conclusions

These preliminary results looking at both perception and memory performance suggest a close relationship between phonetic coding and immediate memory. They also imply that hearing impairment may impact on cognitive tasks requiring access to short-term working memory. Further study with a larger subject population of both normal-hearing and hearing-impaired subjects is warranted to generalize these findings to other processing activities used in spoken language comprehension. If we assumed that spoken language processing makes use of a "limited-capacity" processing system requiring the use of short-term working memory, then impairments in perceptual analysis may propagate up the system and have an impact not only on speech perception performance, but on activities that make use of these neural/cognitive representations as well.

References

House, A.S., Williams, C.E., Hecker, M.H.L, & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response test. *Journal of the Acoustical Society of America*, **3**, 158-166.

Kirk, K.I., Pisoni, D.B., & Miyamoto, R.C. (1995). Effects of stimulus variability on speech perception in hearing impaired-listeners. Manuscript under revision.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Perception & Psychophysics*, **47**, 379-390.

Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-1324,

Tillman, T.W. & Carhart, R. (1966). An expanded test for speech discrimination utilizing CNC monosyllabic words: *Northwestern University Auditory Test No. 6. Technical Report No. SAM-TR66-55*. USAF School of Aerospace Medicine, Brooks Air Force Base, TX.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

The "Easy-Hard" Word Multi-Talker Speech Database: An Initial Report[1]

Gina M. Torretta

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

[1] This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC-00012 to Indiana University-Bloomington. I gratefully acknowledge the assistance of Benedetta Mariotti and Melissa Kluck in preparing the stimulus materials, and Bill Svec for assistance with program and perceptual training system (PTS) development. I am indebted to Ann Bradlow for her generosity and patience in guiding data analyses and to Mitch Sommers for his input on data collection.

The "Easy-Hard" Word Multi-Talker Speech Database: An Initial Report

Abstract. One research strategy used in our laboratory has focused on the development of large speech databases as general purpose resources for use in various experiments. The speech database described here incorporates several variables known to affect spoken word recognition. The database consists of 4500 digital speech files. Five male and five female talkers each produced a list of 150 words at three different speaking rates. The basic word list contains 75 lexically "easy" and 75 lexically "hard" words as defined by the Neighborhood Activation Model (NAM) of Luce (1986). This report provides a description of the database as well as some intelligibility data for the medium and fast rate tokens.

Introduction

The development of large speech databases is one research strategy used in our laboratory. The goal of the present project was to develop a carefully constructed speech database that can be accessible in CD-ROM format for general use. There are three phases in the development of the database. Phase one is stimulus preparation, including word list formulation, speech digitization, and speech file preparation. Phase two is the collection of intelligibility data, and statistical analyses of listener responses to the utterances. Phase three will involve acoustic-phonetic analyses of a subset of the digital speech files. Currently, the project is in phase two. In this initial report, both phase one and phase two are described in detail, as well as the intended progression of phase three.

The Easy-Hard Word Database consists of 4500 digital speech files. The heart of the database is a set of 150 English words that vary in both word frequency and lexical similarity. The set of words was recorded from ten talkers, five males and five females, at three different speaking rates (150 words x 3 rates x 10 talkers = 4500 tokens). The database is intended to provide a set of speech materials for experimentation in areas of speech perception and spoken word recognition, as well as acoustic-phonetic analyses.

Phase 1: Stimulus Preparation

Previous work in our laboratory has shown that lexical characteristics such as word familiarity, lexical frequency and lexical similarity have behavioral consequences in a variety of spoken word recognition tasks (Luce, 1986; Luce et al., 1990; Kirk et al., 1995; Sommers et al., 1995). *Word familiarity* is a subjective rating of familiarity, or perceived commonness of words (Nusbaum et al., 1984). *Lexical frequency* is defined as the average number of times a word occurs in printed text (Kucera & Francis, 1967). *Neighborhood density* is the number of "neighbors" or words that differ by one phoneme from the target word. An example of a lexical neighborhood for the target word "pat" are words such as "cat, bat, pit, pet, spat and pats." The word "pat" has more neighbors than, for example, the word "phone." The lexical neighborhood for "pat" would therefore be considered more "dense" than the lexical neighborhood for the word "phone." Figure 1 shows the relationships among these three stimulus variables.

Insert Figure 1 about here

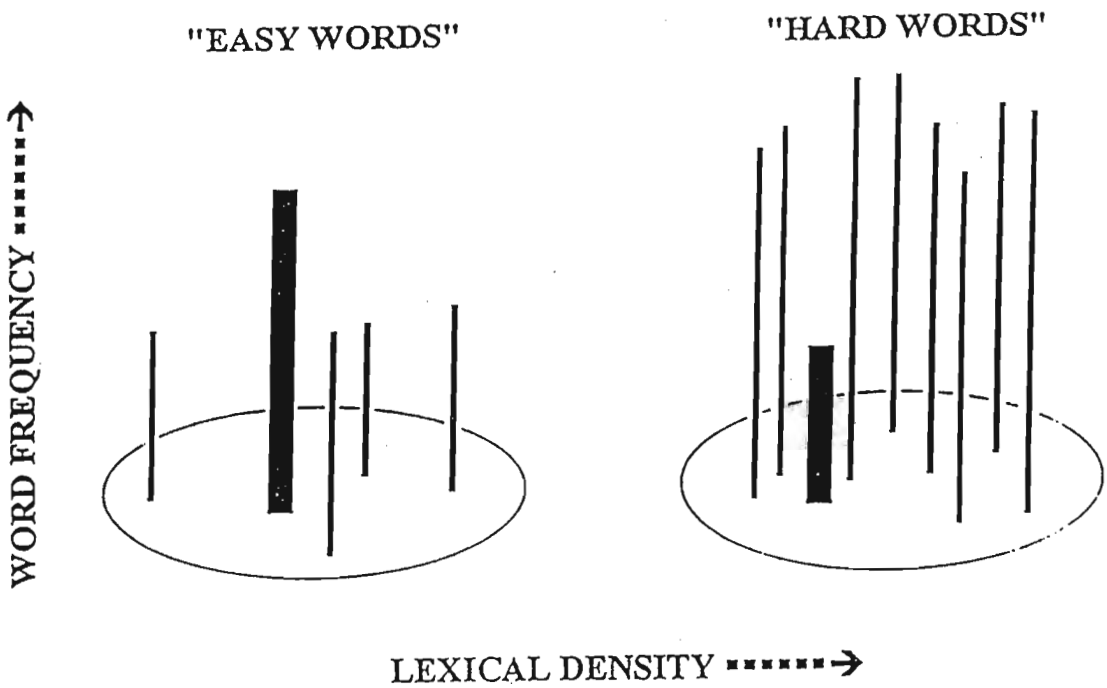


Figure 1: The Neighborhood Activation Model (Luce, 1986): Word frequency, familiarity and lexical density are characteristics of "lexical neighborhoods" that influence spoken word recognition.

In this figure, each bar represents a target word, and the lexical neighborhoods are enclosed by the ovals. Word familiarity is represented by bar width, word frequency is represented by bar height, and neighborhood density is represented by the number of bars within an oval. The Neighborhood Activation Model (NAM) of Luce (1986) provides a computational analysis of the relationship between these factors. According to NAM, the number of nearest neighbors can inhibit or enhance the correct selection of a target word from lexical memory. Lexically "easy" words come from "sparse" similarity neighborhoods, whereas lexically "hard" words come from "dense" similarity neighborhoods. Furthermore, "easy" words have a higher frequency than the mean frequency of other words in the neighborhood, whereas "hard" words have lower frequencies than the mean neighborhood frequency. In other words, "easy" words are perceptually distinctive and "stick out" from their similarity neighborhoods, whereas, "hard" words are "swamped" by their neighbors.

The word list for this database was constructed using a computer-readable version of Webster's Pocket Dictionary (Pisoni et al., 1985; Nusbaum et al., 1984). Using an in-house lexical search program (SRLEX), monosyllabic CVC words with high familiarity ratings (>6.7 on a 7 point scale, where 1 indicated the lowest and 7 indicated the highest degree of familiarity) were selected. Lexical frequency and neighborhood density varied in the list. The 75 "easiest" words and the 75 "hardest" words were then selected from the list. The 75 "easy" words are high frequency words (mean=310 occurrences per million) with relatively few neighbors (mean=14). These are words from "sparse" lexical neighborhoods. In contrast, the 75 "hard" words have relatively low frequency (mean=12 words per million), and more phonetically similar neighbors (mean=27). These words come from "dense" similarity neighborhoods. The mean word frequency for the "easy" words is higher than the mean neighborhood frequency (309.7 versus 38.3 occurrences per million), whereas the opposite is true for the "hard" words (12.2 versus 282.2 occurrences per million for the word frequency and neighborhood frequency, respectively). Table 1 gives the means for lexical frequency, familiarity, neighborhood density and neighborhood frequency for the subsets of 75 "easy" and 75 "hard" words.

The next step was to make audio recordings of this set of words spoken by various talkers. The talkers sat in a sound-attenuated IAC booth in front of a microphone and read a randomized word list displayed on a CRT monitor. Talkers were instructed to speak in a normal speaking voice, varying only the rate of speech as appropriate for each of the three different speaking rate conditions.

All speech samples were low-pass filtered at 10kHz and digitally sampled at 20kHz with 16-bit resolution using a DSC Model 240 analog-to-digital converter interfaced to a VAX station 3500. After digital sampling, the speech files were edited using a digitally controlled waveform editor to remove the silent portions on either side of the word and to visually check the waveform to ensure speech sample integrity. Rerecordings were performed in appropriate cases (e.g., speaking level too loud/ too soft) and retained for the final version of the speech database.

After segmentation, files were prepared for presentation in our newly implemented PTS lab (Hernandez, 1995). Overall RMS (root-mean-square) amplitude levels for each speech file were digitally equated to ensure equal presentation levels. Following leveling, files were converted to WAV format for presentation in the PC laboratory.

Table 1

Means of lexical characteristics of words in the Easy-Hard Word Database.

	Easy	Hard
Frequency	309.7	12.2
Familiarity	7.0	6.8
Density	13.5	26.6
Neighborhood Frequency	38.3	282.2

Phase 2: Intelligibility Assessment

The purpose of this new database is to provide researchers with a large sample of carefully selected spoken words by multiple talkers produced at several speaking rates. In addition to having the database of speech tokens, intelligibility data were also collected. The results of the intelligibility tests for both the medium and fast rate of speech are reported here as well.

Subjects

Two-hundred subjects from Indiana University participated in intelligibility data collection. All subjects received course credit for their participation in the experiment. All subjects had normal hearing and reported no history of speech or hearing disorders at the time of testing.

Procedure

An intelligibility testing program developed for the PTS system (Hernandez, 1995) was used to present the stimuli to subjects and record their responses. For both the medium and fast speaking rates, the 150 words from each of the 10 talkers were presented binaurally in the clear over matched and calibrated DT-100 Beyerdynamic headphones to 10 listeners at a comfortable listening level (75 dB/SPL). For each data collection session, a listener heard words spoken by only one talker, and no two listeners heard the same talker during the same session. All listeners received a set of typed instructions explaining the task. Listeners were informed that they would hear a list of English words, and, following each word, they were required to type the word they heard into the computer. All listeners were informed by the experimenter that there would be ample time to check any possible spelling mistakes and make corrections before proceeding with list presentation. Questions and instruction clarifications were handled by the experimenter prior to each data collection session. Following data collection, subjects were debriefed about the nature of the experiment they participated in, and any questions were answered.

Data Analysis

All output data files were processed twice for error tabulations. The intelligibility data program matched the typed input of each listener to a template list of correct responses to provide an initial correct response assessment. This first pass over listener responses was intended to mark as correct all typed input that exactly matched the template list. Then each response marked as incorrect was examined individually by hand in order to accept or reject "incorrect" marked responses on a case-by-case basis. This second pass was intended to catch any homonyms or misspellings that still formed a possible correct response (e.g., 'mit' for 'mitt'), or miscellaneous responses and misspellings which were identifiable as the intended correct response (e.g., 'wrko' for 'work'; 'lvoe' for 'love'). All errors which were unidentifiable as either misspellings, typing errors, or incorrect responses remained marked as incorrect.

Database tables were compiled to represent the data for each speech rate in several ways. First, responses for all listeners were examined to assess overall performance. Following these analyses, the data were examined by the categories "easy" and "hard" in order to assess lexical as well as talker effects. Second, responses for each random presentation order were compiled so performance could be assessed over the course of the experiment to assess learning effects. Finally, an error database was created in order to compare errors across talkers and listeners.

Results

Overall intelligibility (in terms of percent correct transcription) was scored separately for each speaking rate and for each talker. Figure 2 shows each talker's mean intelligibility score across all 150 words and across all ten listeners for both speaking rates.

Insert Figure 2 about here

The mean intelligibility score for the medium rate was 92% correct with a range of 86.7% to 96.9% correct; the mean score for the fast rate was 88% with a range of 79.2% to 93.7% correct. A repeated measures ANOVA, with listener response as the repeated measure, indicated a main effect of talker for both the medium and fast rates ($F(9,148)=7.41$, $p<.0001$, and $F(9,148)=12.81$, $p<.0001$, respectively) across all 150 words. This pattern indicates significant differences in overall intelligibility across the ten talkers for both speaking rates.

Figure 3 shows the effect of lexical category ("easy" vs. "hard") on intelligibility scores for each talker at the medium (Figure 3a) and fast (Figure 3b) speaking rates. A repeated measures ANOVA, with listener response as the repeated measure, showed a main effect of lexical category for both rates ($F(1,148)=13.19$, $p=.0003$, and $F(1,148)=13.96$, $p=.0004$, for medium and fast rates respectively). There were also significant interactions between talker and lexical category ($F(9,1332)=3.22$, $p=.0007$, and $F(9,1332)=2.33$, $p=.013$, for medium and fast rates respectively). The listeners' responses across words were also assessed. A repeated measures ANOVA showed that individual listener performance did not vary significantly across words for either speaking rate.

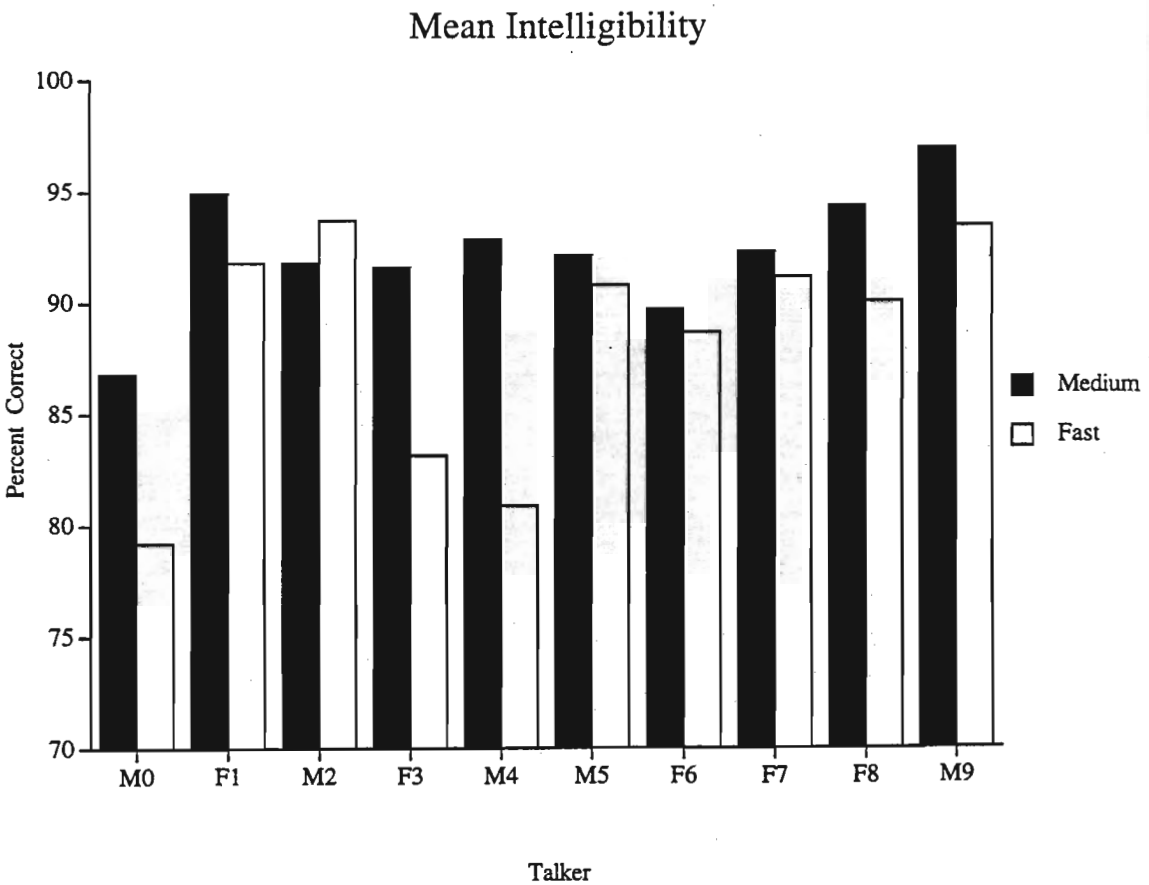


Figure 2: The effect of speaking rate on mean intelligibility scores.

Insert Figure 3 about here

Changes in listener performance over the course of the experiment were also assessed for each speaking rate by comparing intelligibility scores for the first and last quartile (n=38) of the randomly ordered word list presented in each data collection session. Figure 4 shows percent correct transcription scores for each quartile for each talker at the medium (Figure 4a) and fast (Figure 4b) speaking rates. A repeated measures ANOVA, with listener response as the repeated measure, showed a main effect of quartile ($F(1,90)=18.61$, $p<.0000$, and $F(1,90)=28.08$, $p<.0000$, for medium and fast rates respectively). Performance was consistently better in the last quartile than the first, and this effect was independent of the specific test item presentation orders, which differed from listener to listener.

Insert Figure 4 about here

Taken together, the results shown in Figures 3 and 4 indicate that "easy" words were more accurately recognized than "hard" words (shown in Figure 3), and that listener performance improved from the first to fourth quartile of trials presented during a test session (shown in Figure 4). In addition to these main effects, there was also an interaction between quartile and lexical category for both medium and fast speaking rates ($F(1,90)=5.53$, $p<.0209$, and $F(1,90)=5.64$, $p<.0196$, respectively). Figure 5 shows the overall intelligibility scores for the "easy" and "hard" words in the first and fourth quartiles of the randomly ordered word lists at the medium (shown in Figure 5a) and fast (shown in Figure 5b) speaking rates. The graph demonstrates that the transcription scores improved more for the lexically "hard" words than for the lexically "easy" words.

Insert Figure 5 about here

Additionally, difference scores between the first and fourth quartile for "easy" and "hard" words for each listener (fourth quartile minus first quartile percent correct transcription scores) were compared. A paired t-test on these difference scores showed a larger difference for the "hard" words than for the "easy" words. This pattern of results held for both the medium speaking rate ($t(99) = -2.41$, $p<.0178$) and the fast speaking rate ($t(99) = -2.37$, $p=.0198$). Figure 6 displays these difference scores, showing the interaction between lexical characteristics and presentation order for both medium and fast speaking rates.

Insert Figure 6 about here

In summary, results of the intelligibility tests revealed three general patterns. First, "easy" words were more accurately recognized than "hard" words. This was expected based on previous research. Second, listeners' word recognition accuracy improved from the first to the fourth quartile, demonstrating listener adaptation or "tuning" to the specific characteristics of an individual talker. Finally, the adaptation effect to the talker's voice was larger for lexically "hard" words than for lexically "easy" words, indicating an interaction between the process of lexical discrimination and perceptual learning of a talker's voice

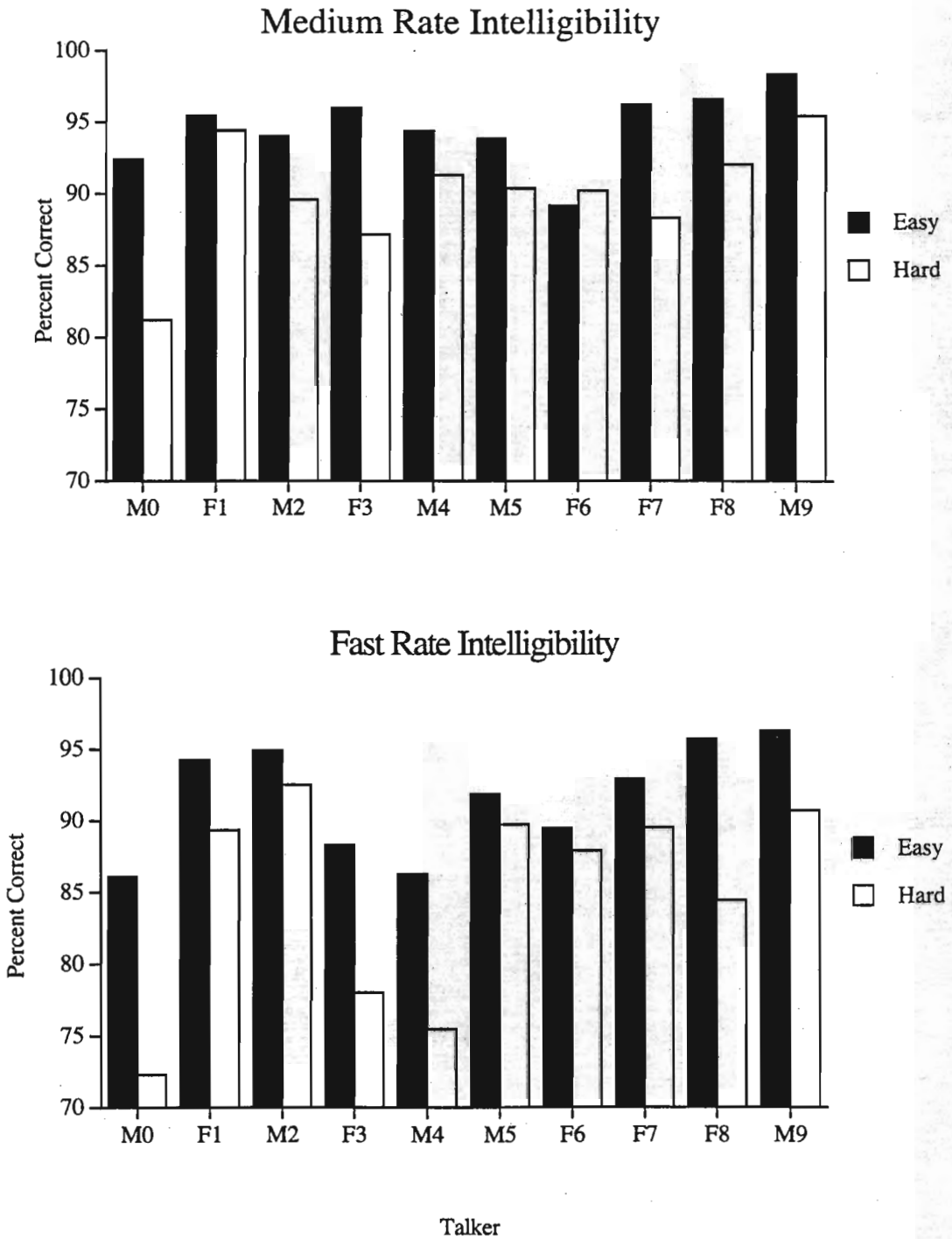


Figure 3: The effect of lexical characteristics on intelligibility scores for medium (top) and fast (bottom) rates of speech.

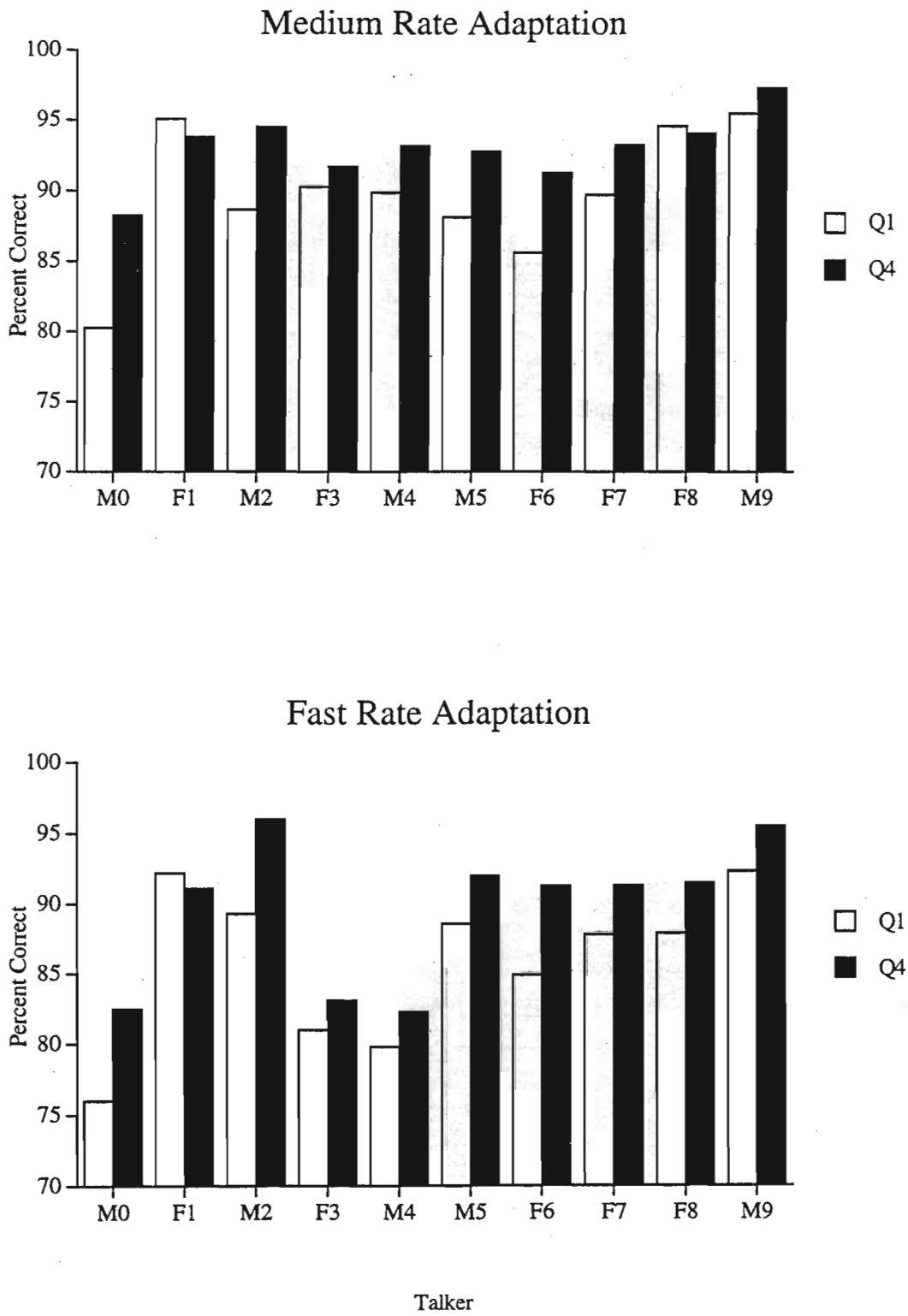


Figure 4: The effect of adaptation from first to fourth quartile of the word lists for medium (top) and fast (bottom) rates of speech.

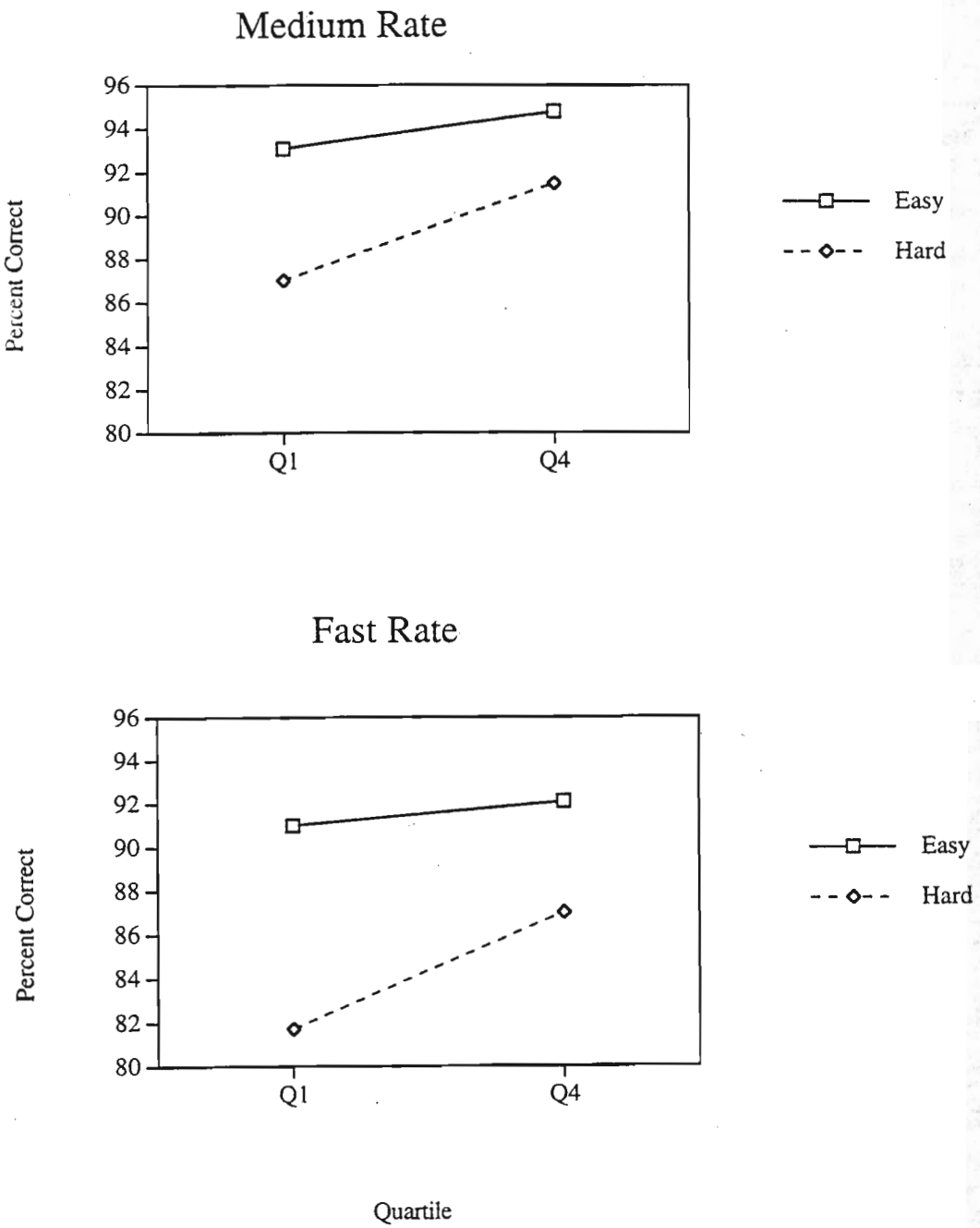


Figure 5: The interaction between lexical characteristics and quartile of the word lists for medium (top) and fast (bottom) rates of speech.

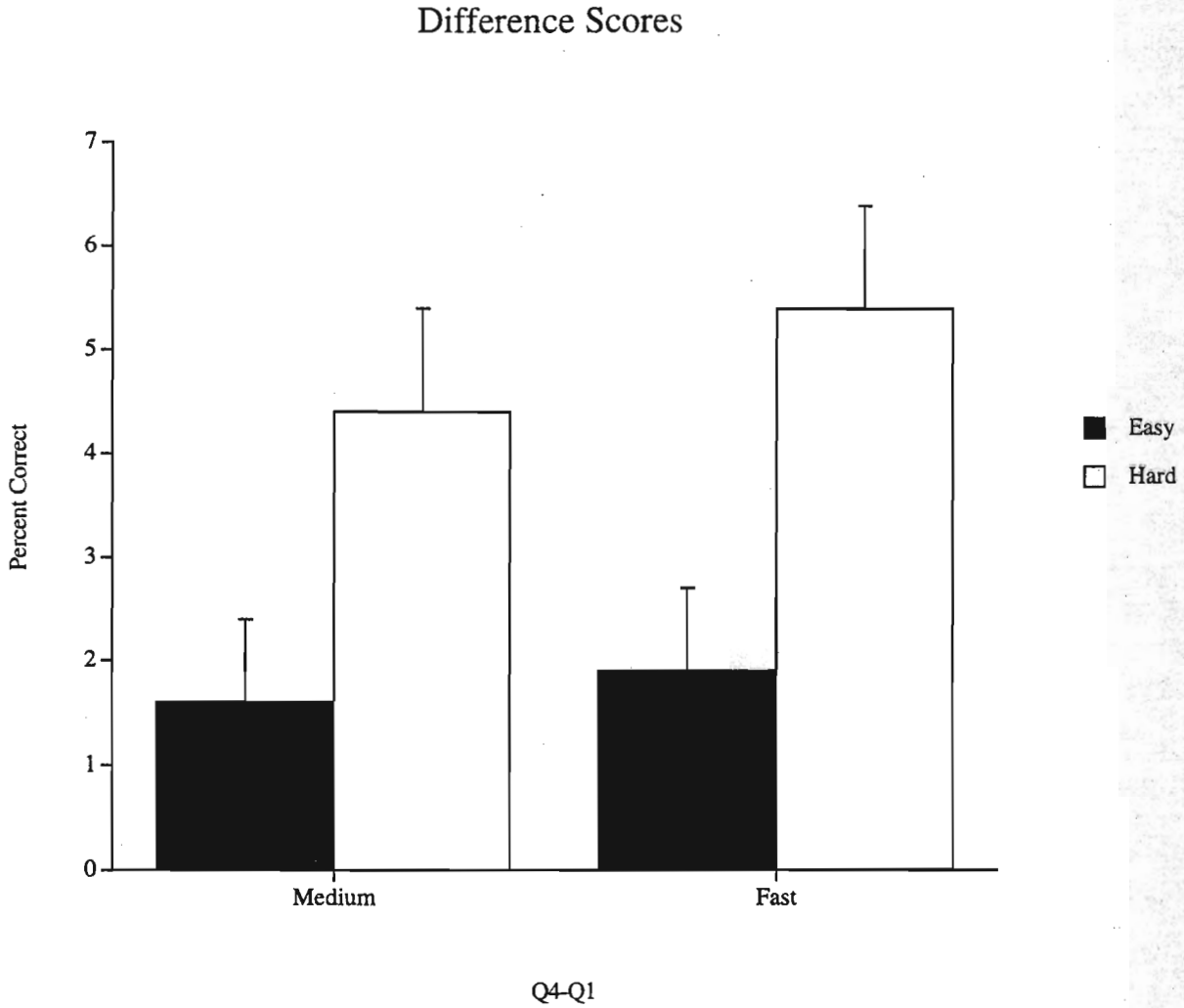


Figure 6: Difference scores showing the interaction between lexical characteristics and quartile presentation order for medium and fast rates of speech.

during the course of the experiment. Larger learning effects for the "hard" lexical category may be due to either lexical complexity, neighborhood characteristics of these words, or ceiling effects of performance for "easy" words. A follow-up study, in which the stimulus tokens are presented in noise, should help to determine which of these accounts is correct.

Phase Three: Phonetic Analysis

Phase three of this project involves acoustic-phonetic measurements of the individual talkers included in this database. A vowel-space measure will be obtained for each talker. A subset of words will be selected to represent a minimum of six tokens for each of the vowels /i, a, o/. First and second formant frequencies will be measured from the steady-state portions of the vowels, and individual vowel spaces will be plotted in order to compare vowel space area across talkers, speaking rates and lexical categories. A phonetic assessment of this type can provide a basis for investigating the relationship between speech production, lexical characteristics and overall intelligibility. This final phase will complete the initial project.

Summary

The overall goal of this project was to provide researchers in our laboratory with a large multiple talker database of isolated spoken words that have specific lexical characteristics. The additional data provided with this digital audio speech database, such as the intelligibility measures and the acoustic-phonetic analyses, are useful data from which researchers may plan further experiments appropriately.

References

Hernandez S., L.R. (1995) Implementation of a PC-based perceptual testing system (PTS): A first milestone. *Research on Spoken Language Processing Progress Report No. 19*. Bloomington, IN: Speech Research Laboratory, Indiana University. Pp. 321-328.

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.

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

Luce, P.A., Pisoni, D. B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 122-147). Cambridge, MA: MIT Press.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10.* Bloomington, IN: Speech Research Laboratory, Indiana University. Pp. 357-376.

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.

Sommers, M.S., Kirk, K.I., & Pisoni, D.B. (1996). Some considerations in evaluating spoken word recognition by normal-hearing and cochlear implant listeners I: The effects of response format. *Research on Spoken Language Processing Progress Report No. 20* (this volume). Bloomington, IN: Speech Research Laboratory, Indiana University. Pp. 31-49.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Implanted Children Can Speak, But Can They Communicate?[1]

Amy McConkey Robbins,[2] Mario Svirsky,[2] and Karen I. Kirk[2]

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

[1] This work was supported by NIH-NIDCD Grant DC 00064 to Indiana University. The assistance of Susan L. Todd, Allyson I. Riley, and Kathy S. Kessler in testing the children, Terri Kerr in the data analyses and Linette Caldwell in manuscript preparation is gratefully acknowledged. We also thank Richard T. Miyamoto and David B. Pisoni for their invaluable suggestions throughout the course of this project.

[2] Also Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Implanted Children Can Speak, But Can They Communicate?

Abstract. English language skills were evaluated in two groups of profoundly hearing-impaired children using the *Reynell Developmental Language Scales–Revised* (RDLS; Reynell & Huntley, 1985). The first group consisted of 89 deaf children who had *not* received cochlear implants. The second group consisted of 23 children wearing Nucleus multichannel cochlear implants. The unimplanted subjects provided cross-sectional language data used to estimate the amount of language gains which were expected on the basis of maturation. The Reynell data from the unimplanted group were subjected to a regression by age. Based on this analysis, deaf children were predicted to make half or less of the language gains of their normally-hearing peers. Predicted language scores were then generated for the implant subjects using their preimplant RDLS scores. The predicted scores were then compared to actual scores achieved by the implant subjects at 6- and 12-months post-implant. At 12 months post-implant, the subjects demonstrated receptive and expressive language gains that exceeded by seven months the predictions made on the basis of maturation alone. Moreover, the average language development rate of the implant subjects in the first year of device use was equivalent to that of normally-hearing children. These effects were observed for implant children using both the Oral and Total Communication methods.

Introduction

Multi-channel cochlear implants have been shown to provide substantial benefit to many children in the acquisition of speech skills (Kirk, Diefendorf, Riley, & Osberger, 1995; Osberger, Robbins, Todd, & Riley, 1995; Tobey, Geers, & Brenner, 1994). Although improved speech production is often the focus of parents and teachers working with deaf children, speech skills do not necessarily ensure language competence. Speech refers to oral productions, whereas language is the internalized, abstract knowledge system that is the basis for communication. Language ability is a very strong predictor of reading achievement and, hence, academic success in children (Goldgar & Osberger, 1986). If it could be shown that implants enhance language development, this would be compelling evidence as to the usefulness of cochlear implants in the pediatric population.

Assessing the effects of cochlear implant use on language development is difficult because some improvement in language skills occurs over time as a result of maturation. Due to the considerable variability in language scores over time, an ideal research paradigm would involve comparing the scores of an implanted subject to scores which the same child would have achieved if he had not received a cochlear implant. Given that this is impossible, an alternative method is to make informed predictions about each subject's language performance in the absence of a cochlear implant, and then to compare those predictions to the scores the child actually achieves.

In this paper, we first describe a method used to predict the development of receptive and expressive language skills in deaf children that might be expected with maturation. Then, we longitudinally compare predicted and observed language scores in a group of children using multi-channel cochlear implants to assess the effect of implant use on language development. If the observed language performance

over time exceeds that which has been predicted on the basis of maturation alone, then this would suggest that cochlear implant use enhances language development.

Methods

Subjects

Two groups of hearing-impaired subjects participated in this investigation.

Unimplanted Subjects. The first group consisted of 89 profoundly deaf, unimplanted children ranging in age from 16 to 95 months. All subjects wore either a hearing aid or tactile aid, and were audiologically suitable for a cochlear implant. All experienced early onset of deafness; 62 subjects were congenitally-deafened, whereas 27 of the 89 subjects were deafened between birth and 2 years, 11 months of age. Of these subjects, 61% used Total Communication (TC) and 39% Oral communication. The subjects in the unimplanted group provided cross-sectional data on language scores as a function of age that were used to generate predictive equations for language development in deaf children.

Cochlear Implant (CI) Subjects. This group consisted of 23 children wearing multichannel cochlear implants. All CI subjects were prelingually deafened; 11 of the 23 subjects were congenitally-deafened, whereas 12 experienced early, acquired deafness, prior to the age of three years. The average age at onset of deafness was 10 months, and the average age at time of implantation was 4 years, 11 months. Fourteen of the CI subjects used TC, the remaining nine subjects used Oral Communication. The Oral and TC groups were well-matched for age at onset of deafness (mean = .72 years for Oral subjects; .9 years for TC subjects) and age at implantation (mean = 4.98 years for Oral subjects; 4.86 for TC subjects). All subjects wore the Nucleus multichannel cochlear implant; five used the FO/F1/F2 strategy, 11 used the MPEAK strategy, and seven used the SPEAK strategy.

Test Instrument

The *Reynell Developmental Language Scales-Revised* (RDLS; Reynell & Huntley, 1985) was used to assess the English language abilities of the subjects. This assessment tool was chosen for several reasons: it evaluates receptive and expressive language independently, an important criterion, according to child language experts. It has been used extensively with deaf children (Moeller, Osberger, & Eccarius, 1986) and is appropriate for a broad age range, one to eight years of age, allowing repeated test administrations over a relatively long period of time. Normative data are available on 1319 hearing children (Reynell & Huntley, 1985). In addition, the test format involves object manipulation and description based upon questions which vary in their length and complexity. This format reflects real-world communication to a greater degree than do many other language tests (Muma, 1978). Thus RDLS results are considered to more accurately portray a child's communicative competence than does, for example, a single-word vocabulary test (Moeller et al., 1986; Robbins, Osberger, Miyamoto, & Kessler, 1994).

Prediction of Language Development in Deaf Children Without Cochlear Implants

The RDLS was administered once to each of the 89 subjects in the unimplanted group. The test was administered in whatever modality of English each child used, including spoken English, TC (i.e., simultaneous spoken English and Signing Exact English; Gustason & Zawolkow, 1993) or Cued Speech. Each child's responses were converted to a receptive and expressive language age in months. We then performed separate linear regressions of receptive and expressive language as a function of age at time of

testing. This allowed us to estimate the rate of language development in deaf children without implants. In other words, we used cross-sectional data from a group of deaf children to obtain regression slopes that could be used to predict the longitudinal changes in individual subjects. We shall refer to the slopes obtained from this regression analysis as "deaf slopes," one for receptive and one for expressive language. The prediction method used in this study assumes that the language age of implanted children would follow a straight line, starting at the child's language age and chronological age at the pre-implant testing session, and increasing according to the corresponding deaf slope.

Comparison of Observed and Predicted Language Performance in CI Children

The RDLS was administered to the 23 implanted children at three intervals using the administration and scoring procedures described above. The preimplant (PRE) measure was obtained approximately 0 to 3 months before initial hook-up. The two postimplant assessments were carried out at approximately 6- and 12-months postimplant (referred to as the POST1 and POST2 intervals, respectively). These three assessments yielded "observed" language scores for each child. Predicted scores were generated for each subject at the same test intervals using the regression equations calculated in the previous analysis. That is, we assumed that these CI subjects' language scores would have increased over time at a rate described by the deaf slopes, if they had not received a cochlear implant. Therefore, predicted scores for a given subject over time are described by a straight line that starts at his pre-implant score and age, and increases according to the deaf slope. At the PRE interval, predicted scores are, by definition, identical to observed scores. Using these data, we performed a two-way repeated measures ANOVA (repeated measures both on the "interval" and the "observed vs. predicted" variables).

One difficulty in comparing the observed and predicted language scores of the implanted children is that communication mode varied across subjects. It is possible that the relationship between observed and predicted scores in the TC and Oral groups were different. To investigate this possibility, we performed an additional analysis. We first separately calculated the average receptive and expressive language gains that were made by each one of the TC and Oral groups from the PRE to the POST1 interval, and from the POST1 to the POST2 interval. These gains then were compared to the language progress that would have been predicted by maturation alone (using the "deaf slopes" to make this prediction) within the Oral and TC groups separately.

Insert Figure 1 about here

Results

Prediction of Language Development in Deaf Children Without Cochlear Implants

In Figure 1, each of the 89 subjects' chronological age is plotted against his or her receptive language age. The dotted, diagonal line indicates the language change expected by a normally-hearing child; that is, language age and chronological age increase in synchrony. The slope of the normal-hearing, diagonal line is 1. The solid line shows a regression by age of the Reynell receptive language data, expressed in age equivalent scores. The r^2 of this regression was 0.61, and the slope was 0.5, suggesting that the receptive language gains to be expected from profoundly deaf children are roughly half those of normally-hearing peers. Thus, we predict that deaf children will show about six months' receptive language growth in one year.

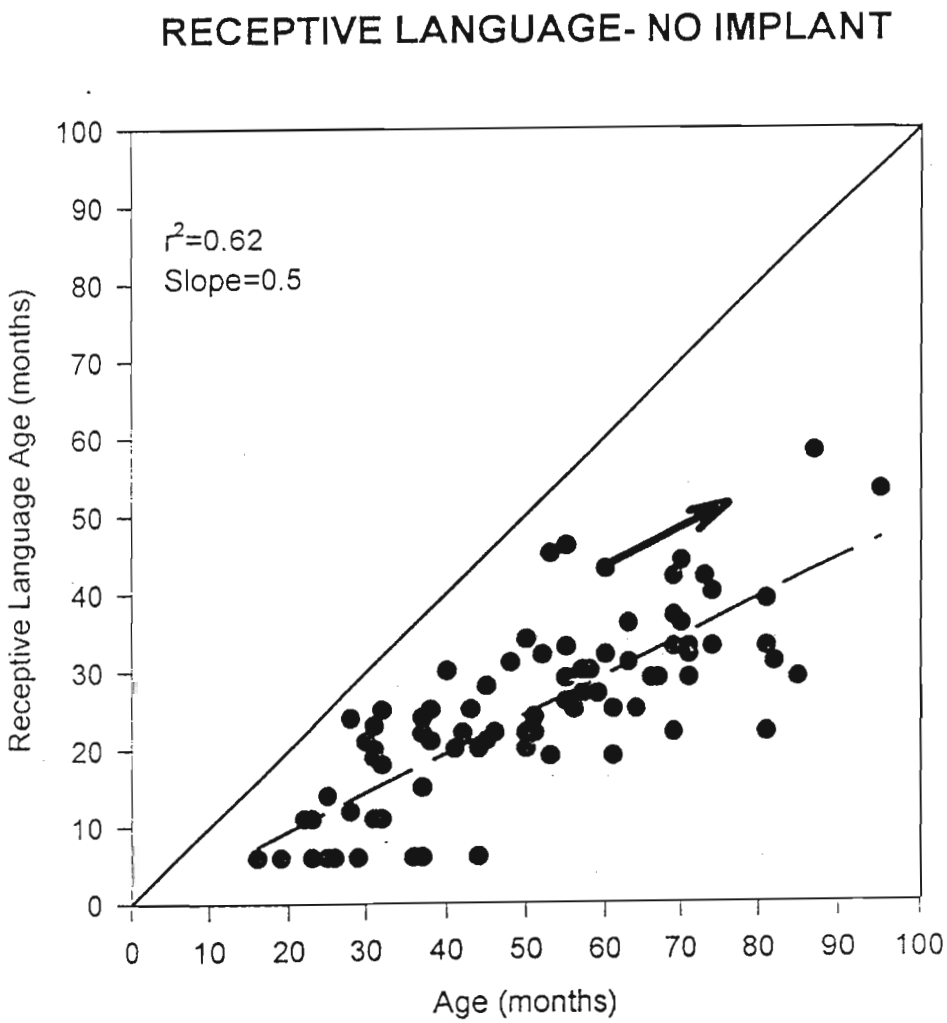


Figure 1. Reynell receptive language data for 89 unimplanted deaf children. Chronological age (in months) is plotted on the x axis and receptive language age is plotted on the y axis. The linear regression is shown by a dashed line, and the solid, diagonal line illustrates receptive language growth expected of a normally-hearing child.

The prediction method described above is illustrated by the arrow in Figure 1, showing the prediction line that corresponds to one specific subject.

The results for the expressive language scores on the RDLS are shown in Figure 2. Note the similar pattern as that for receptive skills, although the r^2 is 0.53 and the expressive deaf slope is only .42, lower than the receptive deaf slope. Based on these expressive language data, we predict that deaf children will show about five months of expressive language progress in one year.

Insert Figure 2 about here

Comparison of Observed and Predicted Language Performance in CI Children

The predicted and observed mean receptive language scores for the group of implanted subjects are shown in Figure 3. The horizontal axis represents chronological age and the vertical axis shows receptive age-equivalent scores, measured in months. The solid diagonal line running through the graph represents the language development expected of a normally-hearing child. The first filled dot is the group average of age-equivalent scores obtained by the children at the pre-implant interval. The open circles represent the mean scores that would be predicted at 6- and 12-months post-implant using the prediction method described above. The slope of language growth predicted for these children without an implant is shown by the broken line. Notice the shallowness of this line relative to that for the normally-hearing child (the solid diagonal line). The two later filled circles represent the language scores actually observed in these implanted children when they were tested at 6 and 12 months post-implant. As a group, these subjects showed language growth that, at one year post-implant, exceeded by 7.1 months the predictions that were made on the basis of maturation alone.

Insert Figures 3 and 4 about here

The pattern of results for expressive language in the implanted subjects was similar to that for receptive language (Figure 4). Note the pre-implant score, and the two open circles which are the scores predicted based upon the formula described earlier. The actual scores that the implanted subjects achieved are represented by the filled circles. Comparing the predicted with the actual scores after one year of implant use, there is a 6.9 month advantage of the actual over the predicted score. Note that although the implant subjects' language scores are higher than the corresponding predictions for unimplanted deaf children, their language scores remain significantly below those of their normally-hearing peers. This may be seen in Figure 4 by comparing the position of the closed circles (post-implant language performance) relative to the solid line (normally-hearing subjects).

The variables used in the two-way repeated measures ANOVA analyses were interval (6 or 12 months post-implant) and observed/predicted. There were significant ($p<0.05$) interaction effects for both receptive and expressive language scores, indicating that the difference between observed and predicted values depends on the interval. Post-hoc Student-Newman-Keuls tests indicated that both receptive and expressive observed scores were significantly higher than the corresponding predicted scores at the 12

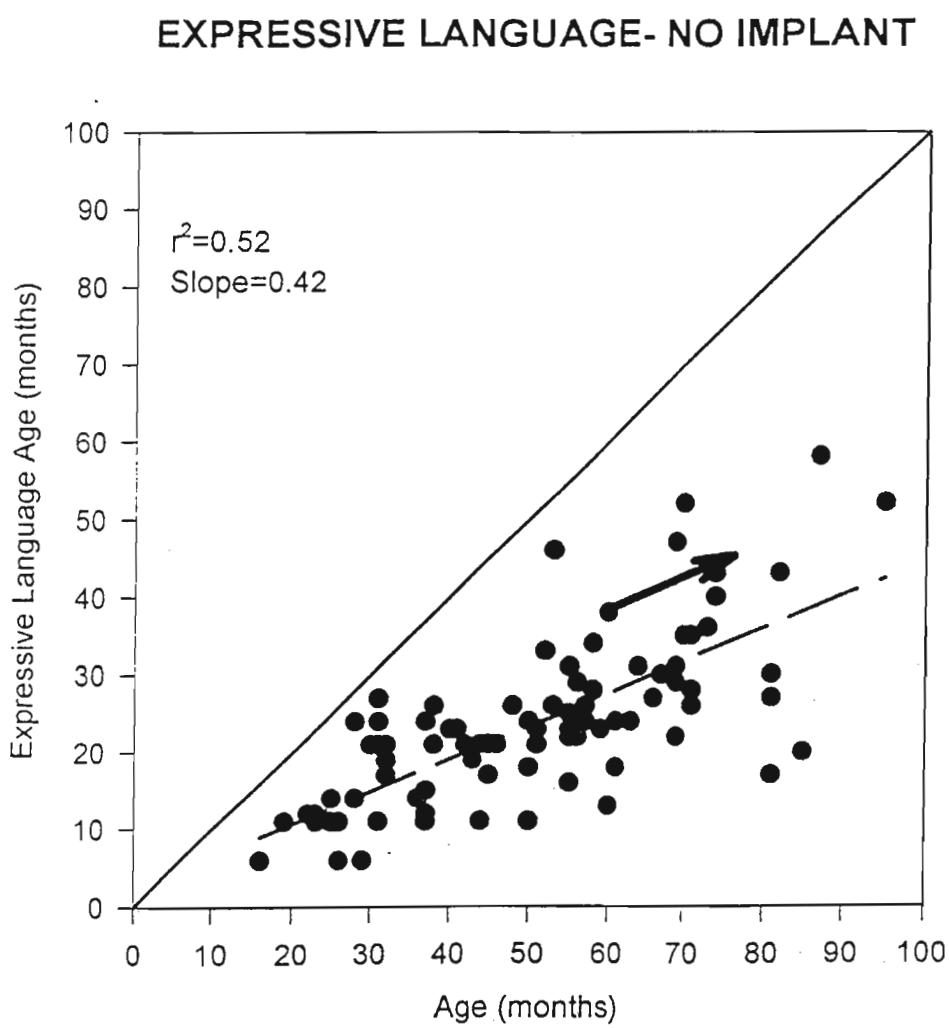


Figure 2. Reynell expressive language data for 89 unimplanted deaf children. Chronological age (in months) is plotted on the x axis and expressive language age is plotted on the y axis. The linear regression is shown by a dashed line, and the solid, diagonal line illustrates the expressive language growth expected of a normally-hearing child.

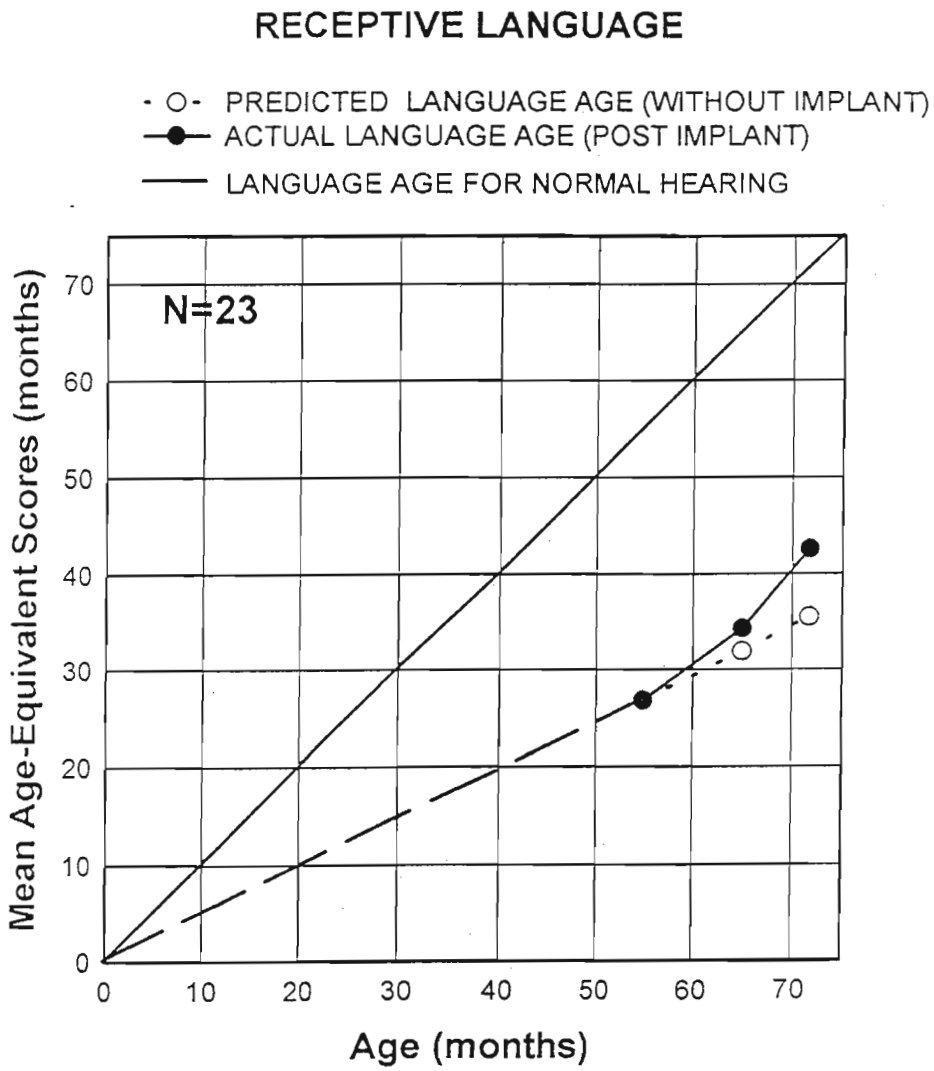


Figure 3. The open circles show the receptive language development that would be predicted for unimplanted children while the filled circles show actual scores obtained from the 23 CI subjects. The solid, diagonal line illustrates receptive language growth expected of a normally-hearing child.

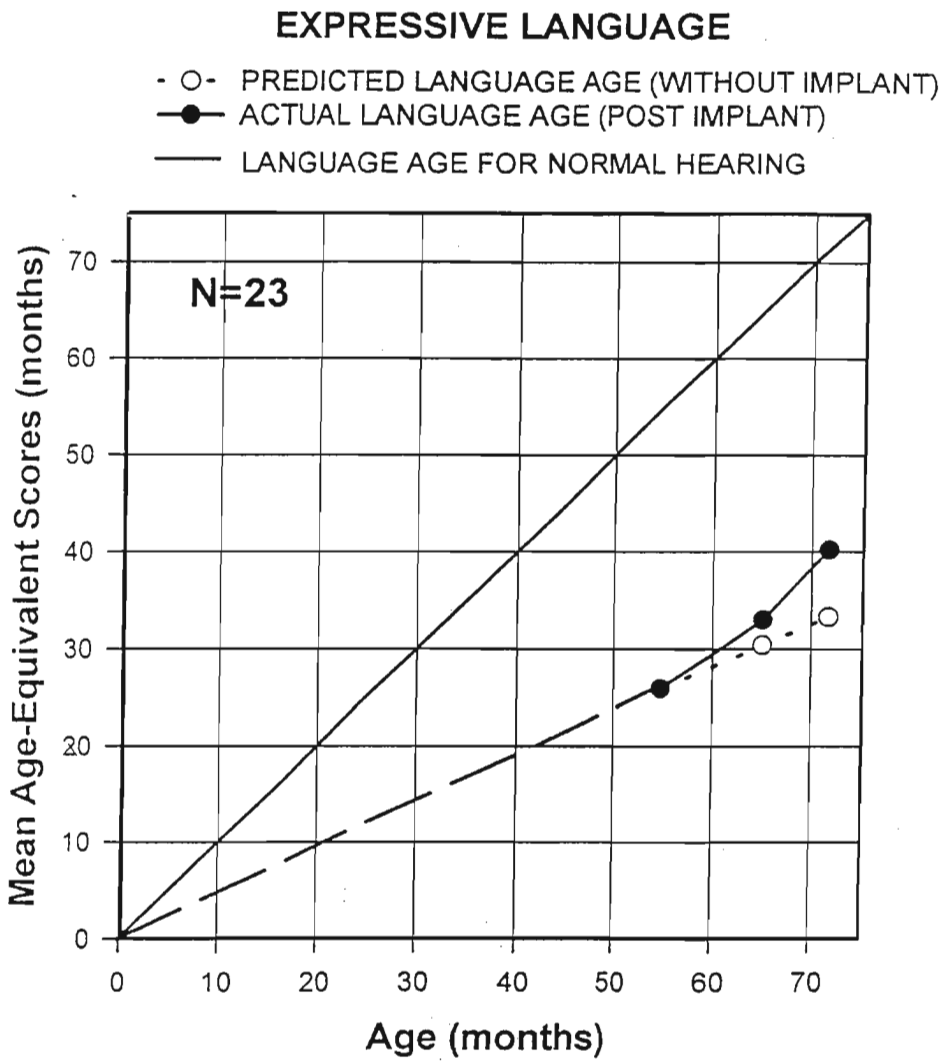


Figure 4. The open circles show the expressive language development that would be predicted for unimplanted children while the filled circles show actual scores obtained from the 23 CI subjects. The solid, diagonal line illustrates the expressive language growth expected by a normally-hearing child.

month interval ($p<0.05$). At the 6-month interval, however, the difference between observed and predicted scores failed to reach statistical significance.

Figure 5 illustrates the average amount of language growth that occurred in the Oral and TC implant groups beyond that expected on the basis of maturation. If the progress achieved was equivalent to that expected by maturation, the bars would be at zero. The top panel of Figure 5 indicates that at 6 months post-implant, the TC children (open bars) had made three months progress beyond that expected on the basis of maturation. The oral subjects (solid bars) also made progress beyond that expected by maturation, but only one month beyond. At 12 months post-implant, the TC group had made an average of seven months gain and the Oral group averaged almost 8 months of language gain beyond that expected through maturation alone. Expressive findings for the two communication groups are shown in the bottom panel of Figure 5. Note the mixed pattern of progress seen in the Oral and TC groups at the two intervals, although by 12-months post-implant, both groups show considerable increases in expressive language beyond those due to maturation.

Insert Figure 5 about here

Discussion

These data suggest that the English language skills of profoundly deaf subjects without cochlear implants improved at a rate that was markedly slower than that of normally-hearing children, a finding that is in agreement with previously-reported studies on language development in deaf children (Osberger, 1986; Levitt, McGarr, & Geffner, 1987). Data from the non-implanted subjects allowed us to calculate rates of language development (deaf slopes) that would be predicted for a deaf child on the basis of maturation alone. Recall that these deaf slopes were .5 for receptive and .42 for expressive language, respectively, suggesting that deaf children would be expected to make language gains at half or less the rate of normally-hearing children.

We found that observed language scores for the subjects with cochlear implants were significantly higher than the predictions made for the same subjects if they had not received cochlear implants. Specifically, after about one year of device use, the implant subjects' mean receptive and expressive language scores were approximately seven months better than those predicted on the basis of maturation. This suggests that the cochlear implant promoted both receptive and expressive language development to a greater extent than would be predicted by maturation alone. In addition, the findings show that the longer the children used their implants, the greater the difference between the observed and predicted scores. Both of these findings are in agreement with our earlier investigations (Robbins et al., 1994; Robbins, 1993).

Both children using Oral language and those using TC demonstrated an increased rate of language learning with the cochlear implant. Given that the implant is an auditory sensory aid, one might expect to see greater benefit for Oral subjects, whose language learning is primarily mediated through the auditory modality, than for TC subjects, whose language learning is strongly, though not exclusively, visual. That the cochlear implant provided language benefit for both groups is particularly encouraging, in light of the fact that at least 75% of deaf children in the United States are educated in Total Communication programs.

The possibility also exists that the cochlear implant has a global, multi-sensory effect on language learning. This would be consistent with the recent findings of Quittner, Smith, Osberger, Mitchell, and

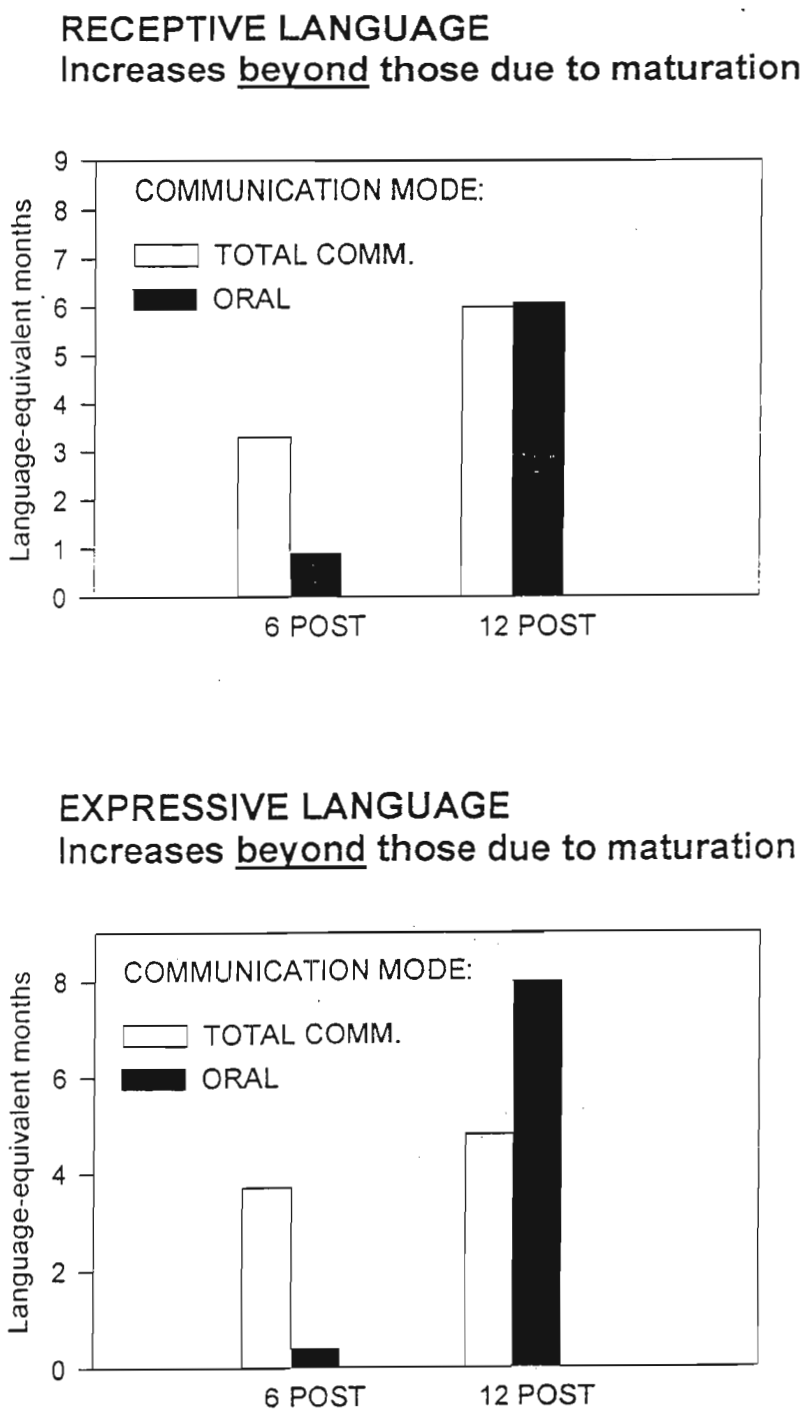


Figure 5. Average amount of language growth (in months) beyond that due to maturation in implant subjects using Total Communication (open bars) and Oral Communication (solid bars). The top panel shows receptive language results, the bottom panel shows expressive language results.

Katz (1994) who reported increases in selective visual attention in children following cochlear implantation, and with anecdotal reports by parents and teachers of children's improved attention to task following implantation. Such an effect might explain the benefit demonstrated by both Oral and TC subjects in the present study.

It is also possible that the rapid change in the rate of language learning rate following implantation may relate to the children's new-found ability to acquire language incidentally, through the overhearing of everyday conversations. This natural exposure to spoken language communication is the avenue by which normally-hearing children learn their language skills, and is generally unavailable to profoundly deaf children who must be directly and explicitly taught every spoken language structure they know. Access to natural conversation through the implant would allow these children the exposure to language that previously was inaccessible.

It is interesting to note that, at least for the first year after implantation, increases in receptive and expressive language scores for implanted children matched that of the normally-hearing group. In consequence, the gap in absolute scores between implanted and normal children remained roughly constant during that first year post-implant, instead of increasing. If this result happened to be true for subjects implanted earlier, and was consistent beyond one year post-implant, the case for earlier implantation would be considerably strengthened. The younger a deaf patient is, the smaller the gap between his language age and chronological age. If the auditory information provided by a cochlear implant prevented the language gap from increasing (as our data suggest), early implanted children would have an excellent chance of achieving near-normal language development. In consequence, it is crucial to extend these studies in two directions: looking at subjects implanted earlier in life, and following them for longer periods of time post-implantation.

The results of the present study demonstrate an important consequence of cochlear implants, namely, the foundation of language development above and beyond that anticipated from maturation alone. Thus, not only do children display improvement in speech perception and speech intelligibility with cochlear implants, but they show significant increases in their rate of language development.

References

Goldgar, D. & Osberger, M.J. (1986). Factors related to academic achievement. In M.J. Osberger (Ed.) *Language and Learning Skills of Hearing-Impaired Students. ASHA Monograph,* **23**, 87-91.

Gustason, G. & Zawolkow, E. (1993). *Signing Exact English*. Los Alamitos, CA: Modern Signs Press.

Kirk, K.I., Diefendorf, A., Riley, A, & Osberger, M.J. (1995). Consonant production by children with multichannel cochlear implants or hearing aids. *Advances in Otorhinolaryngology*, **50**, 154-159.

Levitt, H., McGarr, N. & Geffner, D. (1987). Development of language and communication skills in hearing-impaired children. *ASHA Monographs*, **26**.

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 Monograph,* **23**, 41-53.

Muma, J.R. (1978). *Language Handbook*. Englewood Cliffs, NJ: Prentice Hall.

Osberger, M.J. (1986). Language and learning skills of hearing-impaired students. *ASHA Monographs,* **23**.

Osberger, M.J., Robbins, A.J., Todd, S.L. & Riley, A.I. (1995). Speech intelligibility of children with cochlear implants. *The Volta Review*, **5**, 169-180.

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

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

Robbins, A.M. (1993). Language evaluation of children with cochlear implants. Paper presented at the *Third International Conference on Pediatric Otolaryngology*, Jerusalem, Nov. 1993.

Robbins, A.M., Osberger, M.J., Miyamoto, R.T. & Kessler, K.S. (1994). Language development in young children with cochlear implants. *Advances in Otorhinolaryngology*, **50**.

Tobey, E.A., Geers, A., & Brenner, C. (1994). Speech production results: Speech feature acquisition. *The Volta Review*, **96**(5), 109-129.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Lexical Discrimination and Age of Cochlear Implantation: A First Report[1]

Karen I. Kirk[2] and David B. Pisoni

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

[1] This work was supported by NIH-NIDCD Research Grant DC00423 to Indiana University School of Medicine and NIH-NIDCD Training Grant DC00012 to Indiana University-Bloomington. We thank Susan Todd, Amy Robbins, Allyson Imber Riley, and Erin Diefendorf for assistance with data collection, and Terri Kerr and Linette Caldwell for technical assistance.
[2] Also Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Lexical Discrimination and Age of Cochlear Implantation: A First Report

Abstract. This paper examines the effects of age at time of implantation on word recognition and word discrimination by two groups of prelingually deafened pediatric cochlear implant (CI) users. An 'early' group consisted of children receiving implants at earlier than 6 years of age, while a 'late' group was implanted at 6 years or later. Both groups were tested on both lexical discrimination tasks (LNT, MLNT) and language measures (PPVT, Reynell). For monosyllabic words, significant age effects were found for both word and phoneme recognition, while significant effects of lexical difficulty were found for word recognition but not phoneme recognition. Similar results obtained for multisyllabic words, although there was no effect of lexical difficulty on either word or phoneme recognition. In addition, word recognition was moderately correlated with some measures of receptive and expressive language abilities in both the early CI and late CI groups. These preliminary findings suggest that early implantation produces better word recognition performance in young, prelingually-deafened children and that children who are better at making fine discriminations among phonetically similar words also exhibit better language skills.

Introduction

Recent research (e.g., Waltzman et al., 1994) suggests that prelingually-deafened children who receive a cochlear implant (CI) at an early age (i.e., before age 5-6 years) may obtain greater speech perception and language benefits than those children implanted at a later age. There are at least two accounts of this advantage. First, children implanted at an early age may still be within the critical period for the acquisition of spoken language skills. Second, because these children had briefer amounts of auditory deprivation, they have had more listening experience than their peers who were implanted at a later age. Both of these explanations would predict differences in the word recognition abilities of children implanted at earlier vs. later ages -- specifically improvements in lexical discrimination performance. That is, we predict that children who receive a CI prior to six years of age will demonstrate better word recognition performance, and will also be better at making fine lexical discriminations among phonetically similar words than children who are implanted at or after six years of age.

Purpose

The goals of the present investigation were:

1) To examine the effects of age at time of implantation on word recognition and lexical discrimination by pediatric cochlear implant users, and

2) To investigate the relationship between lexical discrimination and language skills in pediatric cochlear implant users.

Method

Subjects

Two groups of pediatric cochlear implant users participated as subjects in this investigation. All subjects were prelingually deafened (i.e., < 3 years) and received a cochlear implant because they essentially derived no benefit from conventional amplification. The *early group* consisted of 25 children who received their cochlear implant between the ages of 2.0 and 5.9 years. The *late group* consisted of 12 children who received their cochlear implant between the ages of 6.0 and 8.9 years. Subject characteristics are presented in Table 1 and device characteristics in Table 2.

Table 1

Subject Characteristics

	Early CI Group N=25		Late CI Group N=12	
	Mean	(SD)	Mean	(SD)
Age at Onset (yrs)	0.3	(0.7)	0.4	(0.6)
Age Fit with CI (yrs)	4.3	(0.9)	7.2	(1.1)
Length of Auditory Deprivation (yrs)	4.0	(1.2)	6.9	(1.1)
Age at Time of Testing (yrs)	7.9	(2.1)	9.9	(2.7)
Length of CI Use (yrs)	3.5	(1.9)	2.7	(2.1)
Unaided PTA (dB HL)	>110		>110	
Communication Mode	TC =13 Oral = 12		TC = 6 Oral = 6	

Stimulus Materials

Lexical Discrimination Tasks. The Lexical Neighborhood Test (LNT) and the Multisyllabic Lexical Neighborhood Test (MLNT) (Kirk, Pisoni & Osberger, 1995) are word lists constructed to allow systematic examination of the effects of word frequency (i.e., the frequency of occurrence of each word) and lexical difficulty (i.e., the number of phonemically similar words or neighbors to a target, as determined by counting the number of words that can be generated by adding, subtracting, or deleting a single phoneme) on spoken word recognition in children. The LNT contains two lists of 50 monosyllabic words, and the MLNT contains two lists of 30 two- and three-syllable words. Each test has an equal number of easy and hard words. Easy words have high word frequency and have few phonemically similar words with which they can be confused. Hard words have low word frequency and have more phonemically similar neighbors. The percent of words and phonemes correctly identified is determined separately for easy and hard words.

Table 2

Device Characteristics

	Early CI Group N=25	Late CI Group N=25
Processor Type	WSP = 0 MSP = 10 Spectra = 15	WSP = 1 MSP = 5 Spectra = 6
Processing Strategy	F0F1F2 = 0 MPEAK = 10 SPEAK = 15	F0F1F2 = 1 MPEAK = 5 SPEAK = 6
Stimulation Mode	CG[a] = 10 BP [b] = 3 BP+1[c] = 7 BP+2[d] = 3 BP+3[e] = 2	CG[a] = 8 BP [b] = 0 BP+1[c] = 3 BP+2[d] = 1 BP+3[e] = 0
Number of Active Electrodes	13-22	5-22

[a] Common Ground, [b] Bipolar, [c] Bipolar + 1, [d] Bipolar + 2, [e] Bipolar + 3

Language Measures. The *Peabody Picture Vocabulary Test* (Dunn, 1965) was used to assess receptive vocabulary skills in both groups. The single-word stimuli were presented in the child's communication mode, along with a written and/or signed representation, as appropriate. Children responded by pointing to one of four words pictured on a single plate, and a receptive vocabulary age was determined for each child.

The *Reynell Developmental Language Scales - Revised* (Reynell & Huntley, 1985) is a measure designed for normally-hearing children ages 1-7 years. The test is organized in a hierarchy of difficulty, utilizing a variety of question forms, and yields both a receptive and expressive language quotient (language age/chronological age).

Procedures

All subjects were tested in a quiet room, seated across a table from the examiner. Stimuli were presented via live voice at approximately 70-75 dB SPL by one of five experienced audiologists or speech-language pathologists. Language assessments were carried out in the childs communication mode, either Total Communication using Signed English, or orally; written cues also were provided during vocabulary assessment. Word recognition/lexical discrimination was assessed in the auditory-only modality. Children responded by repeating the word they heard. If the child's verbal response could not be understood, it was phonemically transcribed and scored; no credit was given for word recognition.

Results

Spoken Word Recognition / Lexical Discrimination

Figures 1 and 2 present the percent of words and phonemes correctly identified on the LNT and MLNT. Separate two-way analyses of variance were performed for each test, with word and phoneme scores as the dependent measures and age at implantation and lexical difficulty as the independent variables. Because of the relatively small number of subjects in the Late implantation group, alpha levels between .05-.1 were considered marginally significant.

Insert Figures 1 and 2 about here

Monosyllabic Stimuli (LNT). Results from administration of the Lexical Neighborhood Test were as follows:

- Significant effects of age at time of implantation were found for both monosyllabic word and phoneme recognition. The average scores for the Early and Late groups were 40% vs. 29% for word identification ($p < .06$), and 60% vs. 48% for phoneme identification ($p < .07$).

- Significant effects of lexical difficulty were found for word recognition ($p < .02$) but not for phoneme recognition. On average, 40% of the Easy words and 30% of the hard words were correctly identified. The percent of phonemes correctly identified in the Easy and Hard words was 56% vs. 52%, respectively.

- The interactions between age at implantation and lexical difficulty for word or phoneme recognition were not significant.

Multisyllabic Stimuli (MLNT). Results from administration of the Multisyllabic Lexical Neighborhood Test were as follows:

- The effects of age at implantation were significant for both multisyllabic word ($p < .02$) and phoneme ($p < .007$) recognition. Average scores for the Early and Late groups respectively were 56% vs. 40% for word identification, and 70% vs. 53% for the phoneme identification.

- There was no effect of lexical difficulty on multisyllabic word or phoneme recognition, and no interaction between age at implantation and lexical difficulty.

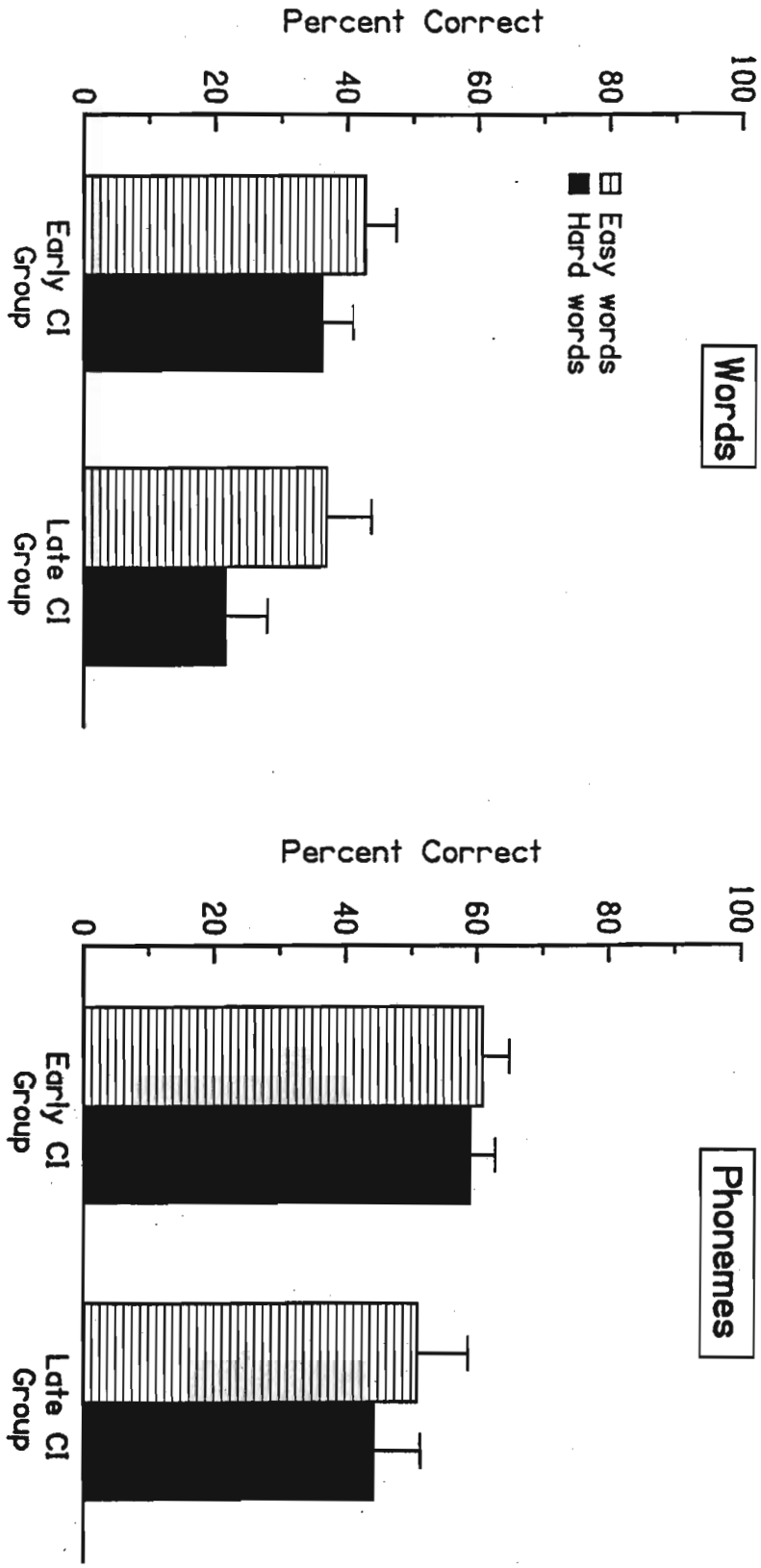


Figure 1. The percent of monosyllabic words and phonemes correctly identified as a function of age at time of implantation and lexical difficulty.

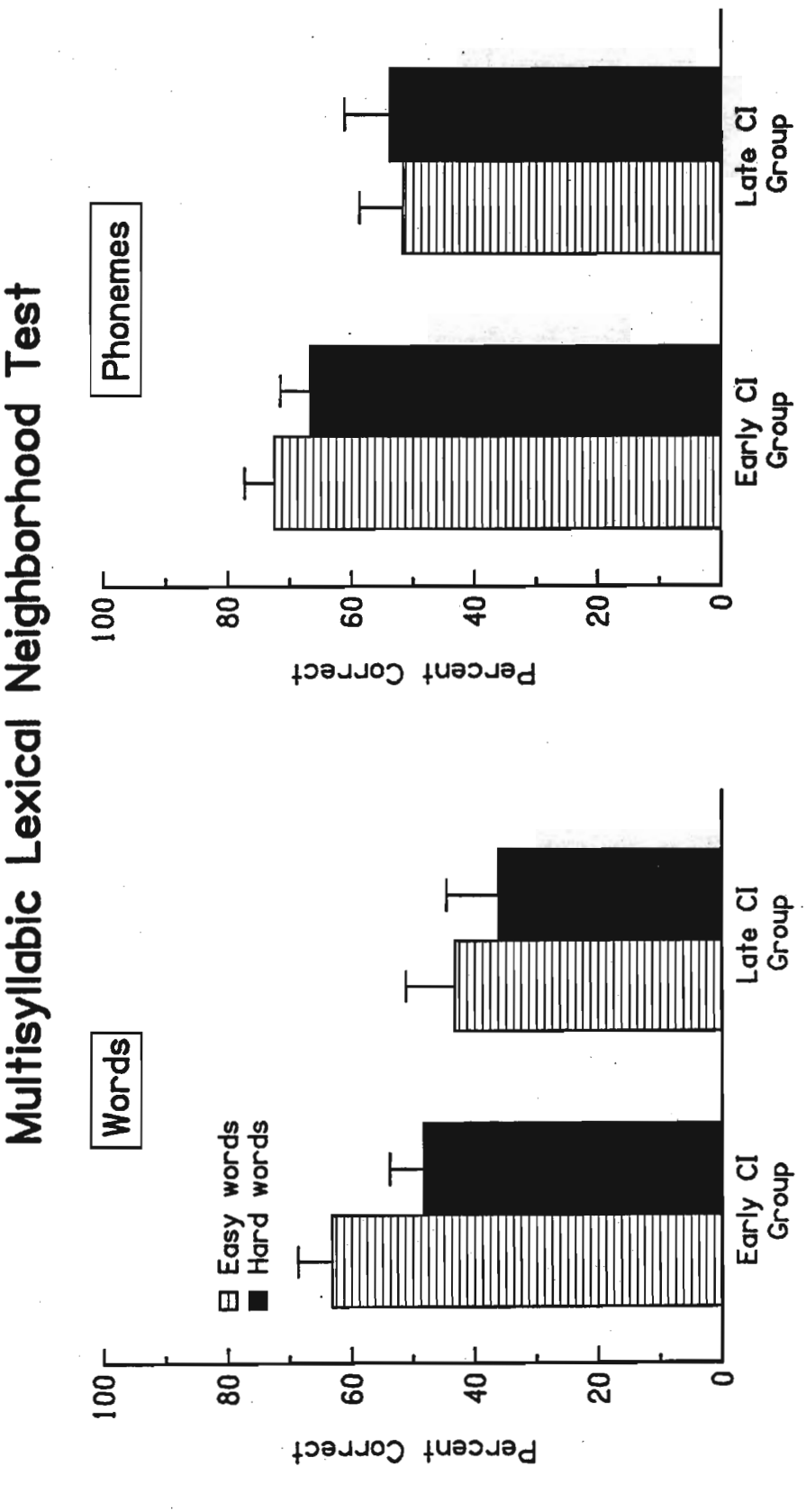


Figure 2. The percent of multisyllabic words and phonemes correctly identified as a function of age at time of implantation and lexical difficulty.

Correlations with Language Measures. To examine the relationship between word and language abilities, performance on the LNT and MLNT was correlated with the language quotients derived from the PPVT and the receptive and expressive portions of the Reynell Language Development Scales, as shown in Table 3. Only 15 children from the Early CI group were included in this analysis as the remaining had not been tested with the Reynell. The results demonstrated:

• Early CI Group - Performance on the Reynell Developmental Language Scales was moderately correlated with performance on the LNT Easy words ($r = +.43$ for both expressive and receptive quotients, not significant) and MLNT Hard words (receptive: $r = +.58$, $p < .02$; Expressive: $r = +.50$; $p < .06$). There was no relationship between performance on the PPVT and word recognition for this group of subjects.

• Late CI Group - The PPVT language quotient was moderately correlated with performance on the LNT Hard words ($r = +.55$) and the MLNT Hard words ($r = +.49$), but only the former was marginally significant ($p < .07$).

Table 3

Correlations of word recognition and language measures by age at time of implantation.

	Early CI Group (N=15)			Late CI Group (N=12)
	PPVT	Reynell - Expressive	Reynell - Receptive	PPVT
LNT - Easy	.07	.43	.43	.24
LNT - Hard	.14	.39	.35	.55# #(p<.07)
MLNT - Easy	-.02	.35	.34	.27
MLNT - Hard	.08	.50* *(p < .06)	.58+ +(p < .02)	.49

Discussion

These preliminary findings suggest that early implantation (i.e., by age 5 years) produces better word recognition performance in young prelingually-deafened children. Children in the early CI group had higher average word recognition scores, and were better at identifying lexically difficult words, than children who received their CI after 5 years of age.

Lexical factors influenced word recognition performance, but not phoneme recognition. Easy words were identified with significantly greater accuracy than Hard words, but this was significant only for monosyllabic words, possibly because of the small numbers of subjects tested in the Late CI group. The finding that Easy words were identified with greater accuracy than Hard words on the LNT is in agreement with the earlier results of Kirk et al. (1995).

Word recognition was moderately correlated with some measures of receptive or expressive language abilities for subjects in the Early and Late CI groups. The strongest correlations were found for the Hard words on the LNT and MLNT tests. Thus, the present findings suggest that that children who are better at making fine discriminations among phonetically-similar words stored in their lexicons are also those with better language skills. Word recognition is an important component of spoken language processing, as it represents the interface between sensory input and linguistic knowledge. In other words, it represents the interface between speech perception and spoken language comprehension.

References

Dunn, LM (1965). *Peabody Picture Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Kirk, KI, Pisoni, DB, & Osberger, MJ (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing*, **16**, 470-481.

Reynell, JK, & Huntley, M (1985). *Reynell Developmental Language Scales, Revised Edition 2*. Windsor: NFER Publishing.

Waltzman, SB, Cohen, NL, Gomolin, R, Shapiro, WH, Ozdamar, S, & Hoffman, R (1994). Long-term results of early cochlear implantation in congenitally and prelingually deafened children. *American Journal of Otology*, **15**, 9-14.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Acoustic and Glottal Excitation Analyses of Sober vs. Intoxicated Speech: A First Report[1]

Kathleen E. Cummings,[2] Steven B. Chin, and David B. Pisoni

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

[1]This work supported in part by grants to Indiana University from the Alcoholic Beverage Medical Research Foundation and the National Institutes of Health (NIH–NIDCD), Training Grant DC00012.

[2]Also Digital Signal Processing Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA.

Acoustic and Glottal Excitation Analyses of Sober vs. Intoxicated Speech: A First Report

Abstract: This is a first report of results from acoustic and glottal excitation analyses of speech produced both with and without alcohol. The new analyses were designed to determine whether there are significant and identifiable differences in phonation between the two types of speech. Non-nasal vowels extracted from eight isolated words produced by four talkers in both a nonalcohol and an alcohol condition were examined in terms of (1) direct measures of acoustic speech waveform parameters, (2) perturbation measures of acoustic speech waveform parameters, and (3) measures of the glottal excitation waveshape. Parameters related to the steadiness of speech production, as reflected in perturbations in adjacent pitch periods, exhibited differences between alcohol and nonalcohol speech. Specifically, we found consistent differences between the two types of speech on several measures of jitter, although the amount of variation between the alcohol and nonalcohol speech appeared to be talker-dependent.

Introduction

The goal of the research reported here was to analyze sober versus intoxicated speech in order to determine whether speech produced while a person is intoxicated is significantly and identifiably different from speech produced while a person is sober. To this end, several measures related to voiced excitation have been extracted and studied for four of the speakers in the Indiana University Alcohol Speech Database (Pisoni & Martin, 1989; Pisoni, Yuchtman, & Hathaway, 1986). These measures can be divided into three categories: direct measures of the acoustic waveform, perturbation measures of the speech waveform, and measures of the glottal excitation waveform. Because of physiological differences between speakers, most of these parameters may vary a great deal from speaker to speaker. The present research effort attempts to identify significant parameter variation trends in intoxicated versus sober speech for a particular speaker that are consistent for all of the speakers studied.

Of all of the parameters we have extracted thus far, those that are related to the steadiness with which a person produces speech are the parameters that best reflect the differences between intoxicated and sober speech. For example, several measures of jitter appear to be consistently different for intoxicated versus sober speech. The amount of variation between sober and intoxicated speech also seems to be speaker-dependent. In one speaker in particular, KM, all parameters vary less drastically between sober and intoxicated speech. We are currently investigating the possibility that he is a tolerant drinker.

Method

Materials for the acoustic and glottal excitation measurements described here were taken from a digital database of isolated monosyllabic words, isolated spondaic words, isolated sentences, and connected passages spoken by nine young adult male talkers in two conditions: without alcohol and under alcohol at .10% BAC or higher. Full details regarding talker selection and preparation, speech materials, and

elicitation procedures can be found in Pisoni et al. (1986) and Pisoni and Martin (1989). Analyses of up to four talkers are reported here. Table 1 shows subject data for these four talkers, including age, initial BAC, final BAC, and self-reported alcohol intake.

Table 1

Talker Characteristics

Age = age in years at last birthday; Initial BAC = BACs at beginning of recording in g/100 ml blood, as measured by Smith & Wesson Breathalyzer (Model 900A); Final BAC = BACs at end of recording; Alcohol Intake = self-reported total alcohol intake during 30 days prior to recording session, converted to oz 200-proof alcohol. (From Pisoni, Yuchtman, & Hathaway, 1986.)

Talker	Age	Initial BAC	Final BAC	Alcohol Intake
DP	21	0.15	0.10	8.94
JB	26	0.10	0.10	6.15
JS	22	0.16	0.10	3.53
KM	21	0.17	0.10	16.80

Speech tokens of isolated words were elicited in a shadowing task in both the alcohol and nonalcohol conditions; auditory stimuli were presented via audio tape playback over headphones, and talkers were instructed to simply repeat words aloud as quickly as possible. Vowels from eight isolated words ('chaff' (Word 11), 'chap' (Word 12), 'cheese' (Word 13), 'chest' (Word 14), 'chief' (Word 15), 'choose' (Word 17), 'chops' (Word 18), and 'heath' (Word 61)) in each of the two conditions, sober and intoxicated, were used for the analyses described here. All of the parameters have been extracted and studied for four speakers designated DP, JB, JS, and KM.

Each utterance was pitch-marked using a semi-automatic cepstrum-based pitch detector on a Sun 4 or Sun SPARCstation 20. The boundaries of the voiced sections were marked, and a pitch contour was determined. Parameters were then extracted using each original utterance, the voicing boundaries, and the pitch contours. Three types of measures of voiced excitation were extracted:

1. direct measures of acoustic speech waveform parameters,

2. perturbation measures of acoustic speech waveform parameters, and

3. measures of the glottal excitation waveshape.

The pitch contours (in samples, sampling rate of 20 kHz) are shown in Figures 1–8. Each figure contains four plots, each of these comparing the pitch contours for sober and intoxicated versions of the same word for a given speaker.

Insert Figures 1 through 8 about here

Parameters

Direct Measures of the Acoustic Speech Waveform

Typically, parameters that are used to distiguish between different speaking styles involve measures of the energy, or RMS intensity, in a given segment of speech, and measures of the pitch period (see Cummings, 1992). Several such parameters were extracted directly from the acoustic speech waveforms. These included:

1. **pave**: mean of the pitch contour for an utterance
2. **pc**: measure of the flatness of the pitch contour
3. **aint**: mean(total RMS intensity per pitch period) for an utterance
4. **mint**: max(total RMS intensity per pitch period) for an utterance
5. **naint**: mean((total RMS intensity per pitch period)/(pave)) for an utterance
6. **nmint**: max((total RMS intensity per pitch period)/(pitch at the max location)) for an utterance.

Sample distributions were determined for each of these parameters for each speaker in each of the two speaking conditions (sober and intoxicated).

Summary of Results. Results from direct measures of the acoustic waveform for the four talkers were as follows:

- The average pitch was higher for intoxicated speech than for sober speech. Average pitch was measured in samples; thus, the frequency of the average pitch was lower for intoxicated speech than for sober speech.
- The average pitch was more tightly clustered (i.e., varied less about the mean) for sober speech than for intoxicated speech.
- The pitch contour measure was somewhat mixed: more flat in sober speech for three speakers (DP, JB, and JS) but more flat in intoxicated speech for one speaker (KM). This result can also be observed by looking at the pitch contours in Figures 1–8. As a rule, the intoxicated and sober pitch contours had similar general shapes for a given word.
- As expected, all four of the RMS intensity measures exhibited the same behavior. Like the pitch contour measure, the results were mixed: RMS intensity was lower in sober speech for three speakers (DP, JB, and KM) but higher for the fourth speaker (JS). For speaker JS, word 15 ('chief') in the sober condition had a much higher RMS intensity than any other word.

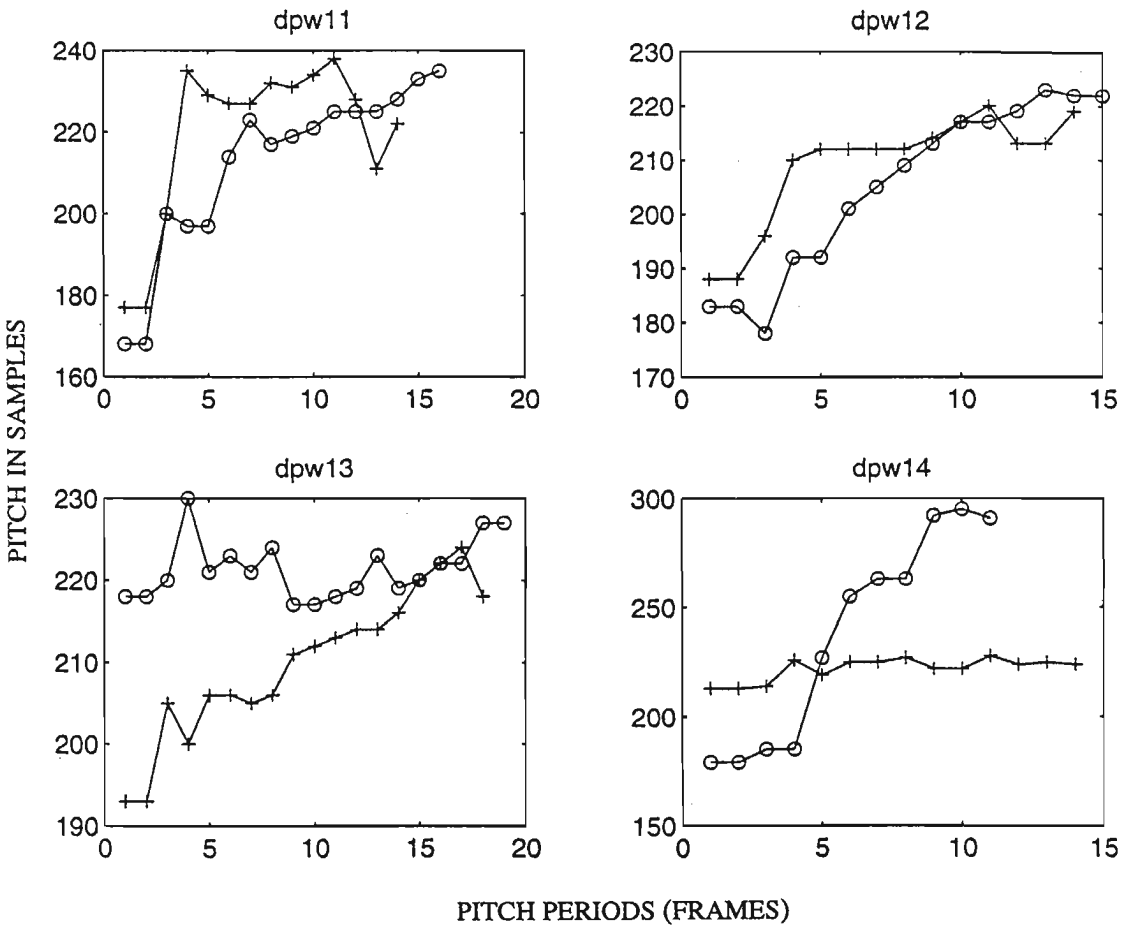


Figure 1: Sober (+) and intoxicated (o) pitch contours for subject DP for the words 'chaff' (11), 'chap' (12), 'cheese' (13), and 'chest' (14).

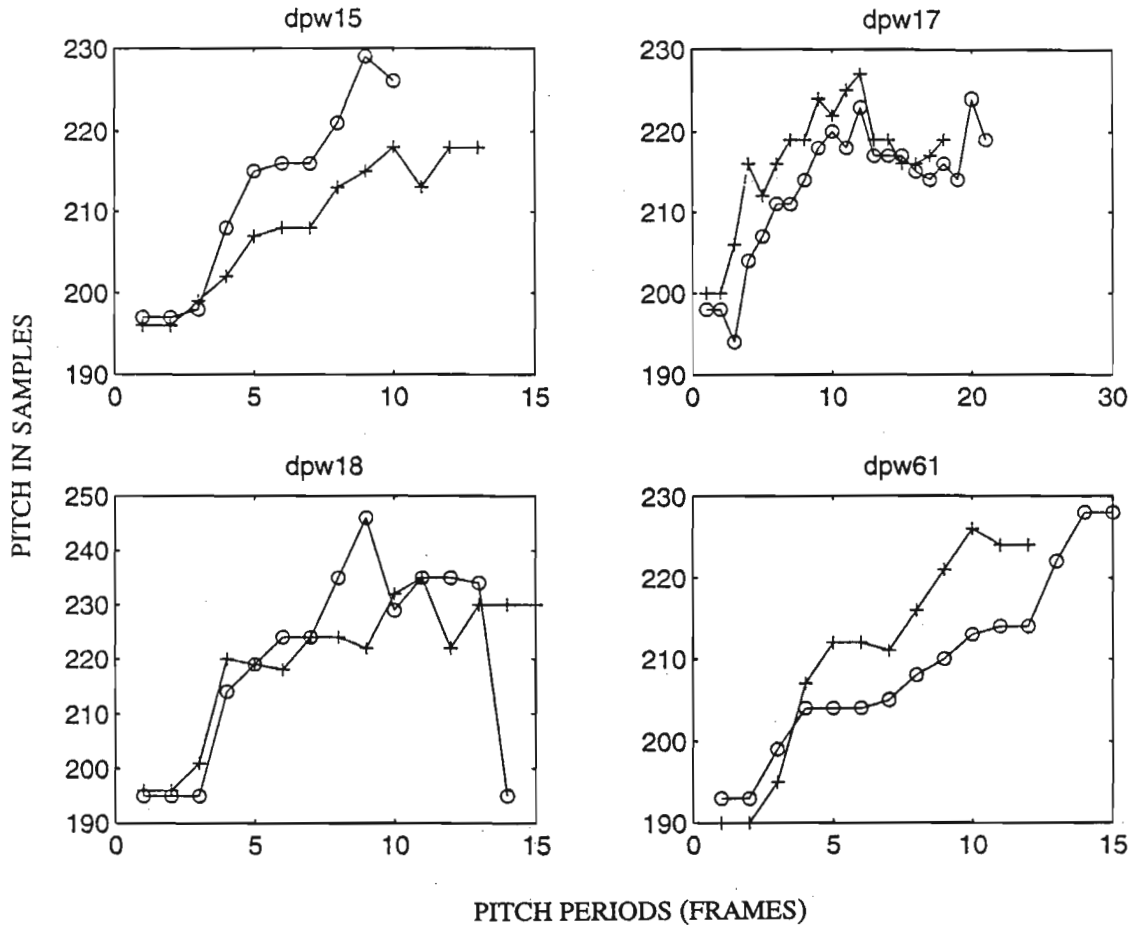


Figure 2: Sober (+) and intoxicated (o) pitch contours for subject DP for the words 'chief' (15), 'choose' (17), 'chops' (18), and 'heath' (61).

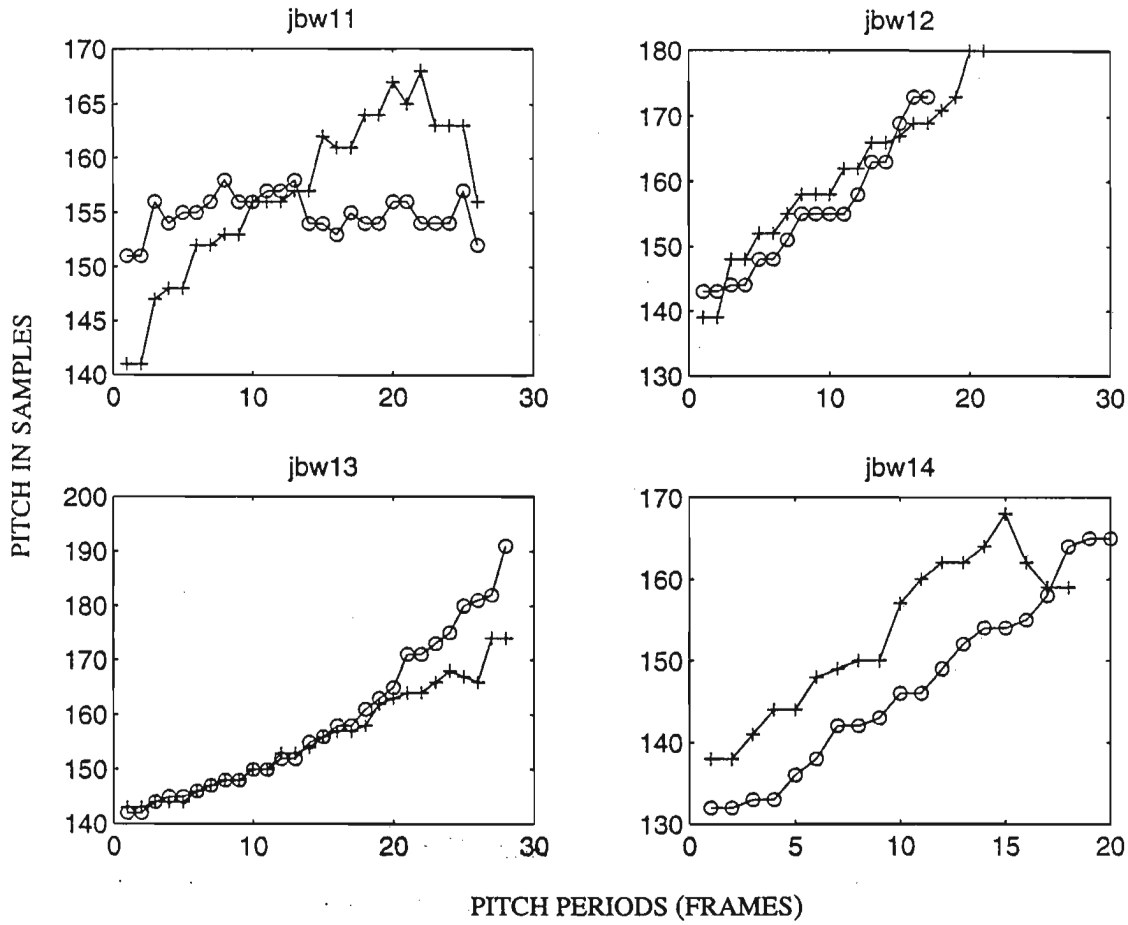


Figure 3: Sober (+) and intoxicated (o) pitch contours for subject JB for the words 'chaff' (11), 'chap' (12), 'cheese' (13), and 'chest' (14).

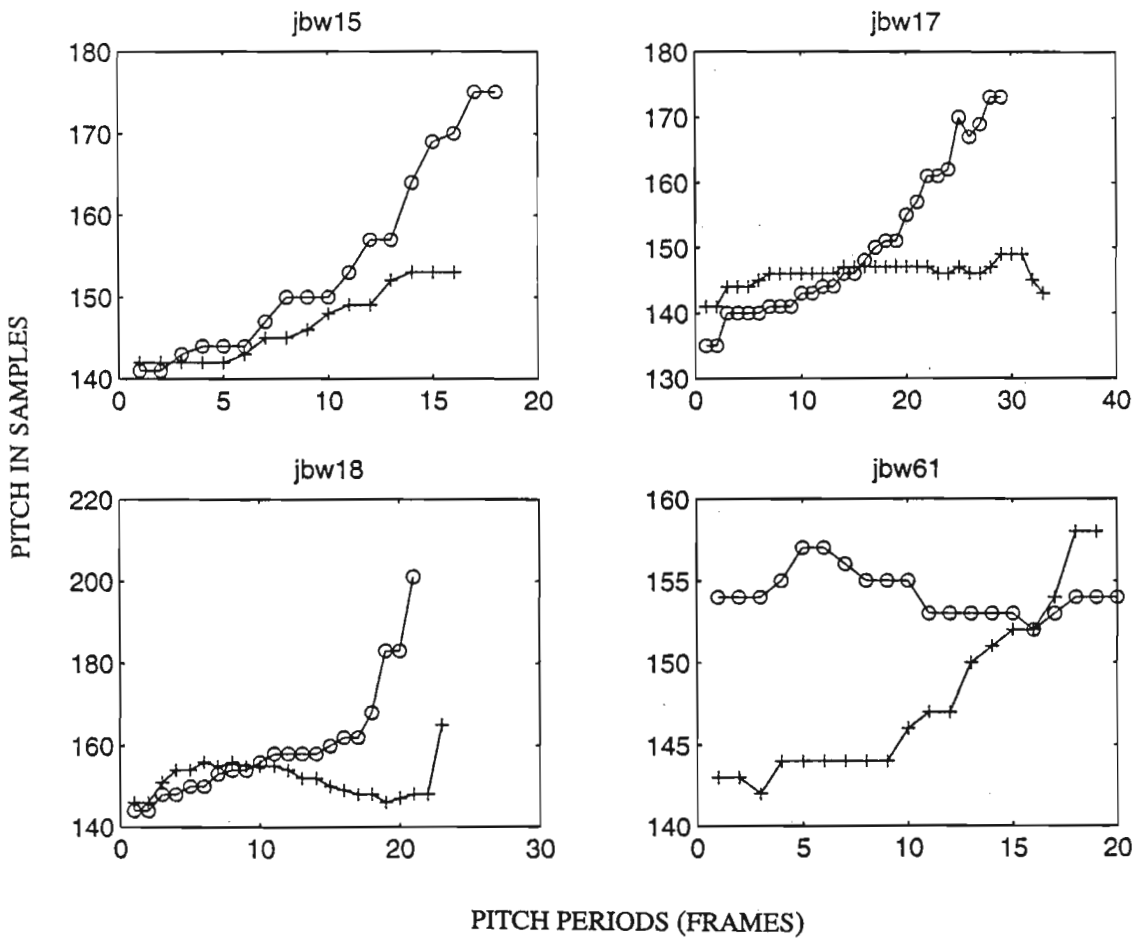


Figure 4: Sober (+) and intoxicated (o) pitch contours for subject JB for the words 'chief' (15), 'choose' (17), 'chops' (18), and 'heath' (61).

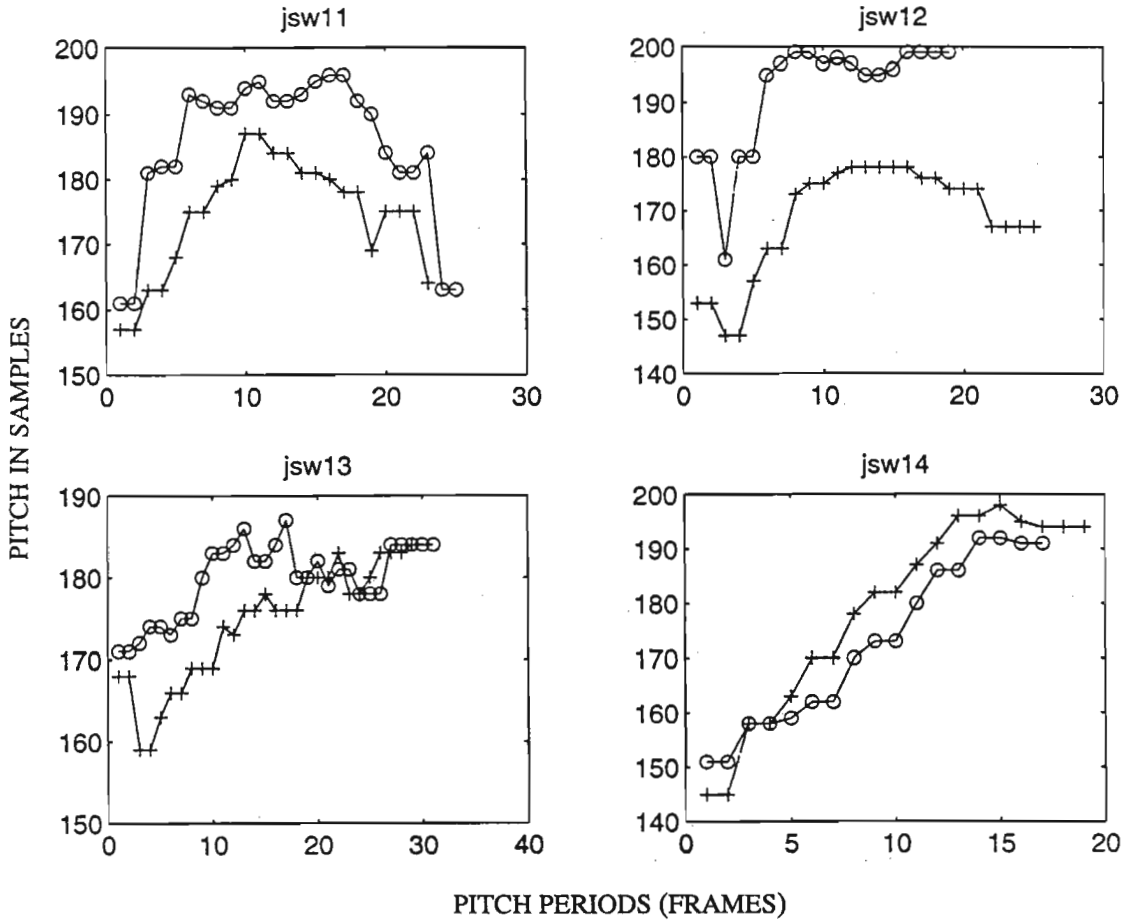


Figure 5: Sober (+) and intoxicated (o) pitch contours for subject JS for the words 'chaff' (11), 'chap' (12), 'cheese' (13), and 'chest' (14).

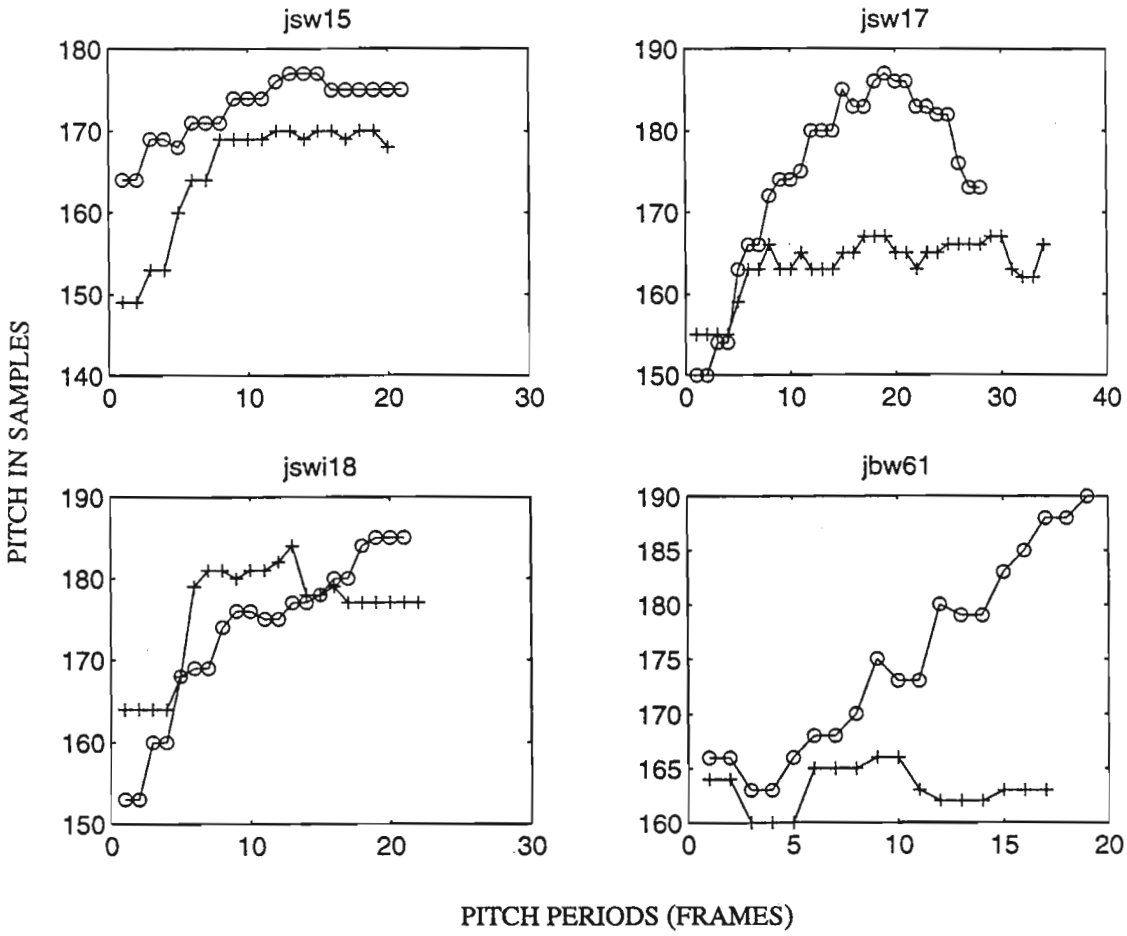


Figure 6: Sober (+) and intoxicated (o) pitch contours for subject JS for the words 'chief' (15), 'choose' (17), 'chops' (18), and 'heath' (61).

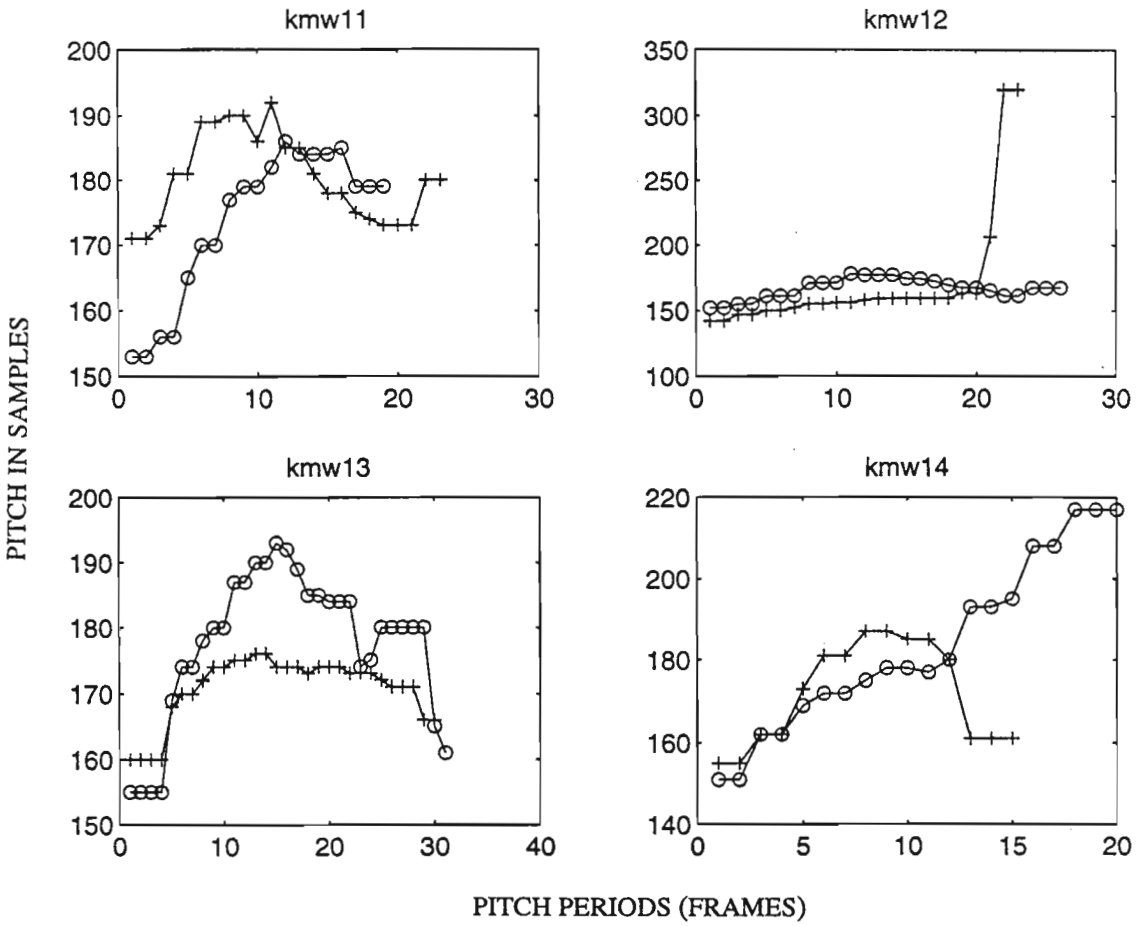


Figure 7: Sober (+) and intoxicated (o) pitch contours for subject KM for the words 'chaff' (11), 'chap' (12), 'cheese' (13), and 'chest' (14).

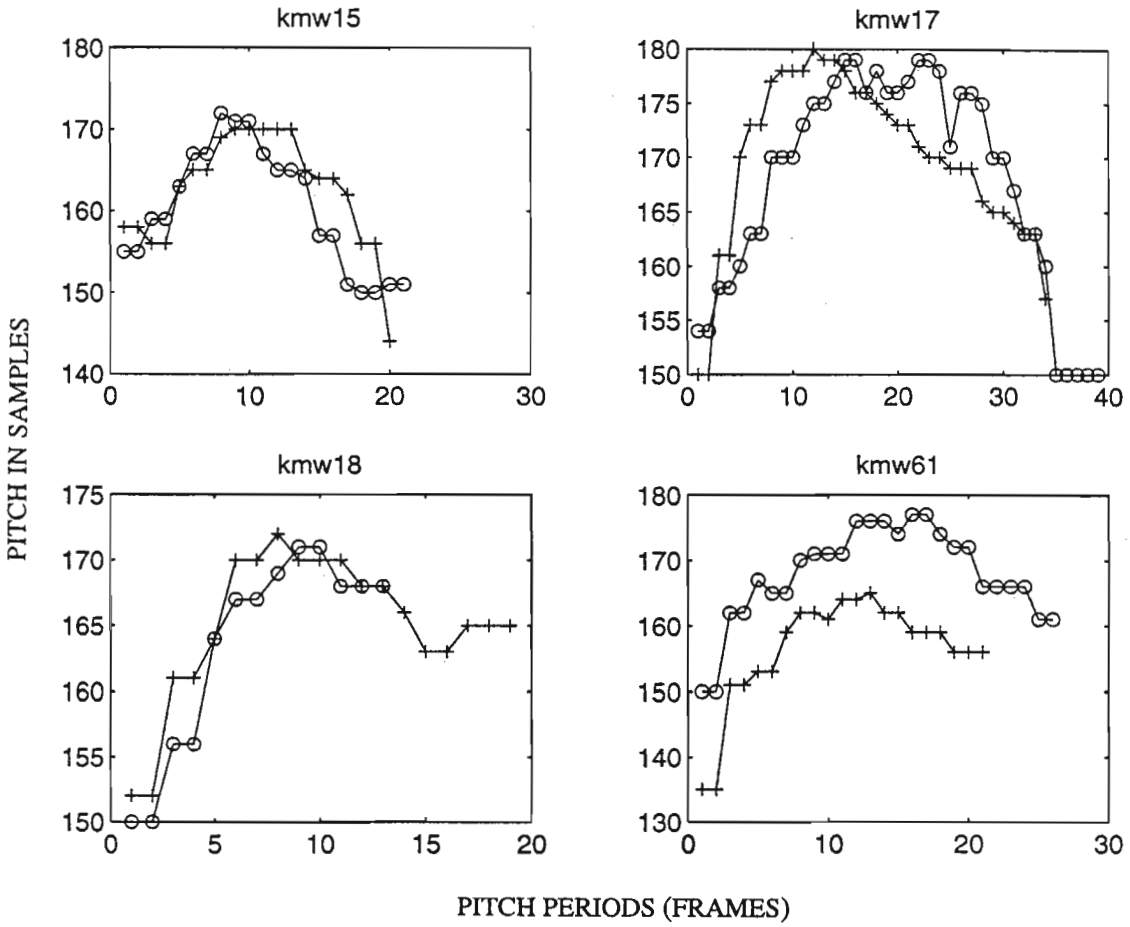


Figure 8: Sober (+) and intoxicated (o) pitch contours for subject KM for the words 'chief' (15), 'choose' (17), 'chops' (18), and 'heath' (61).

Perturbation Measures of the Acoustic Speech Waveform

In addition to direct measures of speech production such as those discussed in the preceding section, a number of perturbation measures were also extracted and examined in detail. Perturbation measures reflect the steadiness with which one produces speech. Since alcohol use affects a person's motor control (see Starmer, 1989), it is reasonable to expect perturbation measures to reflect differences in the speech production process when a person is intoxicated as compared to speech produced in the sober condition.

Initially, a series of perturbation parameters based on the six direct measures of the acoustic speech waveform discussed above were extracted. These were, generally, measures of the magnitude and direction of change in pitch or RMS intensity from pitch-period to pitch-period. The resulting parameter distributions displayed increases in variability in the intoxicated over the sober speech conditions.

A full perturbation analysis used measures based on those suggested by Pinto and Titze (1990). There are many definitions of 'jitter' and 'shimmer' in the literature (see Baken, 1987), but in a general sense, jitter is a measure of pitch frequency variability. In other words, jitter measures how much the pitch frequency changes from pitch-period to pitch-period. Shimmer is a measure of the change in energy from pitch-period to pitch-period.

In order to measure jitter and shimmer, in fact, in order to measure the changes in pitch and energy in a variety of ways, perturbation analysis was performed on three parameters:

$$\text{F0val} = \frac{F0}{\text{mean}(F0) \text{ over the utterance}}$$

$$\text{Ampval} = \frac{|\text{mean(amplitude in a pitch period)}|}{\text{mean}(|\text{mean(amplitude in a pitch period)}|)\text{over the utterance}}$$

$$\text{Intval} = \frac{\sqrt{\Sigma_{i=1}^{L} s^2(i)}}{\text{mean}(\sqrt{\Sigma_{i=1}^{L} s^2(i)}) \text{ over the utterance}}$$

where L is the length of a given pitch period.

Perturbation measures of F0val are related to jitter, while perturbation measures of Ampval and Intval are related to measures of shimmer.

Standard perturbation analysis was carried out on each of these three parameters. Zero-th, first, and second order perturbation vectors were calculated for each of the three parameters for each utterance. Letting a_i be the cyclic parameter (either F0val, Ampval, or Intval) in the ith cycle of N total cycles of the waveform

$$\bar{a} = \frac{1}{N}\Sigma_{i=1}^{N} a_i$$

$$\begin{aligned} p_i^0 &= a_i - \bar{a} && \text{for} \quad \text{i=1,...,N} \\ p_i^1 &= p_i^0 - p_{i-1}^0 = a_i - a_{i-1} && \text{for} \quad \text{i=2,...,N} \\ p_i^2 &= p_{i+1}^1 - p_i^1 = a_{i+1} - 2a_i + a_{i-1} && \text{for} \quad \text{i=2,...,N-1} \end{aligned}$$

For each perturbation vector (p^0, p^1, and p^2) for each of the three parameters (F0val, Ampval, and Intval) for each utterance, four perturbation measures were calculated. N_k is the length of the kth order perturbation vector, p^k, and k = 0, 1, and 2. These four measures are

1. MR_k – rectified mean (centroid of the histogram)

2. MER_k – median (not rectified)

3. rms_k – root-mean-squared value, where

$$rms_k = \sqrt{\frac{1}{N_k}\Sigma_{i=1}^{N_k} p^k(i) \cdot p^k(i)}$$

4. ZCR_k – zero-crossing rate (number of sign changes in p^k divided by $N_k - 1$)

The first three values are measures of perturbation extent; the last value, the zero-crossing rate, is a measure of perturbation rate.

A number of common measures of jitter and shimmer can be related to these four measures of the perturbations when the original parameter is pitch frequency or energy, respectively. For example, measures of jitter from Hollien, Michel, and Doherty (1973) and Jacob (1968) can be related to $MR^1/\bar{a}$, while Ludlow, Coulter, and Gentges's (1983) deviation from linear trend and Koike's (1973) relative average perturbation can be related to $MR^2/\bar{a}$. It is thought that the ratios rms^1/rms^0, rms^2/rms^0, and rms^2/rms^1 are related to the temporal nature of the perturbations.

Results of the perturbation analysis are shown graphically in Figures 9 through 17. Each graph compares the sample distributions for each speaker for intoxicated (I) versus sober (S) speech. The mean and plus and minus one standard deviation are shown. These include the following:

1. rms_0 – F0val
2. MR_0 – F0val
3. MER_1 – F0val
4. MER_0 – F0val
5. ZCR_0 – Ampval
6. coefficient of variation – F0val
7. direct measure of shimmer – Ampval
8. period variability index – F0val
9. rms_1/rms_0 – F0val

Insert Figures 9 through 17 about here

Summary of Results. The results involving F0val were larger than the results involving either Ampval or Intval. A number of the perturbation measures involving F0val showed the same variation from sober to intoxicated speech for all four speakers. For example, rms_0 for F0val, which is a measure of jitter rate, was consistently higher in intoxicated speech than in sober speech for all four speakers. As another example, MR_0 for F0val, which is a measure of jitter extent, was also higher in intoxicated speech than in sober speech for all four speakers. The coefficient of variation and the period variability index, two other measures of the extent of jitter, were also higher in intoxicated speech than in sober speech for all four speakers. Also, we found that one speaker, KM, showed less change in intoxicated versus sober speech for all measures.

Interestingly, the one exception to this rule was that the rms-ratio parameters (which are believed to be related to the temporal nature of the perturbations) were significantly different only for speaker KM. We are currently investigating the possibity that this speaker was a more tolerant drinker. Many of the results were not consistent across all four speakers but did show significant differences between sober and intoxicated speech.

Glottal Excitation Waveshape

The final portion of this phase of the project involved analyzing the glottal excitation waveshapes of sober versus intoxicated speech. Thus far, glottal waveforms have been extracted from two speakers (DP and JB) for two utterances. The speech was downsampled to 10 kHz prior to inverse-filtering. The glottal waveforms were then extracted using an adaptation of Wong, Markel, and Gray's (1979) closed phase glottal analysis. In this method, glottal closure is roughly identified from the covariance linear prediction

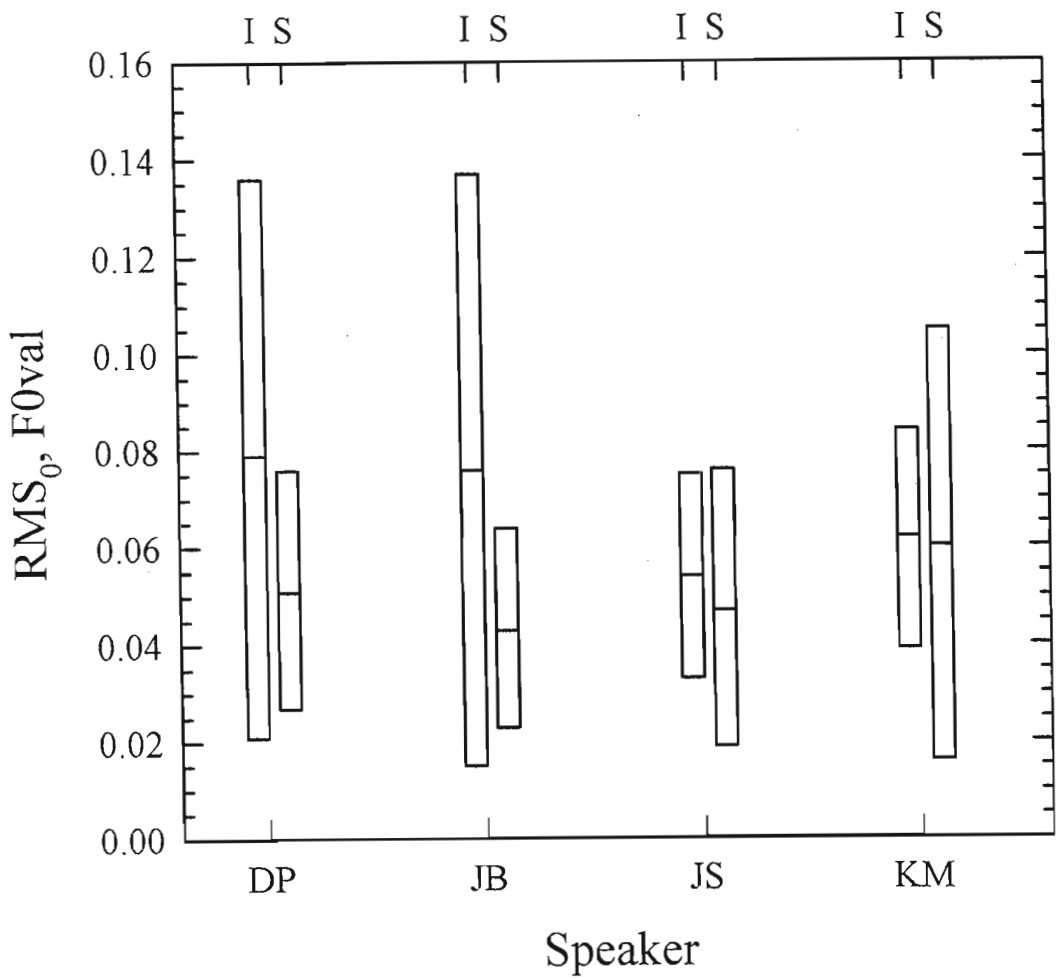


Figure 9: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure RMS_0, F0val.

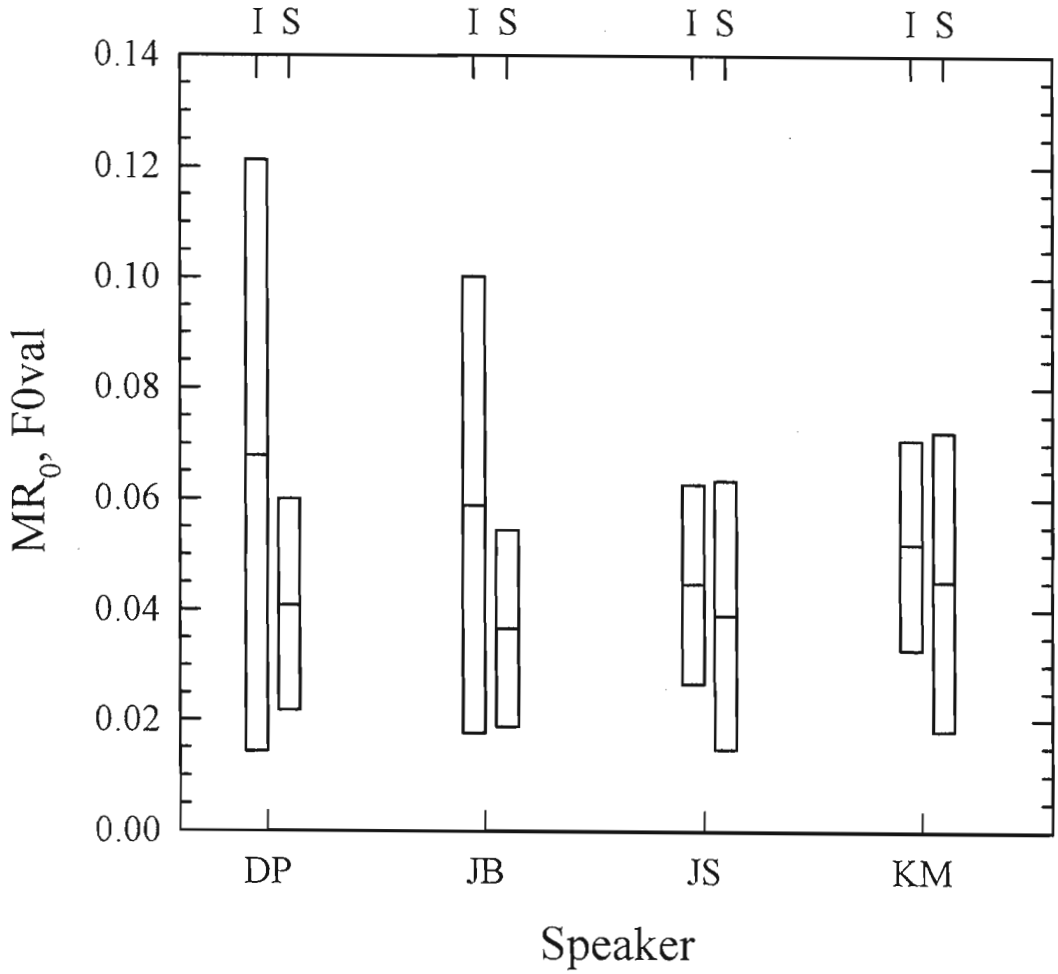


Figure 10: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure MR_0, F0val.

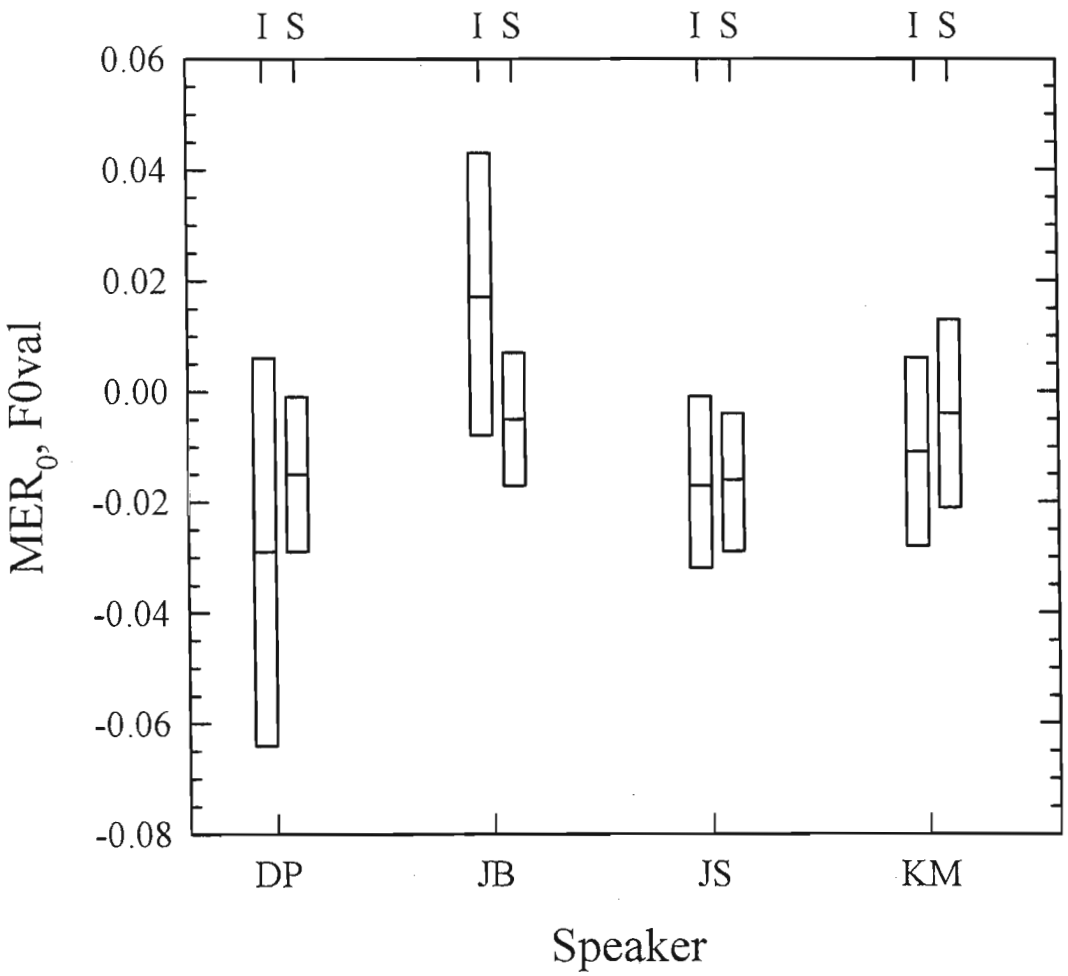


Figure 11: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure MER_0, F0val.

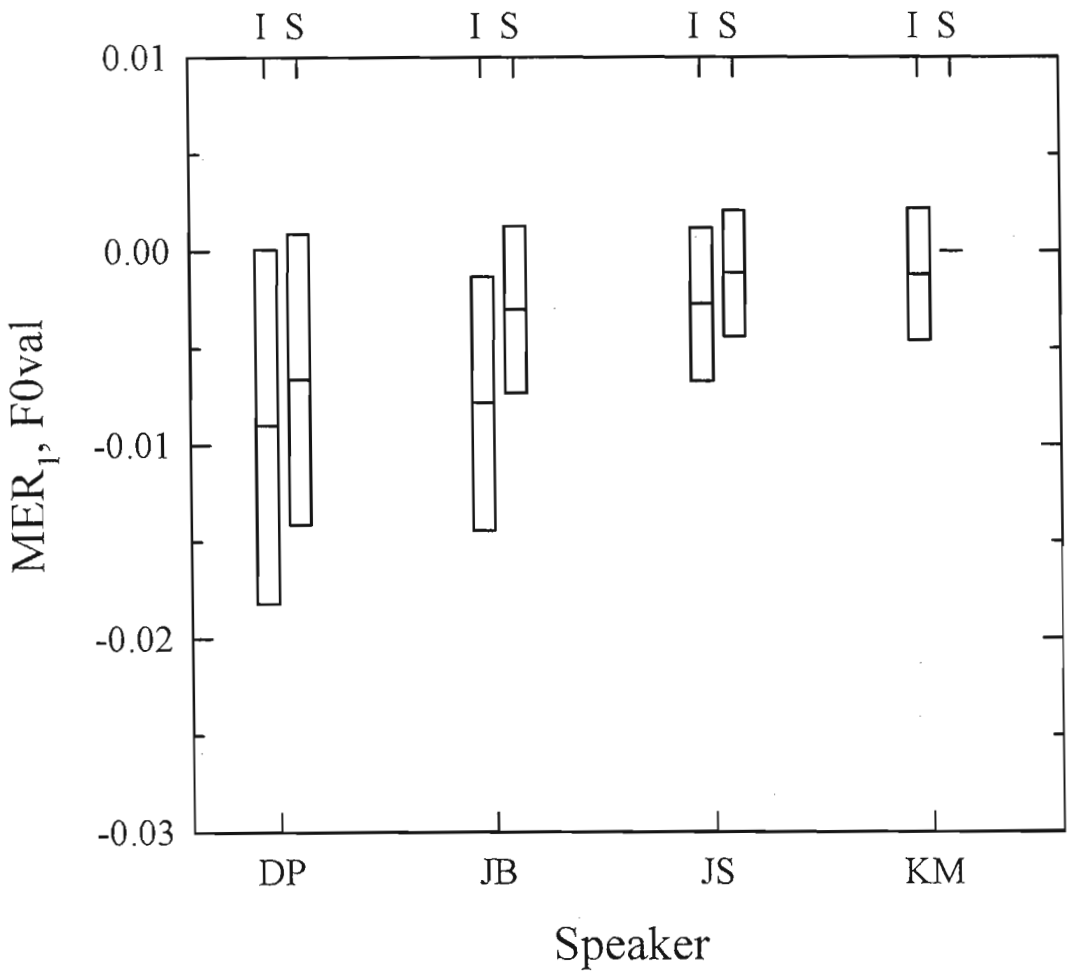


Figure 12: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure MER_1, F0val.

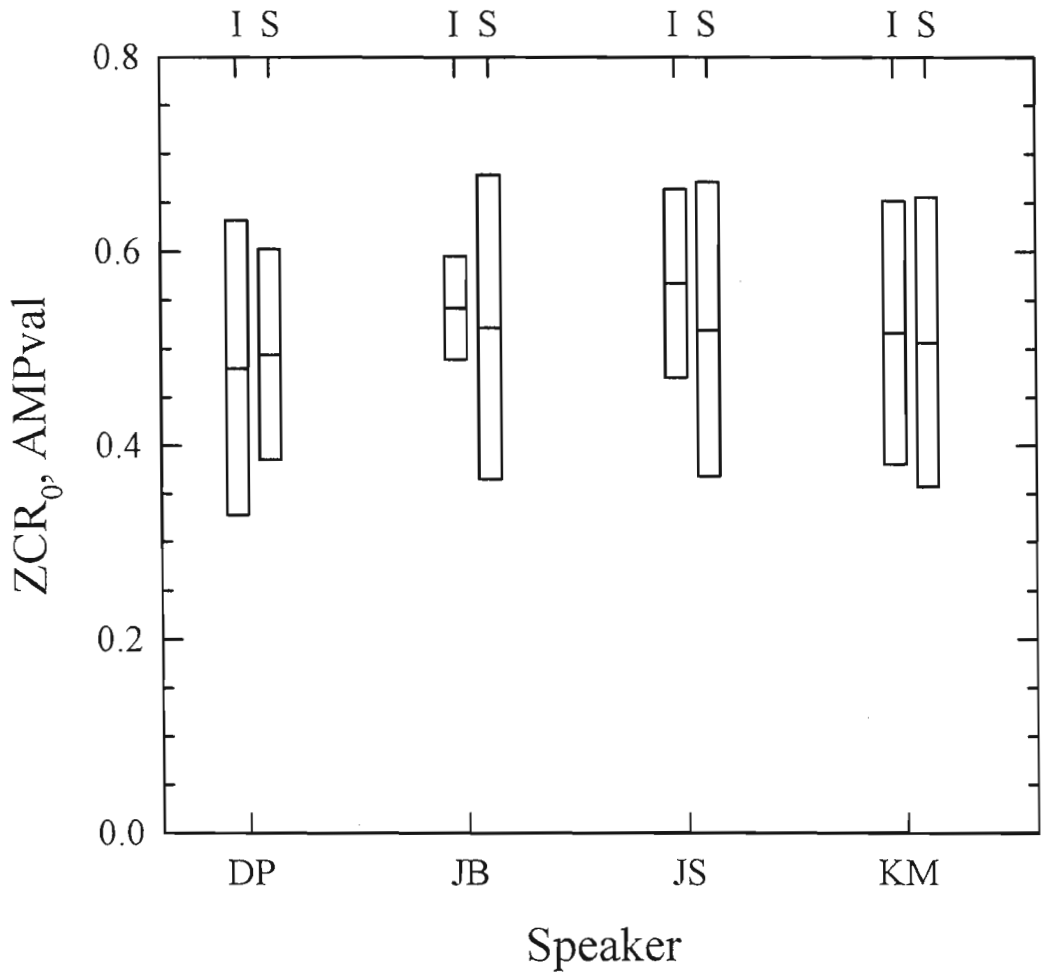


Figure 13: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure ZCR_0, AMPval.

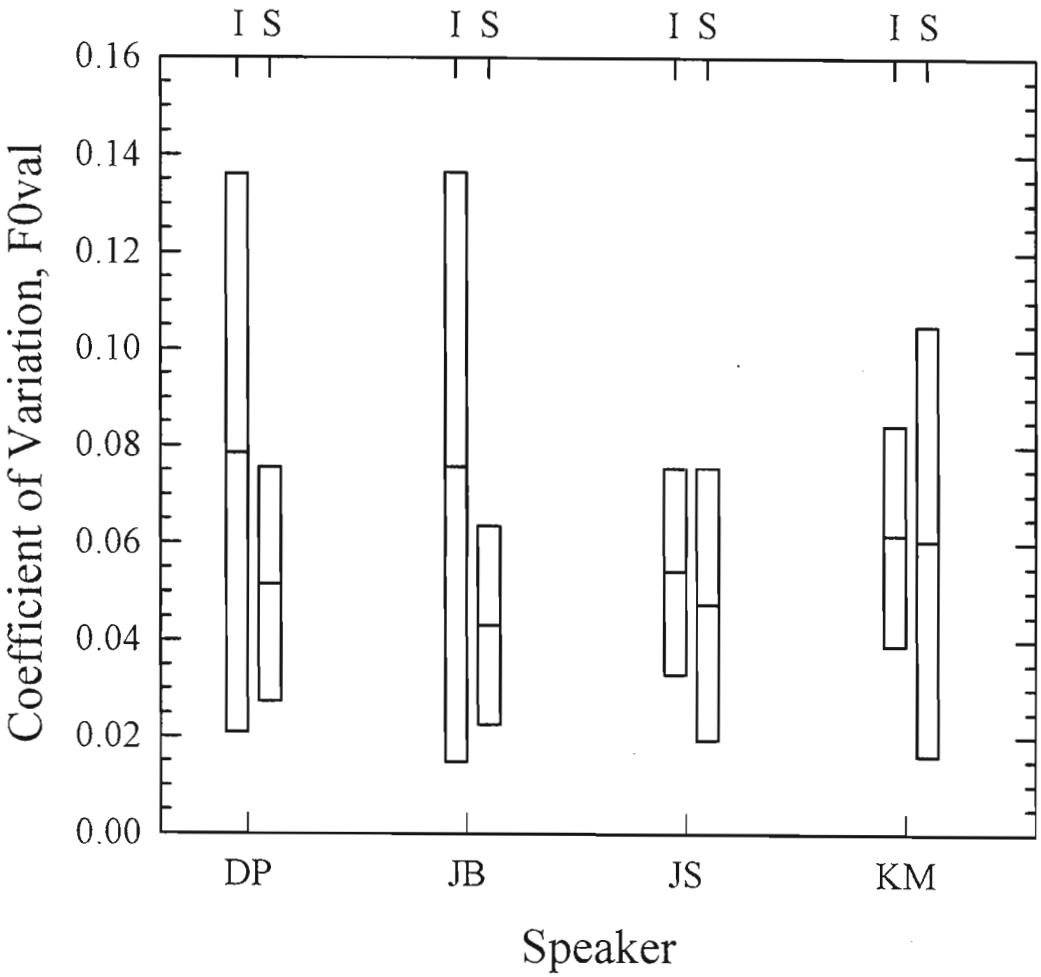


Figure 14: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure Coefficient of Variation, F0val.

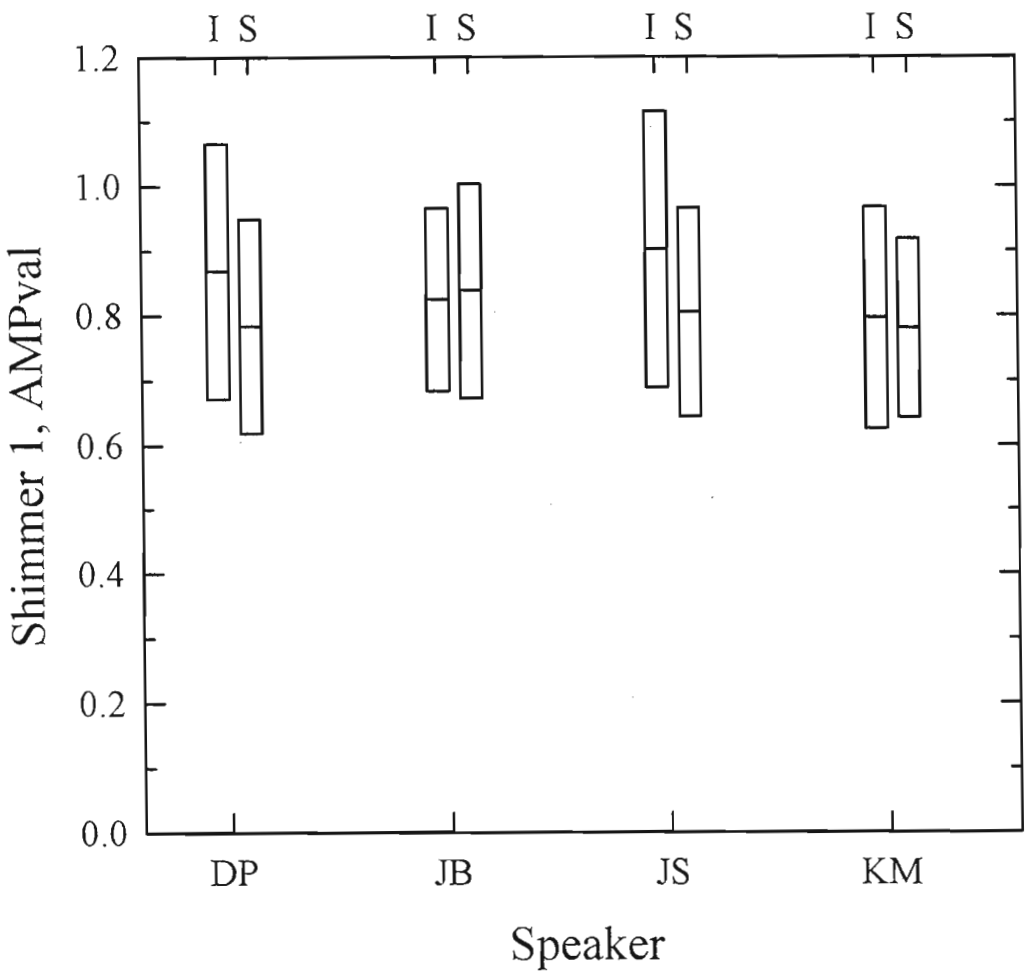


Figure 15: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure Shimmer 1, AMPval.

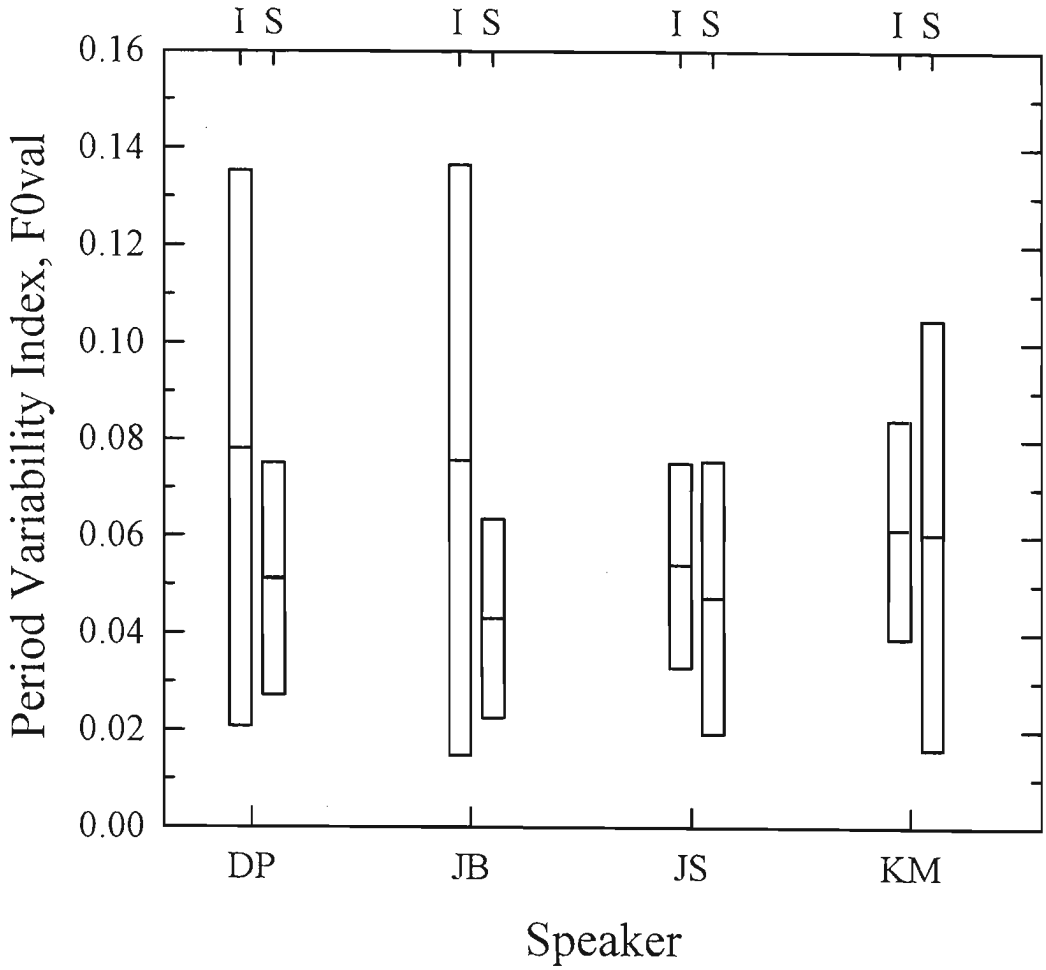


Figure 16: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure Period Variability Index, F0val.

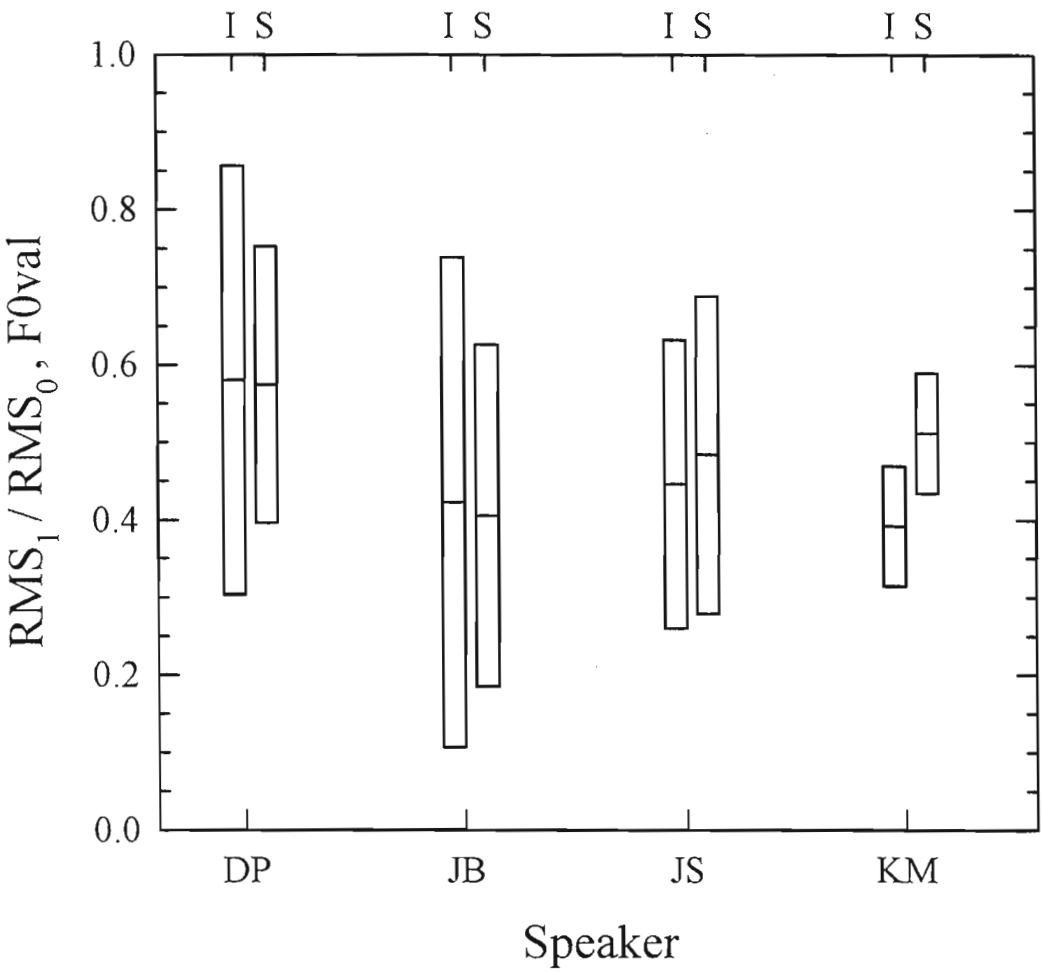


Figure 17: Sample distributions (mean and ± one standard deviation) for 4 speakers in intoxicated (I) and sober (S) conditions for perturbation measure RMS_1/RMS_0, F0val.

analysis error waveform. Sample-by-sample covariance LP modelling is performed on the speech waveform using a segment of speech that is shorter than the expected length of glottal closure. When the error is closest to zero, the speech segment most closely resembles an all-pole model. This segment is assumed to include only the vocal tract resonances; thus, the glottis is closed during this segment.

Once a segment of speech during which the glottis is closed is identified, a vocal tract model is determined using 12-pole covariance LP analysis. Ten of the poles represent the ten poles of the vocal tract model. The other two are the poles of the two-pole, two-zero model of radiation at the lips. The speech is inverse-filtered with the 12-pole model and with a two-zero filter representing the two zeros in the radiation model. The resulting waveform is assumed to be a representation of the glottal excitation waveform.

The z-Transform representation is

$$G(z) = \frac{S(z) \cdot (1 - \Sigma_{k=1}^{10} a_k z^{-k})(1 - \Sigma_{l=1}^{2} \alpha_l z^{-l})}{1 - \Sigma_{j=1}^{2} \beta_j z^{-j}}$$

where

a_k = LP coefficients of the vocal tract model

α_l, β_l = coefficients of the polynomials representing the two-zero, two-pole model of radiation at the lips

$S(z)$ = z-Transform of the windowed speech signal

$G(z)$ = z-Transform of the glottal excitation waveform.

Some examples of extracted glottal waveforms are shown in Figure 18. Several qualitative observations can be made concerning the differences between sober and intoxicated glottal excitation waveforms, as indicated in the following section.

Insert Figure 18 about here

Summary of Results. Results from the glottal excitation waveshape analysis for two talkers were as follows.

- The vocal tract is less stationary in intoxicated speech than in sober speech. The LP vocal tract model can only be used to inverse-filter about four pitch priods of intoxicated speech. After that, the model is no longer accurate enough to generate a " clean" glottal waveform. With sober speech, a given LP model accurately represents the vocal tract for several pitch periods (often as many as eight or nine).

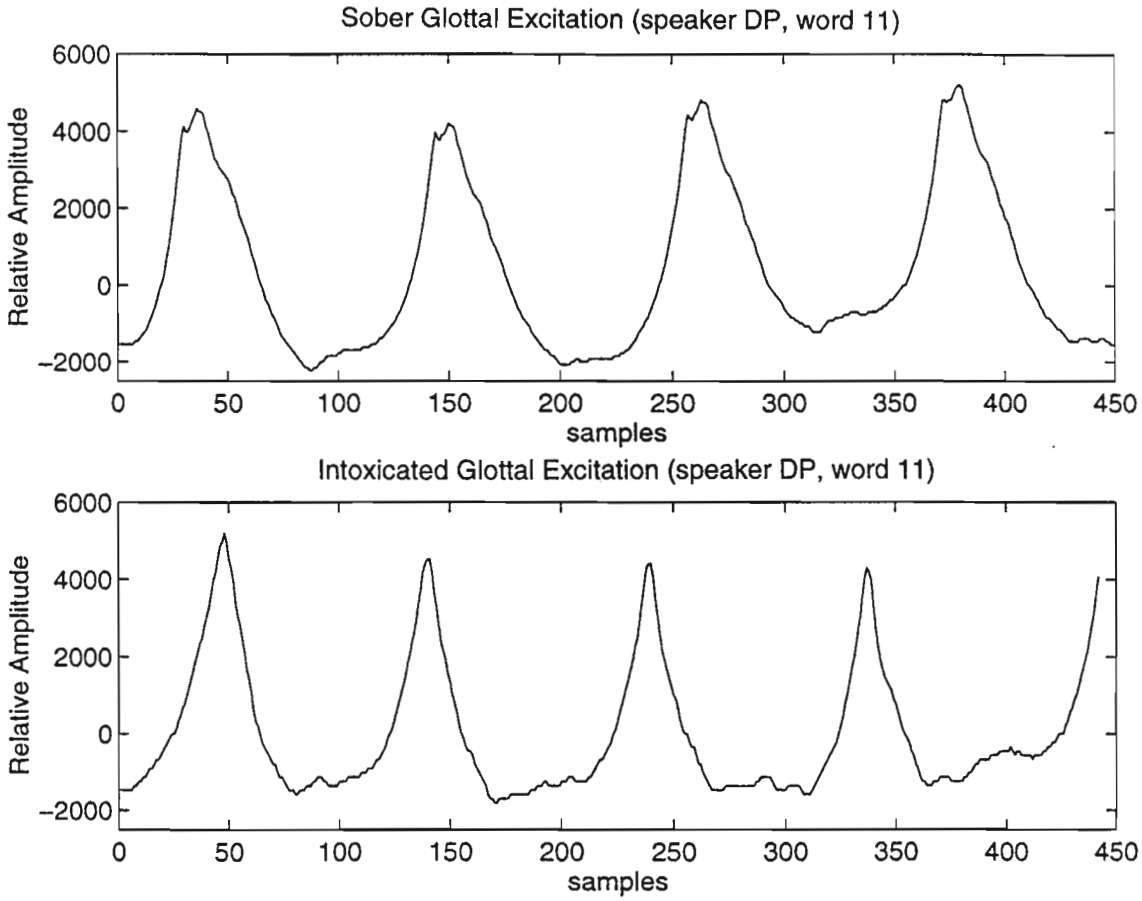


Figure 18: Examples of extracted glottal waveforms.

- The glottal waveshape is less consistent through an utterance in intoxicated as opposed to sober speech. More variance in amplitude, slopes, durations, and overall waveshape are seen in the intoxicated glottal waveforms than in the sober glottal waveforms.

- The glottal pulse shape appears to be more symmetric and more triangular in shape for intoxicated speech than for sober speech.

- Glottal closure appears to have more ripple in intoxicated speech than in sober speech.

Future Work

The work remaining on this project include the following.

- Direct measures of the acoustic speech waveform:

  Extract and analyze these parameters for the remaining five speakers in the alcohol database.

- Perturbation measures of the acoustic speech waveform:

  Add a shimmer parameter based on the peak amplitude within a pitch period.

  Extract and analyze the perturbation measures for the remaining five speakers in the database.

- Glottal excitation waveshape:

  Continue to extract glottal waveforms from all nine speakers (do the four speakers DP, JB, JS, and KM first).

  Parameterize and mark all glottal waveforms for analysis.

  Calculate and compare the cepstra of sober versus intoxicated speech.

- Statistical Analysis:

  Based on means and standard deviations, perform statistical analysis to determine which parameters are significantly different in a statistical sense.

  Develop a parameter set that can be used to distinguish between sober and intoxicated speech.

References

Baken, R.J. (1987). *Clinical Measurement of Speech and Voice*. Boston, MA: College-Hill Press.

Cummings, K.E. (1992). *Analysis, Synthesis, and Recognition of Stressed Speech*. Ph.D. dissertation, Georgia Institute of Technology.

Hollien, H., Michel, J., & Doherty, E.T. (1973). A method for analyzing vocal jitter in sustained phonation. *Journal of Phonetics*, **1**, 85–91.

Jacob, L. A. (1968). A normative study of laryngeal jitter. Master's thesis, University of Kansas.

Koike, Y. (1973). Application of some acoustic measures for the evaluation of laryngeal dysfunction. *Studia Phonologica*, **7**, 17–23.

Ludlow, C., Coulter, D., & Gentges, F. (1983). The differential sensitivity of measures of fundamental frequency perturbation to laryngeal neoplasms and neuropathologies. In D.M. Bless & J.H. Abbs (Eds.), *Vocal Fold Physiology: Contemporary Research and Clinical Issues* (Chap. 33, pp. 381–392). San Diego, CA: College-Hill Press.

Pinto, N. B., & Titze, I. R. (1990). Unification of perturbation measures in speech signals. *Journal of the Acoustical Society of America*, **87**, 1278-1289.

Pisoni, D.B., & Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*, **13**, 577–587.

Pisoni, D.B., Yuchtman, M., & Hathaway, S.N. (1986). Effects of alcohol on the acoustic-phonetic properties of speech. In *Alcohol, Accidents, and Injuries* (pp. 131–150). Warrendale, PA: Society of Automotive Engineers.

Starmer, G. A. (1989). Effects of low to moderate doses of ethanol on human driving-related performance. In K.E. Crow & R.D. Batt (Eds.), *Human Metabolism of Alcohol, Volume I: Pharmacokinetics, Medicolegal Aspects, and General Interest* (pp. 101–130). Boca Raton, FL: CRC Press.

Wong, D., Markel, J., & Gray, A., Jr. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP–27, pp. 350–355.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Current Computer Facilities in the Speech Research Laboratory[1]

Luis R. Hernández

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

[1] Equipment and software development supported, in part, by NIH-NIDCD Research Grant DC-00111, and, in part, by NSF Research Grant IRI-86-17847.

Current Computer Facilities in the Speech Research Laboratory

Abstract. This report describes changes and new developments of the computer software and hardware in the Speech Research Laboratory for the period from 1990 through 1995. Our computer enviroment has slowly changed from a mainframe to a distributed architecture. Workstations in a Client/Server environment, configured to perform specific tasks have become better and more economical tools than our older Mainframe/Teminal configuration. We outline here the core of our computing facilites and describe some of the more important computer software. New systems descibed include the Local Area Network, a new Perceptual Testing System, a Speech Acquisition Program, new stimulus preparation and signal processing capabilities, and a cross-modal perceptual testing system.

The Speech Research Laboratory at Indiana University has created a unique environment for the study of virtually any aspect of speech. We have expanded and improved many areas of our computer facilities in a continuing effort to provide the necessary tools to conduct and explore this research. Because many changes have occurred since our last report (Bernacki, Feaster, Hernandez, & Forshee, 1989), we will briefly detail here some of the more important hardware and software development efforts. Additional information on specific aspects of any of these systems may be obtain by contacting our laboratory.

The major change in our computing environment has been the migration from a DECnet networked VAX/VMS mainframe to a local area network (LAN) supported by a central Novel Netware server with personal computers as clients. As a result of this move, we can now take advantage of the personal computer for word-processing, statistical analysis, e-mail, and other desktop applications. This change has allowed our researchers to configure and tailor their personal computers according to their specific needs. In addition, our workstations can devote more of their processing time to their specific tasks, since they no longer need to support the load of personal accounts.

Although we continue to support our VAX/VMS workstations for speech processing and stimulus preparation, we have expanded our capabilities by introducing two UNIX workstations. The new workstations are equipped with analog-to-digital converters that allow digitization from various sources. In addition, we have installed software to perform more complex speech processing analysis than what is available on our VAXes.

Over the past few years, we have developed a PC-based perceptual testing system (PTS) that has replaced our PDP-11/34 (Hernández, 1994). We are currently implementing the last phase of this project (Hernandez, Carrell, Reutter & Bernacki, 1992) and hope to have the system completed by the end of 1996. Various other software utilities developed in this lab for the purpose of digitizing and preparing stimuli have been ported or rewritten to accommodate new file formats and hardware.

A new direction in research has led us into the development of an auditory and visual experimental control system (see Hernández & Marcinkovich, 1996 (this volume)). This system gives us the ability to present digital audio and visual cross-modal stimuli and record reaction times very accurately.

Local Area Network (LAN)

The most significant change in the lab since our last report is the implementation of a Novell Netware LAN. Each person's personal computer can be networked and can perform work much like a non-networked computer. An important difference, however, is that laboratory staff can access files from more than just their local drives while still running the operating system of their choice. The Netware server coordinates all of the computers and regulates the way they share network resources.

Currently there are two servers. One handles everyday desktop needs, and the other handles file-sharing for our PTS system (Hernández, 1994). The desktop server is a 486 33MHz PC-compatible with 16MB of RAM and 1.2 gigabytes (Gig) of disk-space running Netware 3.11. Users can access this server from their desks using either PC-compatibles or Macintosh computers. Other clients are several specially equipped personal computers located throughout the lab. The desktop server also supports three laser printers and a DAT backup device available to anyone in the lab. Commercially available PC and Macintosh software is licensed and maintained centrally, making it easier to keep track of and update. Electronic mail and internet access are also available from each client. Machines that are not directly connected to the server can still access its files using FTP.

The other server is a 100Mhz Pentium computer with 32MB of RAM and 1.2 Gig hard disk and CD-ROM. This server acts as a file-sharing system for our PC and Macintosh perceptual testing system. It runs Novell Netware 4.11 and supports 32-bit addressing from its clients. Experiments that require accurate measures of timing intervals are unable to use stimuli from the LAN directly. In these instances, files need to be copied to each individual client.

Perceptual Testing System (PTS)

This PC-based experiment control system is a replacement for the aging PDP-11/34-based real-time systems used in our laboratory (see Forshee & Nusbaum, 1984). The old system has become virtually nonexistent in terms of both hardware and software. Furthermore, the architecture is 25 years old, and by modern standards, it is an extremely difficult environment in which to develop robust applications.

The current computers supporting our PTS configuration are 133MHz Pentium PC-compatibles with 16MB of RAM and 1.2 Gig hard disks. Each computer contains a SoundBlaster16AWE sound board, a timer board, a specialized parallel port and ethernet. A set of routines has been developed to control timing, presentation, and input into this system to perform perceptual experimental paradigms. For a detailed description of the hardware and software components of the set-up, see Hernández (1994).

Speech Acquisition Program (SAP)

To maintain compatibility with our new PTS hardware, we have ported a version of the Speech Acquisition Program (SAP: Dedina, 1987) to run on PCs. This program digitizes utterances into individual files under benign or manipulated speaking environments. The new program provides expanded visual and auditory cues; pictures in different graphics formats as well as auditory and character string cues are now available. Also, the program has the ability to measure reaction times from other inputs such as button-boxes or voice-activated keys with measures accurate within 3 msec. As the recordings are being made, the experimenter can monitor and control the presentation of cues. The program is able to repeat or skip cues on demand.

The hardware consists of a PC-compatible 486 with 8MB of RAM and 650MB of disk space. High quality 16-bit stereo recording is achieved with Tucker-Davis Technologies (TDT) System II hardware that can achieve sampling frequencies of up to 170kHz. The recordings are done in an IAC booth. In the booth, cues are presented with a 15-inch non-interlaced SVGA monitor and Beyerdynamic DT100 headphones. The microphone used is a Shure SM98.

Stimulus Preparation and Signal Processing

Our current VAX/VMS environment running the ILS package for signal processing and preparation has been useful over the years, but is also difficult to work with because of the lack of a good graphical user interface (GUI). To expand our capabilities, we have acquired two Sun Microsystems SPARCstation 5 UNIX workstations loaded with the Entropics Waves+/ESPS software package, giving us a broad range of analysis capabilities. The machines are equipped with 32MB of RAM and 1.2-Gig hard disks, using 19-inch color monitors and CD-ROMs. Also, each has a 16-bit 44.1kHz A/D and D/A converter for digital sound capabilities.

In order to take full advantage of this new hardware, we have began to update and port various in-house software packages previously developed on our VAX/VMS computers written in C and Motif. Among the programs ported, the most important one is Mwaves (Motif Waveform Editor). This program provides a user-friendly environment for precise and speeded waveform editing with improved abilities to tailor the entire display layout to specific needs. Editing capabilities have been expanded to add noise, insert silence, generate tones, measure waveform lengths, scale waveform amplitudes, and perform a few simple analyses. The modular design of this software allows us to easily add processing and analysis tools.

Cross-Modal Perceptual Testing System

Because of an interest in experimental perceptual paradigms that would present audio and visual (A+V) cross-modal stimuli with control over stimulus presentation and accuracy over latency measures, we have developed a digital system that supports such paradigms. The system was developed on a Macintosh Quadra 950 because of Apple's long history of robust A+V capabilities. A Radius VideoVision board is used to achieve real-time video capable of playing full-screen QuickTime movies, and a 40-channel input/output (I/O) card from Strawberry Tree is used for timed button-box response. The programs were compiled using CodeWarrior C compiler (see Hernández & Marcinkovich, 1996 (this volume), for further details).

Following our philosophy of maintaining compatibility with other laboratories, we have also constructed a system that supports stimulus presentation from various databases available on Laserdisc. This system has total control over the stimulus presentation but does not have the ability to measure reaction times. It consists of a 486-PC with 8MB of RAM and 250MB of disk space connected to a Pioneer LD-V4400 Laserdisc via the serial port. The output of the Laserdisc is routed to a Panasonic CT-2084 color video monitor for stimulus presentation while the audio is presented through Beyerdynamic DT-100 headphones. A collection of routines were developed in Visual C/C++ for ease of programming and software maintenance.

Summary

In summary, as technology and reasearch demands have changed over the years, so has the sophistication of computer software and hardware resources. Ideas and techniques are constantly being generated and re-evaluated in this type of enviroment, and research tools are continuously challenged. To meet these demands, we combine commercially available tools with in-house programs and spend a considerable amount of effort develping modular reusable code on each hardware platform to reduce development time.

References

Bernacki, R.H., Feaster, D.M., Hernandez, L.R., & Forshee, J.C. (1989) Current Computer Facilities in the Speech Research Laboratory. *Research on Speech Perception Progress Report No 15*, Indiana University, Bloomington, IN, pp. 505-512.

Forshee, J.C., & Nusbaum, H.C. (1984). An update on computer facilities in the Speech Research Laboratory. *Research on Speech Perception Progress Report No. 10*, Indiana University, Bloomington, IN, pp. 453-462.

Hernandez, L.R., Carrell, T.D., Reutter, J.G. & Bernacki, R.H. (1992). A new PC-based real-time experiment control system. *Research on Spoken Language Processing Progress Report No. 18*, Indiana University, Bloomington, IN, pp. 243-253.

Hernandez, L.R., Marcinkovich, C.R. (1996). An Auditory and Visual Experimental Control System. *Research on Spoken Language Processing Progress Report No.20*, Indiana University, Bloomington, IN, pp. 395-402 (this volume).

Hernandez, L.R. (1995). Implementation of a PC-based perceptual testing system (PTS): A first milestone. *Research on Spoken Language Processing Progress Report No. 19*, Indiana University, Bloomington, IN, pp. 321-328.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

An Auditory and Visual Experimental Control System[1]

Luis R. Hernández and Caleb R. Marcinkovich

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

[1] The development of this system was supported by NIH-NIDCD Research Grant DC00111 to Indiana University-Bloomington, IN. The authors would like to thank Jim Sawusch for his contribution and assistance in developing this software.

An Auditory and Visual Experimental Control System

Abstract. This report describes the development of a new multi-modal auditory and visual (A+V) testing experiment system. We outline the hardware requirements and describe the implementation of a system that can present full-motion video and record accurate reaction times from subjects responses. A Macintosh equipped with a video board and an analog input/output board was selected as the preferred platform. A set of general purpose high-level library calls were written to simplify the implementation of future experimental control programs.

Overview

The goal of the project was to design and implement a set of experimental perceptual paradigms that would present digital audio and video cross-modal stimuli. Some examples of the paradigms that we had in mind are identification, discrimination, recognition memory, recall and lexical decision tasks. Previous experiments have been carried out using video tape and in some instances Laserdiscs. However, this traditional analog media limits control over stimulus presentation and accuracy over latency measures. Furthermore, we wished to assemble a collection of our own movie-playing, randomization and event-timing routines as a high-level library set to provide other programmers convenient tools that would facilitate future development. An experimenter's ability to precisely control the onset and offset of stimuli presentation, as well as, control over timing between events is crucial to the success of any experiment. All stimulus events are recorded on a log file.

Stimuli consist of a full-motion video of a face articulating words or sentences. These stimuli could be edited and modified according to the needs of the particular experiment. We are planning to develop the ability to edit and modify the audio and the video tracks of the digitized movies independently of each other, maintaining the ability to synchronize the tracks. For this report, we will not discuss how the stimuli are digitized and edited, but will refer to stimuli as files that consist of segmented preprocessed movies of different duration varying from two to six seconds.

Design Criteria

With these goals in mind, we have decided on a number of design criteria. The system should rely, as much as possible, on standard, off-the-shelf components so that all users may take advantage of competitive pricing as well as simplified maintenance. The system should also be designed in such a way that it can be implemented inexpensively for simple, frequently-used experimental paradigms in our research program. However, as the sophistication and resolution of the experiments increases, the cost of additional equipment will increase as well. The system should be easy to maintain and manage, preferably benefiting from a graphical user interface (GUI). Although it is not our immediate goal, software should be as transportable as possible, and hardware components should be available in different platforms.

Functional Requirements

The principal function of our system is to play full-screen, full-motion video at NTSC standards. This usually means 30 frames per second (fps) at a resolution of 24-bit color on a 640 by 480 resolution monitor. We would like to display the stimuli as close to life size as possible while maintaining the resolution desired. The video image needs to be computer-controlled as precisely as possible in order to

capture millisecond accuracy between any two chosen events. The audio requirements are 16-bit resolution with sampling rates up to 44.1kHz synchronized with the video track. Given that the video is playing at 33.3 msec cycles, we would expect the latencies to be only this accurate. But we would still want to maintain a one millisecond resolution for audio-alone experiments and other non-video events.

The types of subject responses that the system needs to handle are: (1) handwritten, (2) keyboard, and (3) button-presses. For handwritten input, the system needs only to play the video at a fixed inter-stimulus interval (ISI) specified by the experimenter. Keyboard input constitutes a string response prompt with a time-out period after the stimuli are presented. Finally, a button-press response needs to capture which button was pressed and accurately measure time between certain events (e.g., stimulus onset to button press). The response boxes also need to have feedback lights above each button, as well as a cue light in the middle top portion of the box. Button responses will be signaled on the down press only, with a short travel distance. Button size is not critical, but will be consistent. For experiments in which latency is measured, keyboard responses will not be acceptable due to small, unpredictable variations in response time. Moreover, mouse responses, voice-operated switches, and touch-screen inputs will never be used to measure latency. A log file needs to record randomization order, current trial, key of correct responses, subject's response, and response latency. Each subject will have control over the pace of the experiment and perform the task independent of other subjects.

Modular software design is essential to insure well-written code, ease of support, ease of maintenance, portability and reusability of our libraries. The software library will have at a minimum, capabilities analogous to those of the existing Perceptual Testing System (PTS) calls (see Hernández, 1995). However, the routines will require extended capabilities given that they do not support video.

System Configuration and Components

We chose the PowerPC 8100 Macintosh platform for the development and implementation of these applications because of Apple's long history of robust A+V capabilities. The system would benefit from the inherited designed GUI providing an easily navigable setup-and-run environment. 36MB of RAM provides enough memory to maintain the throughput required. Since compressed digitized movies of this resolution require approximately 4MB (27MB uncompressed) of storage space per second, we have equipped this computer with a 5 gigabytes (GB) hard disk. A 17-inch Trinitron tube non-interlace monitor was chosen for the display.

To achieve real-time video capable of playing full-screen QuickTime movies, we decided on the Radius VideoVision board. The Radius board provides flicker-free, smooth playback for full-screen, playing 30 frames per second (fps), 24-bit color movies using special compression and decompression (codec) algorithm. This particular board can play and record.

The input/output (I/O) component utilizes the Strawberry Tree ACM2-I/O analog I/O card for timed button box response. This board contains 40 I/O lines and six counter/timers. We utilize two of the six hardware counters to produce millisecond accuracy for response latency. The timer routines were provided by James Savusch at the State University of New York at Buffalo. The first 12 of the 40 digital I/O lines are currently dedicated to button response and feedback lamps. The button-response boxes were developed by the Indiana University Psychology Department Electronic Shop. Detailed description of what is needed to connect from the boxes to the I/O card can be found in the Strawberry Tree manuals. For most experiments, we will use two- or seven-button response boxes with three or eight feedback lamps.

However, the button-box routines are capable of employing response boxes with anywhere from one to eight buttons and feedback lamps.

We picked Metrowerks CodeWarrior Gold version 5.0 for its capability to compile and generate executables for different platforms and for its appealing debugger. We complemented the C compiler with Apple's QuickTime2.0 developers kit.

Implementation Phases

Phase 1: Handwritten Response

We usually set out to develop routines and put together systems with a particular set of experimental paradigms in mind. In this case, however, we started by implementing a basic identification task (MovieID). For this first phase, the application would merely play movies with a fixed ISI. A skeleton of code from a movie controller provided by QuickTime SDK was taken as a model and standard C timing routines were used. This application would operate in the following steps: First, a stimulus-set-file (SSF) is created. Next, all experimental settings are typed or modified from the default, and finally the experiment runs, and timestamps are logged to a file.

The SSF is a text file containing the list of filenames of the movies to be played. The user creates this file before hand using a general text editor (entering the name of each movie per line). MovieID uses this file to read in the names of the movies to be played. The Mac simplifies the process of typing in a long list of names by its copy-and-paste (i.e., copying from the desktop and pasting to the SSF text file) desktop editing feature. This file is saved in the same subdirectory where the movies reside.

We developed a user-friendly GUI dialog box that allows for easily modifiable variables which the experimenter can modify to meet the requirements of a particular experiment. These settings are recorded to a log file to maintain an accurate report of experimental setup. The setup dialog box allows the experimenter to modify the following settings:

- General comment
- Log file name
- Cue file name (optional)
- SSF name
- Number of blocks
- Number of repetitions
- Randomization (on/off)
- Inter-stimulus interval time (ISI)
- Inter-trial interval time (ITI)

A set of default values is also available with simple click of the mouse. We realize that for this type of experiment, the accuracy for ISI and ITI is not very important given that 20 seconds are usually allowed for subjects to make their responses (ISI) using ITI times of 5 seconds or more.

To maintain an accurate record of the experiment, the data are logged (recorded) to a file for post-experimental analysis. This file contains a header with all of the modifiable settings. Following the header, the actual experimental data are provided including the trial number, repetition number, token number, and movie filename. The footer of this file contains a timestamp showing the date and time of completion of the

experiment. This file is saved in ASCII text format which can be easily loaded into a spreadsheet (e.g., Excel) or statistical software (e.g., Statview) for analysis.

MovieID presents an option for the presentation of a cue-movie between the presentation of stimulus-movies. The purpose of this cue is to prepare the subjects for the next stimulus. We decided to keep the cue in the format of a movie since this would allow not only a dynamic movie as a cue but also a static-movie (an image), a sound-only movie (only sound output with black screen), or a combination of the two, as most movie-making applications would allow the experimenter to create.

The timeline of the MovieID experiment consists of the sequence; stimulus, ISI, cue, and ITI. If a cue is not present, the ITI time is ignored. During any of the delay-intervals (e.g., ISI and ITI), a blank screen is presented. This run-time sequence is very general and is used in the next two applications with minor modifications. This experiment also has the capability of being prematurely terminated during the presentation of any token or cue.

During the development of this first phase, we realized that we did not have much control over exactly when a movie started playing. Another observation was that the application required at least 30MB of RAM before it could display the movie smoothly and without interruptions. To help the processor maintain its throughput, we turned off all unnecessary extensions. Finally, it will also be helpful to defragment the disk when large movies are played.

Phase 2: Keyboard Response (MovieIDkb)

The second step was to modify the first version of the movie-player to accommodate keyboard responses from the subject after the presentation of each movie. All subject responses, consist of a series of keystrokes terminated by a "return." These are logged to a file as they are input into the computer.

The MovieIDkb has the same functionality as the original MovieID identification experiment but has a few minor changes. The setup dialog box now allows the experimenter to modify the instruction for the dialog box in which the subject enters his/her response. Among the body of experimental data being logged to file, the subject's responses are also logged to file as their input. The timeline of the MovieIDkb is exactly the same as MovieID except that the subject is prompted to type their response. The subject's keyboard response is not timed, and therefore has an indefinite period of time to respond.

This particular phase seemed easier to implement that the first, with the exception that we had to worry about insuring that no keyboard combination would interrupt the program. Also, we came across our first non-standard dialog box (one not from the developer's library) and had to deal more closely with the Macintosh event handler. If this software is to be transportable, it needs to be restricted to the Macintosh family of computers.

Phase 3: Button Response (Forced Choice Experiment, MovieID-RT)

For the third phase, we modified the original version of MovieID to accommodate real-time button-box response with feedback. The subject is now forced to make a decision about each stimulus presentation, at which point the button number, latency and current frame would be immediately recorded.

Feedback is also available through the LED sitting on top of each button on the button box. An additional file paired with the SSF is specified in the set-up dialog called stimulus name file (SNF), and is

used to keep track of correct responses. This file contains a list of numbers ranging from 1 through 7, corresponding to response buttons which are labeled accordingly from left to right.

The timeline mimics that of the original version but contains a few modifications to accommodate the paradigm. The response box is deactivated prior to the playback of the token as well as, at any time after a response has been made to avoid recording multiple button responses. If feedback is specified, the correct answer will be illuminated on the button box for a specified time.

In addition to the experimental data being logged to file, the subject's button response number, response latency, and current movie frame are also logged to file. The response latency and movie frame are both included to provide the experimenter with a precise means of detecting any deviation from actual response time and frame number.

To measure precisely the time from the onset of the stimulus to the button press, we had to load the movie into memory and have the first frame waiting in the hardware's buffer. With this strategy, the movie would start playing immediately after the "StartPlaying" routine call. These routines are available in the QuickTime libraries, making it possible for us to develop this strategy without having to program at a lower level. The other component needed for accurate measurement is the digital I/O board. The programming of this board required us to work at a register hardware level. Although we had to study and modify some routines to make this work with our system, we are very grateful to Dr. James Sawusch for providing us with timer and response-box routines he previously developed for his own system using the same I/O board.

Development of Resource Tools

The final goal of this project was to assemble a group of movie-player and Mac interface routines and provide a general code skeleton for future A+V development on the Mac. Two sets of libraries were developed. One library was created and it contains general routines such as displaying a dialog box, creating a randomized list of numbers, delaying a certain length of time, and converting Mac strings (Pascal-style) to "C" strings. In the past, we have noticed that most of the development time is spent fine-tuning the GUI. We hope that this group of routines will allow more time to be dedicated to the functionality component of development. The second library contains full-screen movie-player routines such as initializing the movie environment, starting and stopping a movie, capturing a frame number, and loading the movie into memory. These routines are the core of functions that allow us to perform the tasks with the necessary flexibility and precision.

As with previous libraries developed in our laboratory over the years (Forshee 1975, 1979; Forshee & Nusbaum, 1984; Hernández, Carrell, Reutter, & Bernacki, 1992), we hope to simplify software development for non-programmers. Giving the users the ability to plan and experiment with their own programs, has yielded a better environment for researchers using computerized experimental paradigms.

In summary, The Macintosh system that we have constructed has proven to work for our initial purposes. We feel that we understand the hardware and software components of the system well enough to have control over the stimulus presentation and latency measures to develop these and other perceptual experiments. We feel that we have met the functional and hardware requirements as described and will continue to test and develop the system as newer paradigms arise.

Discussion

There are some design issues that still need to be discussed and are worth mentioning at this time in development. We have purposely ignored accuracy between events that are not measured with the digital I/O card (i.e., ISI and ITI). This assumption is fine for the paradigms described in this paper, but the addition of precise countdown timers and wait routines are unavoidable. In addition, more external testing should be done to insure that internal and external measures are the same.

Another issue that has not been investigated is the position of the raster when the movie is loaded to the screen. Not only do we need to insure that the raster is positioned at the top left corner of the screen before display, but we also need to measure the time it requires to traverse the entire screen as was done previously for our PTS system (Hernández, 1995). It would be convenient at this stage to calibrate all monitors used in a single experiment with each other to insure uniformity of color parameters.

More testing needs to be done on the audio output to establish a base line signal-to-noise ratio, distortion and other parameters that will inevitably be a factor in many of the experiments conducted here.

Finally, the market changed the Macintosh bus standard during the development of the system described above from NuBus to PCI. Although a Radius VideoVision PCI card is available with the same features as the NuBus card, Strawberry Tree does not manufacture an I/O PCI card equivalent to the ACM2. In order to avoid rewriting the I/O component from another company that offers a PCI I/O card, we will use either older NuBus Macintoshes or newer Macintosh-clones which offer both the PCI and NuBus architectures. Again, further external tests will have to be conducted to determine if the different computer models behave the same way. Also during this same period of development, other companies have developed video cards that can compete with VideoVision, both in terms of performance and, importantly, price. The primary product among these is the Targa board from Truevision. We are seriously considering replacing our VideoVision boards with Targa boards. An extremely important consideration, however, is that movie formats compiled with VideoVision are not compatible with Targa, and vice versa. On the other hand, the only changes to the software would be recompiling with the Targa codec. Currently, we have tested our system only on 950AV Macintosh and PowerPC Macintosh 8100 models.

References

Forshee, J.C. (1975). Speech perception research laboratory: The state of the computer system. *Research on Speech Perception Progress Report No. 2*, Indiana University, Bloomington, IN, pp. 202-220.

Forshee, J.C. (1979). Speech perception research laboratory: Current computer resources. *Research on Speech Perception Progress Report No. 5*, Indiana University, Bloomington, IN, pp. 449-473.

Forshee, J.C., & Nusbaum, H.C. (1984). An update on computer facilities in the Speech Research Laboratory. *Research on Speech Perception Progress Report No. 10*, Indiana University, Bloomington, IN, pp. 453-462.

Hernández, L.R., Carrell, T.D., Reutter, J.G. & Bernacki, R.H. (1992). A new PC-based real-time experiment control system. *Research on Spoken Language Processing Progress Report No. 18,* Indiana University, Bloomington, IN, pp. 243-253.

Hernández, L.R. (1995). Implementation of a PC-based perceptual testing system (PTS): A first milestone. *Research on Spoken Language Processing Progress Report No. 19,* Indiana University, Bloomington, IN, pp. 321-328.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

Using CD-ROM as a Storage Medium for Digitized Speech Materials[1]

Jon M. D'Haenens and Luis R. Hernández S.

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

[1]This work supported, in part, by grants to Indiana University–Bloomington: NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012.

Using CD-ROM as a Storage Medium for Digitized Speech Materials

Abstract. This report describes the use of CD-ROM technology for the storage of digitized speech material. Advantages of this technology are discussed, as well as general procedures and particular hardware and software specifications. Finally, examples of specific databases from the Speech Research Laboratory that have been stored to CD-ROM are discussed.

Overview

The options available for the storage of digitized audio files used in the acoustic analysis of speech and as stimuli in perception experiments in the Speech Research Laboratory have been limited to magnetic tape storage. Due to the relatively large size of digitized audio files and the increasing number of files that must be stored, it has become impractical to permanently store speech audio materials on the hard drives of computers in our laboratory. As a result of these space and operational limitations, it is necessary to store the materials on high-capacity magnetic media such as TK-70 tape or DAT (Digital Audio Tape). Although these materials have a high capacity and are efficient for off-line storage, they exist mainly for the purpose of archiving data. Sound files must be retrieved from the tape using a utility program and stored on a hard drive before they can be analyzed or used in perception experiments. The result is that researchers using a large volume of digitized speech materials must deal with space restrictions and long waits for the retrieval of sound files from archive tapes.

With CD-ROM technology (re: ANSI, 1990), it is now possible to store a large amount of data on a small disk that provides the advantages of both the tape archival systems and hard disk storage. Some of these advantages include high capacity, speed, permanence, and the ability to work with files directly from the disk. Because CD-ROM is now the standard for many storage applications, the format is supported by nearly every platform, and CD-ROM readers are supplied with almost every new computer. This simplifies the process of sharing data; digitized sound files stored on a disk can be read by any computer with a CD-ROM reader. The data stored on a CD-ROM cannot be erased, written over, or modified, which may be considered an advantage or disadvantage, depending on the application.

CD-ROM is the ideal medium to solve many of the Speech Research Laboratory's digitized speech storage problems. Each disk allows for the storage of a significant amount of digitized audio: approximately 300 minutes of one channel of raw sound data at a 20 KHz sampling rate and with 16-bit resolution. Access to the data is considerably faster than with tape (the access speed is dependent on the CD-ROM reader). Additionally, it is not necessary to copy the sound files to a hard drive to listen to them or access them from within signal analysis software, eliminating the necessity of shuffling files from one storage medium to another. Other advantages include permanence and safety of data storage (CD-ROMs cannot be erased or written over), and the ability to store other relevant textual data along with the audio signals, eliminating most paper documentation and references to the material.

Process

The hardware used for CD-R (CD Recordable) writing in our lab consists of a Macintosh Quadra 840AV with a Yamaha CDE100 CD-R mastering drive connected through the SCSI bus. The Quadra is

connected to the local area network, allowing files to be transferred from almost every computer within the lab. The software used is GEAR for Macintosh®, Version 2.5 (Elektroson Software).

The first step in the creation of a CD-ROM is the restoration of the sound files to the local hard drive of the native platform, for example, from the TK-70 tape to the VAX or from DAT to a local PC. After the files have been restored, they can be transferred directly over the network by copying them to the Quadra equipped with the CD-R mastering drive. An alternative method used for computers with no direct network drive access is transferring the files via FTP (Internet File Transfer Protocol). Usually, the restoration and transfer must be done in batches, as there is typically not enough space to store all files from an archive on a local hard drive. It is important that all files are transferred to the hard drive of the Quadra used to create the CD-ROM, as network access is too slow for the data transfer to the CD.

Once the files have been restored and transferred to the Quadra used for CD-R writing, the GEAR CD-R mastering software is used to create a physical image of the CD. Before creating the image, it is important to choose the ISO-9660 format for the target CD-ROM, which is currently the standard format and is readable by almost every computing platform.

A physical image is simply a bit-for-bit representation of what the CD will contain. Creating a physical image is a necessary step because it saves critical processing and disk access time while the CD is being written. If there is any delay in the data stream during the CD-R writing process, the resulting CD-ROM will be unreadable. After the image has been created, the final step is to place a CD-R in the CD-R mastering drive and select 'Write' from the GEAR menu. At this point, the physical image is written to the disk . After the process is completed, the CD is automatically ejected and is tested to ensure that it was written correctly. For detailed instructions on the use of the GEAR software, refer to the GEAR User's Manual.

Example Applications

In the Speech Research Laboratory, we have already created two CD-ROMs, each with a slightly different purpose. The PB/MRT database (re: ANSI, 1971; House et al., 1965) of monosyllabic words is used very often by many people in the laboratory, and is often requested by people outside the laboratory. In the past, these data were either stored on a hard drive so other labs could access them over the Internet, or the desired parts would be retrieved from tape and stored on a hard drive for individual use. Putting these data on a CD-ROM makes them much easier to share. The CD-ROM can be mounted on a PC and accessed via FTP from other laboratories, saving retrieval time and disk space. Additionally, whenever anyone needs to access the data within the laboratory, they can simply insert the CD-ROM into any computer with a CD-ROM reader and have immediate access to the data on the disk .

Another CD-ROM that we created was an archive of materials obtained for the Indiana University/General Motors Research Laboratories studies of the effects of alcohol on speech. This archive contains digitized sound files of words, sentences, and paragraphs. The primary motivation for putting these data on CD-ROM was the anticipation of doing a large amount of computer aided acoustic analysis of the files. By storing these data on CD-ROM, not only do we save time and disk space, but our researchers have immediate and random access to all of the materials, making the analysis much faster and easier.

For the GM studies CD-ROM, the data was initially written to the CD-ROM in the ".ils" format (Interactive Laboratory Systems: Signal Technology, Inc.) in which that data were digitized. Because the

analysis was to be performed with software without native support for the ".ils" format, the files were transformed to the Entropic Waves ".sd" format (Entropic Research Laboratory, Inc.) and written to another CD-ROM. This was necessary to eliminate the need to convert the files from one format to another and the need to store them on a hard drive.

Summary

In the process of creating the two CD-ROMs described above, we have arrived at some important points to consider before beginning the process.

- Because a CD-ROM is a write-once medium, it is important to make sure everything to be placed on the CD is in its final form and has been verified. If anything needs to be changed after the CD-ROM has been written, a new CD must be burned in (created).

- Although it is possible to write the data in multiple sessions, some CD-ROM readers will not recognize any session after the first. If planning on writing multi-session CD-ROMs, verify that the CD-ROMs drives to be used are multi-session.

- Before beginning the process, decide on a final directory structure for the organization of the CD-ROM. This makes the process of transfering the files easier, and a good plan makes working with the CD-ROM much easier.

- Use the ISO-9660 format. It is the most universally recognized format.

Conclusions

Although we are in the early stages of utilizing CD-ROM technology, it has already proven to be reliable and time-saving. CD-ROM is not perfect for every storage application because of the overhead in creating the disk , but there are many ways that the Speech Research Laboratory has found it to be useful for archiving large sets of sound files. The next steps are to create CD-ROMs for the remainder of the databases that are often used, and to educate the users of files in our lab on ways it can be utilized.

References

American National Standards Institute. (1971). *Method for measurement of monosyllabic word intelligibility* (American National Standard S3.2-1960 [R1971]). New York: Author.

American National Standards Institute. (1990). *Volume and file structure of CD-ROM for information interchange* (ANSI/NISO/ISO 9660-1990). New York: Author.

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

Pisoni, D.B., & Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analysis. *Alcoholism: Clinical and Experimental Research*, **13**, 577-587.

Pisoni, D.B., Yuchtman, M., & Hathaway, S.N. (1986). Effects of alcohol on the acoustic-phonetic properties of speech. In *Alcohol, Accidents, and Injuries* (pp. 131-150). Warrendale, PA: Society of Automotive Engineers.

IV. Publications

Papers Published:

Bradlow, A.R. (1995). A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical Society of America*, **97**, 1916-1924.

Bradlow, A.R., Pisoni, D.B., Yamada, R.A. & Tohkura, Y. (1995). Acquisition of the English /r/ - /l/ contrast by Japanese speakers: Effects of training in perception on production. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, **4**, 562-565.

Bradlow, A.R., Torretta, G.M. & Pisoni, D.B. (1995). Some sources of variability in speech intelligibility. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, **1**,198-201.

Bradlow, A.R., Port, R.F. & Tajima, K. (1995). The combined effects of prosodic variation on Japanese mora timing. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, **4**, 344-347.

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.

Kirk, K.I., Pisoni, D.B., Sommers, M.S., Young, M., & Evanson, C. (1995). New directions for assessing speech perception in persons with sensory aids. *Annals of Otology, Rhinology and Laryngology*, **104**, 300-303.

Magnuson, J.S., Yamada, R.A., Tohkura, Y., Pisoni, D.B., Lively, S.E., & Bradlow, A.R. (1995). The role of talker variability in non-native phoneme training. *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan*, Pp. 393-394.

Mandel, D.R., Jusczyk, P.W. & Pisoni, D.B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, **6**, 314-317.

Nygaard, L.C. & Pisoni, D.B. (1995). Talker- and task-specific perceptual learning in speech perception. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, Pp. 194-197.

Nygaard, L.C., Sommers, M.S. & Pisoni, D.B. (1995). Effects of stimulus variability on the representation of spoken words in memory. *Perception & Psychophysics*, **57**, 989-1001.

Yamada, R.A., Tohkura, Y., Bradlow, A.R., & Pisoni, D.B. (1995). The effect of perception training on Japanese speakers' production of the American English /r/-/l/ contrast. *Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan*, Pp. 385-386.

Book Chapters:

Bradlow, A.R., Nygaard, L.C. & Pisoni, D.B. (1995). On the contribution of instance-specific characteristics to speech perception. In C. Sorin, J. Marianai, H. Meloni and J. Schoentgen (Eds.), *Levels in Speech Communication: Relations and Interactions*. New York: Elsevier, Pp. 13-24.

Bradlow, A.R., Nygaard, L.C., & Pisoni, D.B. (1995). Indexical and linguistic attributes in speech perception: A review of some recent findings. In B. Kanki and R. Prinzo (Eds.), *Metrics and Methods in Speech Communication.* Office of Federal Aviation Administration, National Aeronautics and Space Administration/Department of Defense. Pp. 67-78.

Dinnsen, D.A. & Chin, S.B. (1995). On the natural domain of phonological disorders. In J.Archibald (Ed.), *Phonological Acquisition and Phonological Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates. Pp. 135-150.

Goldinger, S.D., Pisoni, D.B., & Luce, P.A. (1995). Speech perception: Research and theory. In N.J. Lass (Ed.) *Principles of Experimental Phonetics*. Toronto, Canada: B.C. Decker, Pp. 277-327.

Nygaard, L.C. & Pisoni, D.B. (1995). Speech perception: New directions in research and theory. In J.L. Miller and P.D. Eimas (Eds.), *Handbook of Perception and Cognition, Volume 11, Speech, Language and Communication*. New York: Academic Press, Pp. 63-96.

Pisoni, D.B. (1995). Sources of variability affecting speech perception and spoken word recognition. In C.W. Nixon (Ed.), *Symposium on Speech Communication Metrics and Human Performance (AL/CF-SR-1995-0023).* Wright Patterson AFB, OH: Armstrong Laboratory, Crew Systems Directorate, Pp. 19-40.

Pisoni, D.B. & Lively, S.E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-language Speech Research.* Timonium, MD: York Press, Pp. 429-455.

Ralston, J.V., Pisoni, D.B., & Mullennix, J.W. (1995). Comprehension of synthetic speech produced by rule. In R. Bennett, A. Syrdal, and S. Greenspan (Eds.) *Behavioral Aspects of Speech Technology: Theory and Applications*. New York: Elsevier, Pp. 233-287.

Manuscripts Accepted for Publication (In Press):

Aslin, R.N., Jusczyk, P.W. & Pisoni, D.B. (In Press). Speech and auditory processing during infancy: Constraints on and precursors to language. In W. Damon (series ed.), *Handbook of Child Psychology. Fifth Edition, Volume 2: Cognition, perception & language* (D. Kuhn & R. Siegler, Eds.). New York: Wiley.

Bradlow, A.R. (In Press). A perceptual comparison of the /i/–/e/ and /u/–/o/ contrasts in English and in Spanish: Universal and language-specific aspects. *Phonetica*.

Chin, S.B. (In Press). The role of the sonority hierarchy in delayed phonological systems. In T.W. Powell (Ed.), *Pathologies of Speech and Language: Contributions of Clinical Phonetics and Linguistics*. New Orleans, LA: International Clinical Phonetics and Linguistics Association. Pp. 109–117.

Chin, S.B. & Pisoni, D.B. (In Press). *Alcohol and Speech.* San Diego, CA: Academic Press.

Kirk, K.I., Pisoni, D.B. and Miyamoto, R.C. (In Press). Effects of stimulus variability on speech perception in hearing impaired listeners. *Journal of Speech and Hearing Research.*

Kirk, K.I., Diefendorf, A.O., Pisoni, D.B. and Robins, A.M. (In Press). Assessing speech perception in children. Chapter to appear in L.L. Mendel and J.L. Danhauer (Eds.), *Assessing Speech Perception*. San Diego: Singular Press.

Miyamoto, R.T., Kirk, K.I., Robbins, A.M., Todd, S.L., Riley, A.I., & Pisoni, D.B. (In Press). Speech perception and speech intelligibility of children with multichannel cochlear implants. *Advances in Oto-Rhinolaryngology*, Basel, Switzerland.

Mullennix, J.W., Goldinger, S.D., & Pisoni, D.B. (In Press). Some characteristics of talker normalization. In J. Charles-Luce, P.A. Luce, and J.R. Sawusch (Eds.) *Theories in Spoken Language: Perception, Production and Development*. Norwood, NJ: Ablex.

Pisoni, D.B. (In Press). Perception of synthetic speech produced by rule: A selective review and interpretation of research over the last 15 years. In J.P.H. van Santen, R.W. Sproat, J.O. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis*. New York: Springer-Verlag.

Pisoni, D.B. (In Press). Some Thoughts on "Normalization" in Speech Perception. Chapter to appear in K. Johnson and J.W. Mullennix (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press.

Rosenblum, L.D., & Saldaña, H.M. (In Press). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance.*

Rosenblum, L.D., Johnson, J.A., & Saldaña, H.M. (In Press). Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research.*

Ryalls, B.O. & Pisoni, D.B. (In Press). The effect of talker variability on word recognition in preschool children. *Developmental Psychology.*

Saldaña, H.M., Nygaard, L.C. & Pisoni, D.B. (In Press). Episodic encoding of visual speaker attributes and recognition memory for spoken words. In D. Stork (Ed.), *Speech Reading by Man and Machine: Models, Systems, and Applications*. Berlin: Springer-Verlag.

Saldaña, H.M., Pisoni, D.B., Fellowes, J.M., & Remez, R.E. (In Press). Audio-visual speech perception without speech cues: A first report. Chapter to appear in D. Stork (Ed.), *Speechreading by Man and Machine: Models, Systems, and Applications*. Berlin: Springer-Verlag.

Sommers, M.S. & Kewley-Port, D. (In Press). Modeling formant frequency discrimination of female vowels. *Journal of the Acoustical Society of America.*

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