RESEARCH ON SPEECH PERCEPTION

Progress Report No. 16
(1990)

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-00111-14

National Institutes of Health
Training Grant No. DC-00012-12

National Science Foundation
Research Grant No. IRI-86-17847

and

U. S. Air Force
Armstrong Aerospace Medical Research Laboratory
Contract No. AF-F-33615-86-C-0549

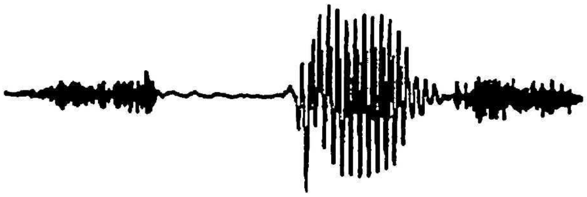
This is the sixteenth annual report summarizing the research activities on speech perception, analysis, synthesis, and recognition carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize various research activities 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 support 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 our own research.

We are distributing 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 speech processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis, synthesis, and recognition and, therefore, we would be grateful if you would send us copies of your own 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
USA
(812) 855-1155, 855-1768
E-mail (BITNET) "PISONI@IUBACS"
E-mail (INTERNET) "PISONI@UCS.INDIANA.EDU"

Copies of this 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.

The information contained in the 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.



RESEARCH ON SPEECH PERCEPTION Progress Report No.16 (1990)

Table of Contents

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Speech Perception and Spoken Word Recognition: Research and Theory[1]

Stephen D. Goldinger, David B. Pisoni and Paul A. Luce[2]

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

[1] To appear in N.J. Lass (Ed), *Principles of Experimental Phonetics*. Toronto: B.C. Decker, Inc. This research has been supported by NIH Research Grant #DC-00111-14 to Indiana University.

[2] Now at Department of Psychology, State University of New York at Buffalo, Buffalo, NY 14260.

Speech Perception and Spoken Word Recognition: Research and Theory

Introduction

Speech Perception and Spoken Word Recognition

The study of speech perception is concerned with the listener's ability to perceive the acoustic waveform produced by a speaker as a string of meaningful words and ideas. By this definition, speech perception has been researched since at least the turn of the century, when one of the earliest empirical studies of language perception was published by William Chandler Bagley (1900-1901; see Cole & Rudnicky, 1983). Bagley's experiments addressed a surprisingly wide variety of topics that have since been re-discovered, including phonemic restoration, semantic priming, primacy of word-initial information, and sentence context effects on word recognition. A common theme of all Bagley's experiments, however, was their focus on the influence of semantic and lexical knowledge on the perception of distorted words. As Cole and Rudnicky (1983) observed, Bagley anticipated many of the empirical phenomena and theoretical accounts of speech perception and spoken word recognition that remain central to theoretical discussions today.

If Bagley's results and arguments were presented today, they would most likely be considered relevant to *language perception*, rather than to *speech perception*, per se. The term "speech perception" has, for a variety of reasons, come to refer more specifically to *phoneme* perception than to the perception of words or phrases. Unlike a process such as person perception in which the recognition of objects or motion is available to the observer, "speech perception" , as the researcher defines it, is a process of which we are generally unaware. As Darwin (1976) comments, "Our conscious perceptual world is composed of greetings, warnings, questions, and statements; while their vehicle, the segments of speech, goes largely unnoticed and words are subordinated to the framework of the phrase or sentence" (page 175). Despite the truth of Darwin's observation, the bulk of the research conducted on speech perception in the last three decades has applied only to the unnoticed vehicle-- to speech in the narrow sense of phonetic perception. This rather myopic approach to speech perception has resulted in a large theoretical body of literature in speech perception that has been somewhat divorced from more general theories of perception as well as the mainstream of cognitive psychology. For example, only recently have serious efforts been applied to model the process of speech perception not as an end in itself, but as the process or processes subserving word recognition (see Pisoni & Luce, 1987). In this chapter, we will consider the effects of this segregated research and theorizing on our understanding of speech as the "front end" of language.

Overview and Orientation of the Chapter

Numerous papers and chapters have been written reviewing the theories and data in speech perception (e.g., Cutting & Pisoni, 1978; Darwin, 1976; Luce & Pisoni, 1987; Miller, 1990; Pisoni, 1978; Pisoni & Luce, 1986; Studdert-Kennedy, 1974, 1976). In large part, the fundamental issues in speech perception and the data relevant to those issues have remained unchanged over the past several years. Accordingly, although the present chapter will address and elaborate on several of the fundamental issues in speech perception, space limitations preclude a comprehensive review of the empirical literature in the field. Nevertheless, our approach to speech perception in this chapter is fairly eclectic, and we hope to address a sufficiently wide range of topics. We do not attempt to marshall evidence for one particular theory or class of theories at the expense of all others; instead, we attempt to examine and evaluate a wide range of theories. Finally, we attempt throughout the chapter to examine how research and theory in the field of speech perception has or has not developed over the years with respect to the fundamental "problems" of speech. In the next section, we begin with a review of several of the long-standing, basic issues in speech perception.

Basic Issues in Speech Perception

Linearity, Lack of Acoustic-Phonetic Invariance, and the Segmentation Problem

In the years since the mid-1950s, no finding has influenced speech research and theory more profoundly than the failures of the speech signal to satisfy the linearity and invariance conditions. The *linearity* condition assumes that for each perceived phoneme, there must be a particular, corresponding stretch of sound in the utterance (Chomsky & Miller, 1963). For example, if the listener perceives that phoneme X occurs before phoneme Y, the stretch of sound associated with phoneme X must precede the stretch of sound associated with phoneme Y in the physical signal. The *invariance* condition assumes that for each phoneme X, a specific set of criterial acoustic correlates associated with the phoneme must occur in all phonetic contexts. Under these conditions, recognition of phoneme X implies that the features for X occurred in the speech signal in a discrete time window, and that no other features or temporal distributions of features could have occurred.

Neither the linearity nor the invariance conditions are met in natural speech primarily because of the manner in which speech is produced; the speech articulators move continuously in production in such a way that the shape of the vocal tract for each intended phoneme is influenced by the shapes for both the preceding and following phonemes. The coarticulation of speech results in featural overlap or "smearing" among neighboring phonemes. Hockett (1955) likens the relation between intended phonemes and the physical speech signal to a series of individual Easter eggs that are pushed through a wringer. The effect of the speaker's coarticulatory wringer is to create a speech signal in which there is rarely a stretch of sound that corresponds uniquely to a given phoneme. Instead, the cues overlap in time, resulting in what Liberman, Cooper, and Shankweiler (1967) have termed the "encoded nature of speech." Coarticulation of speech sounds results in complex mappings between acoustic cues and perceived phonemes. Acoustic features for phonemes vary widely as a function of varying phonetic contexts, speaking rates, etc. As an example, Figure 1 shows the variations in second-formant transitions (schematized, as they would be for synthesis on the pattern playback device (Cooper, Liberman, & Borst, 1951)) for the phonemes /b/, /d/, and /g/ as a function of varying vowel contexts (Delattre, Liberman, & Cooper, 1955; Liberman, Delattre, Cooper, & Gerstman, 1954). Although the second-formant transition provides the cues for place of articulation that distinguish these phonemes, the acoustic realizations of the transitions are clearly not invariant.

Insert Figure 1 about here

The failures of the speech signal to satisfy the linearity and invariance conditions is perhaps the most important puzzle in speech perception, and constitutes what Studdert-Kennedy (1983) refers to as the "animorphism paradox"-- the invariant units of perception do not correspond to invariant acoustic segments in the signal. Indeed, although the problems related to acoustic-phonetic invariance have characterized the field of speech research since its beginning, many researchers are still working on resolution of these problems today. And while some researchers have attempted to continue the quest for invariant aspects of the acoustic signal (e.g., Kewley-Port, 1982, 1983; Kewley-Port & Luce, 1984; Stevens & Blumstein, 1978, 1981), and others have attempted to resolve the problems of invariance via theoretical innovations (e.g., Liberman & Mattingly, 1985; McClelland & Elman, 1986), the problem of contextual variability in speech remains central in current research.

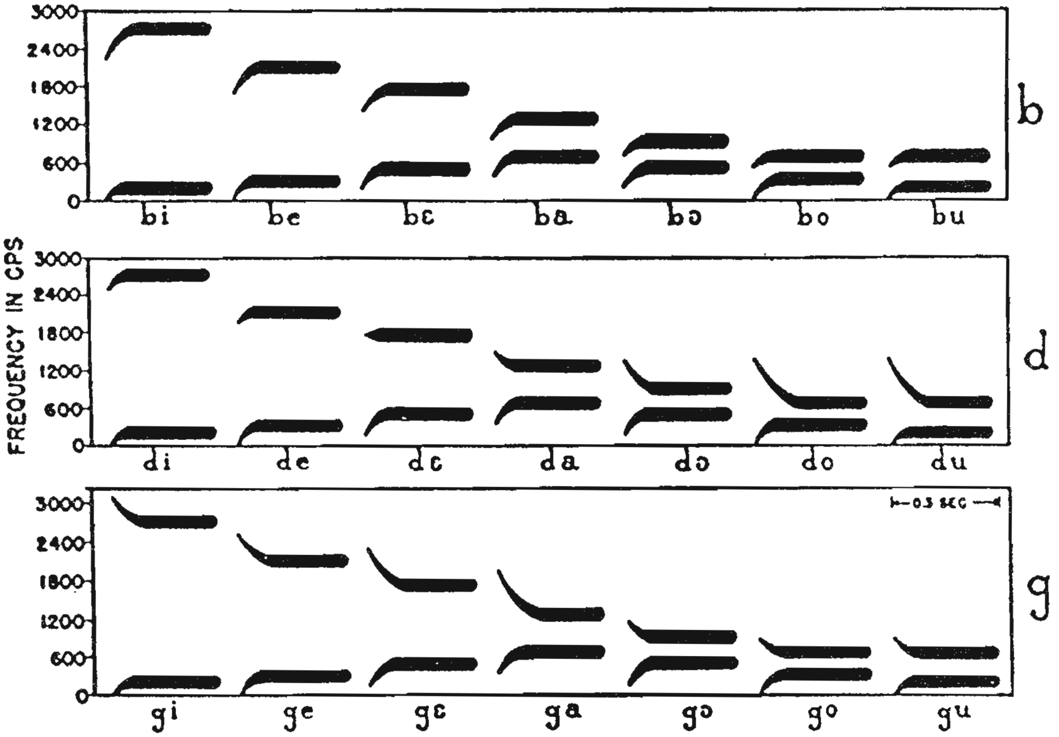


Figure 1. Schematized sound spectrograms showing the formant transitions that are appropriate for the voiced stop consonants /b/, /d/, and /g/ before various vowels. (From Delattre, Liberman, & Cooper, 1955, with permission of the authors and publisher).

The coarticulation of speech also results in another problem for research, namely, the lack of clear segmentation in the speech signal. Although listeners perceive speech as a series of discrete phonemes and words, physical temporal boundaries between phonemes are not reliably found in the spoken utterance. Figure 2 shows a spectrogram of the utterance, "I owe you a yo-yo," which displays almost no physical landmarks on which to base segmentation but does not pose any special difficulties for listeners to parse efficiently. The rate of information transmission in speech is enormous and clearly requires that listeners somehow convert the continuous waveform into discrete, abstract units in order to reduce information for subsequent cognitive processing (Neisser, 1967; Liberman et al., 1967). The speech signal, however, does not lend itself to simple segmental analysis. Although it has been possible to segment speech according to purely acoustic criteria (Fant, 1962), the segmentation provided by such algorithms typically does not correspond to the segmented representation of speech a listener would perceive. The importance and the difficulty of the segmentation problem becomes immediately apparent when considering recent attempts to develop speech recognition devices, in which segmentation, along with lack of linearity and invariance, have proven to be almost intractable problems.

Insert Figure 2 about here

Units of Analysis in Speech Perception

The problems of non-linearity, the lack of acoustic-phonetic invariance, and the non-segmental nature of speech create another difficult problem for theories of speech perception-- the selection of a minimal unit of perceptual analysis. Because of the nature of the information-rich speech waveform and limitations of channel capacities of the auditory system and auditory memory, it is clear that raw sensory information must be encoded into some representation scheme that can be processed efficiently (Broadbent, 1965; Liberman et al., 1967). To appreciate the importance of this initial encoding, consider the estimate given by Liberman, Mattingly, and Turvey (1972) that the conversion of speech sounds into phonemes reduces the information transfer rate of speech from approximately 40,000 bits per second to 40 bits per second. The conversion of phonemes into higher linguistic units of analysis reduces the bit rate further still.

The question for theories of speech perception has typically concerned the selection of the "best" or most natural coding unit; claims of primacy have been made for phonetic features, phonemes, syllables, and words. Researchers from the tradition of generative linguistic theory have even proposed units as large as the clause or sentence (Bever, Lackner, & Kirk, 1969; Miller, 1962). Debates concerning the primacy of various units were widely represented in the literature on speech perception for several years. As an example, a long-standing debate in the 1970s centered around claims that the syllable is a more basic perceptual unit than the phoneme (e.g., Massaro, 1972; Savin & Bever, 1970). Massaro (1972) argued, for example, that syllables are more discretely represented in the speech signal than phonemes are, so the assumption that the syllable is the primary unit of perception resolves the problems of segmentation and invariance quite easily. Unfortunately, numerous problems of invariance again arise with syllable-sized units, to such a degree, in fact, that the problems of syllable-sized units are no more tractable than problems of phoneme-sized units. Furthermore, the information conveyed by syllables may be dependent upon retrieval of their segmental constituents, so the issue of primary units is not unambiguously resolved (see Hawles & Jenkins, 1971; Pisoni, 1978).

Recent theories in speech perception, however, imply that questions regarding the primacy of any particular unit of speech over all other units may not be as important to pursue as questions of the *obligatory* units of speech and their interactions during comprehension of fluent speech (e.g., McClelland & Elman, 1986). Although problems of coarticulation have discouraged researchers from positing the

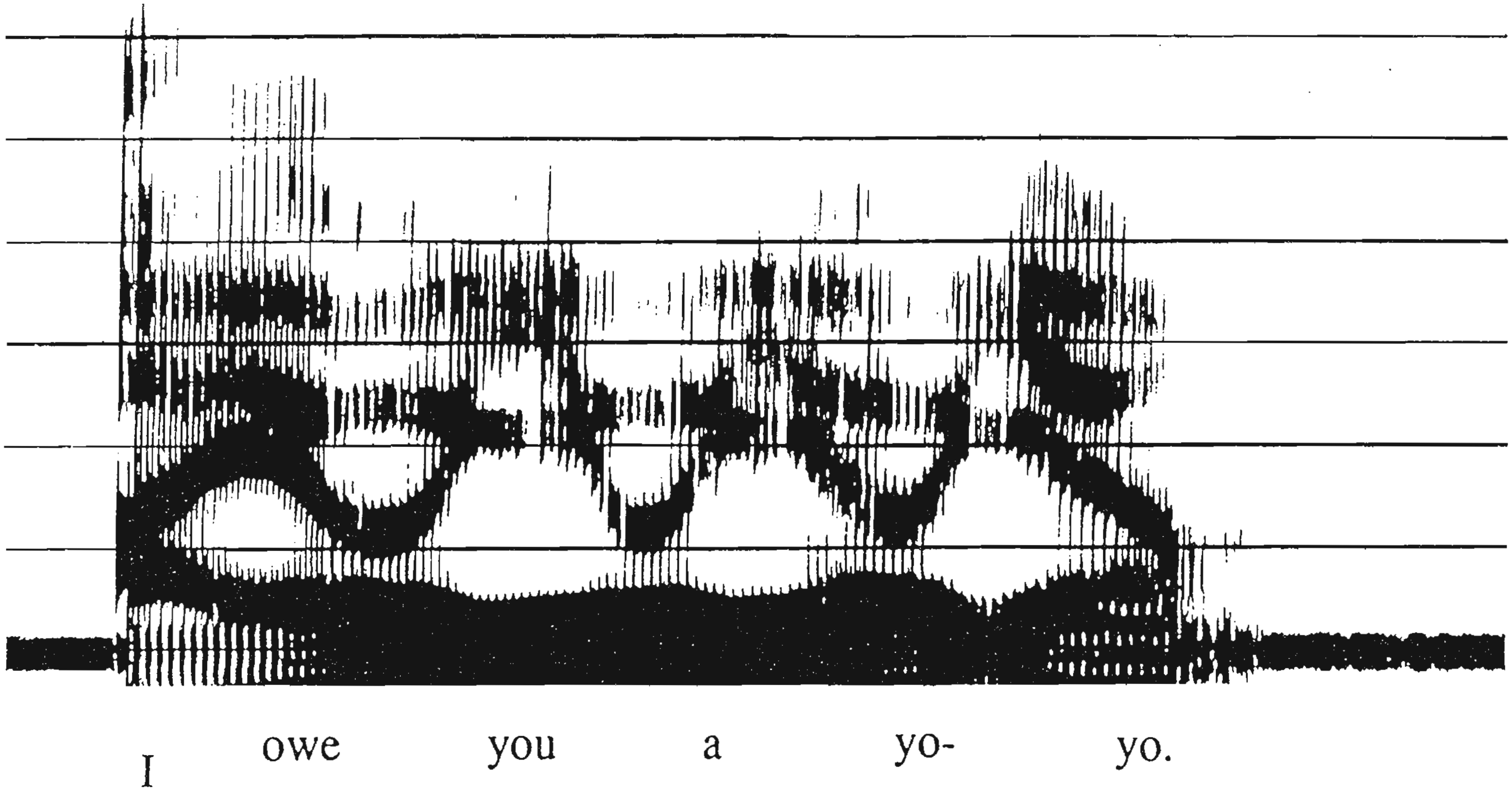


Figure 2. Spectrogram of the utterance, "I owe you a yo-yo," demonstrating that perceptual segmentation is not clearly reflected in acoustic segmentation.

phoneme as an obligatory unit, numerous alternatives have been proposed. Examples include syllables (Cole & Scott, 1974a, 1974b; Massaro & Oden, 1980; Segui, 1984; Studdert-Kennedy, 1974, 1980), context-sensitive allophones (Wickelgren, 1969, 1976), and context-sensitive spectra (Klatt, 1979). All of these approaches have attempted to circumvent the problems of acoustic-phonetic invariance via the proposal of units that are more easily recovered from continuous speech. Although there is still ample reason to consider the importance of segmental representations in speech perception and word recognition (Pisoni & Luce, 1987), these context-sensitive perceptual units incorporate contextual variability directly into their representations, and may therefore prove more robust to the problems of coarticulated speech.

We have discussed four of the most fundamental problems that have shaped speech perception research and theory for nearly four decades and will no doubt figure prominently in future work as well. Further issues aside from the issues we have identified here (linearity, invariance, segmentation, and minimal units of perception), have figured prominently in speech research in recent years. The issues discussed above, however, capture the essence of the "problem" of speech perception: How does the listener convert the continuously varying speech waveform into a series of discrete representations that support further linguistic analysis? This constitutes the fundamental question that any reasonable theory of speech perception must address. In the next section, we turn our attention to somewhat broader issues in speech perception that have been less focal in the research mainstream, but no less interesting than the issues enumerated above. These include the specialization of speech, the problem of perceptual constancy, and the importance of suprasegmental and source information in speech.

Further Issues in Speech Perception

Specialization of Speech Perception

For many years, Liberman and his colleagues at Haskins Laboratories have proposed a view of speech perception as a specialized process, requiring specialized neural mechanisms unique to the human (e.g., Liberman, 1982; Liberman & Mattingly, 1989; Studdert-Kennedy, 1980). The early support for the claim that "speech is special" came from the results of a well-known study conducted by Liberman, Harris, Hoffman, and Griffith (1957). Liberman et al. generated a synthetic continuum of consonant-vowel syllables, ranging from /b/ to /d/ to /g/, by changing the second-formant transitions in graded steps. Although the physical changes between adjacent stimuli were small, when subjects were asked to identify the stimuli, their responses were sharply discontinuous. That is, despite the graded steps in the continuum, subjects' perception of the tokens shifted abruptly, falling into the natural categories for the phonemes /b/, /d/, and /g/. Moreover, when subjects were asked to discriminate tokens drawn from the stimulus continuum, their discrimination of tokens from different phonemic categories was nearly perfect, but their discrimination of tokens from within the same phonemic category was nearly at chance. The phenomenon of discontinuous, categorical perception for speech sounds was markedly different from typical psychophysical experiments employing nonspeech stimuli such as pure tones. Nonspeech continua are perceived continuously, resulting in discrimination functions that are monotonic with respect to the physical scale. These differences between speech and nonspeech perception prompted Liberman et al. to propose that speech perception is subserved by specialized mechanisms that are distinct from mechanisms for general audition (see Repp, 1983a, for a comprehensive review on the categorical perception literature).

Beyond the early findings in categorical perception, a number of empirical phenomena and research paradigms have been purported to demonstrate the specialized nature of speech. These include, among other findings, the findings of phonetic discrimination in infants, the rigidity of adult phonetic categories, cross-modal cue integration, cue trading relations, and the phenomenon of duplex perception. These phenomena are considered in the sections that follow, although we do not discuss the development of speech perception in this chapter (see Aslin & Pisoni, 1980; Eimas, Siqueland, Jusczyk & Vigorito, 1971; or Walley, Pisoni, & Aslin, 1981). We begin our discussion with an examination of categorical perception of speech and nonspeech signals.

Perception of Speech and Nonspeech Signals. As described above, some of the earliest empirical support for the claims of a specialization for speech came from findings of the categorical perception of speech stimuli and comparisons with the continuous perception of nonspeech stimuli. The rationale for these differences offered by Liberman and his colleagues (Liberman, 1970a, 1970b; Liberman et al., 1967; Studdert-Kennedy & Shankweiler, 1970) was based on the framework of the motor theory of speech perception, in which the perception of speech is assumed to be mediated by reference to articulatory knowledge about speech production. So, considering the stimulus continuum examined by Liberman et al. (1957), although the physical scale was composed of many graded steps of second-formant transitions, production of /b/, /d/, and /g/ corresponds to three discrete, discontinuous places of articulation. Listener's perception of these sounds does not follow the continuous physical attributes of the signal, but seems to follow the abstract, discontinuous attributes of place of articulation. In contrast, the fact that nonspeech signals are continuously perceived was taken as further support that, at least for stop consonants, speech perception entails a specialized "speech mode" of perception.

The account of the categorical perception data offered by Liberman and his colleagues, as well as the generality of the data themselves, were challenged by many researchers who believed that the same phenomena could be explained via general principles of auditory perception (e.g., Cutting, 1978; Massaro, 1972, 1987; Pastore, 1981; Schouten, 1980). A problem with using the basic psychophysical studies as the contrast to the speech perception studies was that neither the nonspeech stimuli nor nonspeech categorization tasks were adequately matched to their speech counterparts (see Pisoni, in press). More recently, a large number of experiments using nonspeech analogs of speech have demonstrated that subjects can perceive continuously varying stimuli categorically even though they reportedly hear the stimuli as nonspeech events, such as tones or beeps. Such demonstrations of categorical perception for nonspeech signals have shown that generic psychophysical principles can be invoked to account for categorical perception; perception may be discontinuous, ostensibly without reference to articulatory knowledge. Accordingly, these studies have attempted to account for categorical perception of speech stimuli with reference to general auditory processing of acoustic stimuli, whether speech or nonspeech.

In two well-known studies, Lisker and Abramson (1964, 1967) demonstrated that categorical perception between voiced and voiceless stops (i.e., /b/ vs. /p/, /d/ vs. /t/, /g/ vs. /k/) is determined by voice-onset time (VOT). VOT is the interval between the burst release at the articulators and the onset of voicing. In voiceless stops, there is typically a long lag between the burst release and voicing; in voiced stops, the lag is shorter and may even be negative (such that voicing begins before the stop is released). The findings that the temporal coordination of these articulatory gestures determined categorical perception are consistent with an articulation-based mode of speech perception. However, similar findings have been obtained in experiments using nonspeech materials. Miller, Wier, Pastore, Kelly, and Dooling (1976) created nonspeech VOT analogs by generating stimuli that contained aperiodic noise bursts followed by periodic buzzing. The time between the noise and the buzz was varied in small steps, following Lisker and Abramson's earlier VOT experiments. Subjects asked to classify the stimuli according to a "noise" vs. "no noise" decision showed categorical perception and discrimination functions very similar to those found for speech stimuli. Similarly, Pisoni (1977) employed stimuli that were even less speech-like than those employed by Miller et al. (1976) and still observed categorical perception. Pisoni presented stimuli composed of only two tones, one at 500 Hz and one at 1500 Hz, that varied in their temporal ordering such that the low tone either preceded or followed the high tone by as much as

50 ms, with graded steps in between. The resultant categorical identification and discrimination functions closely resembled those obtained by Lisker and Abramson (1967). In addition, Jusczyk, Pisoni, Walley, and Murray (1980) found that infants also perceive the two-tone stimuli categorically, just as they do for speech stimuli (Eimas et al., 1971).

Beyond categorical perception of stop consonants, comparisons of the perception of speech and nonspeech signals have revealed that other phenomena believed to demonstrate specialized speech processing can be accounted for by general auditory mechanisms. In a study of the effect of perceived speaking rate on phonetic classification, Miller and Liberman (1979) generated a series of synthetic speech stimuli ranging from /ba/ to /wa/ by gradually changing the duration of the formant transitions leading into the steady-state vowel formants. The important manipulation in this experiment involved varying the duration of the syllables, and thereby varying the perceived speaking rate of the syllables. Miller and Liberman found that as perceived speaking rate was increased, subjects' category boundaries shifted toward /wa/, implying that at faster speaking rates, listeners interpret shorter transitions as /w/. Miller and Liberman accounted for these data by proposing that specialized perceptual mechanisms compensate for changes in speaking rate in the perception of stops versus glides. Eimas and Miller (1980) also demonstrated the same compensatory phenomenon with infant subjects, implying that the specialized mechanism is innately specified.

However, just as in the categorical perception findings described above, it was found that the presumed perceptual "compensation for speaking rate" could be obtained using nonspeech analogs of speech. Pisoni, Carrell, and Gans (1983) generated nonspeech analogs (three component tones) of the Miller and Liberman (1979) stimuli. Subjects categorized these stimuli with either "gradual onset" or "abrupt onset" labels, and displayed a category boundary shift dependent upon duration that bore a striking resemblance to the speech data. Figure 3 shows data collected by Pisoni et al. (1983) in a replication of the Miller and Liberman (1979) study using speech materials, and in the analogous experiment using nonspeech materials. From these data, Pisoni et al. (1983) suggested that postulation of specialized, rate-sensitive mechanisms for speech may be unwarranted. Instead, they argued that "... context effects in discrimination may simply reflect the operation of fairly general auditory processing capacities... (pg. 320)." Indeed, Oller, Eilers, and Ozdamar (1990) have recently proposed a simple psychophysical model based on linear regression to account for the compensation effect. Finally, Jusczyk, Pisoni, Reed, Fernald, and Myers (1983) replicated the findings of the Eimas and Miller (1980) experiments, showing that two-month old infants exhibit the boundary shift for nonspeech materials as well as speech materials.

Insert Figure 3 about here

The results of these speech-nonspeech comparison studies imply that the specialized mechanisms proposed to account for speech perception data may be unwarranted. However, it should be noted that, despite the studies demonstrating the similarities of speech and nonspeech perception, important differences between these modes of perception have nevertheless been observed (see Pisoni, in press). A number of empirical demonstrations have shown that when listeners are induced to process auditory signals in a "speech mode" (for instance, when they are told they will hear poor quality synthetic speech and should label the stimuli using phonetic categories), their perception changes markedly. These demonstrations are typically provided by between-subjects experiments using a common pool of perceptually ambiguous stimuli that can be heard as either speech or nonspeech, depending upon the listener's expectations. In one condition, subjects are told they will hear synthetic speech and in the other they are told they will hear, for example, tones. After some performance measure is collected from

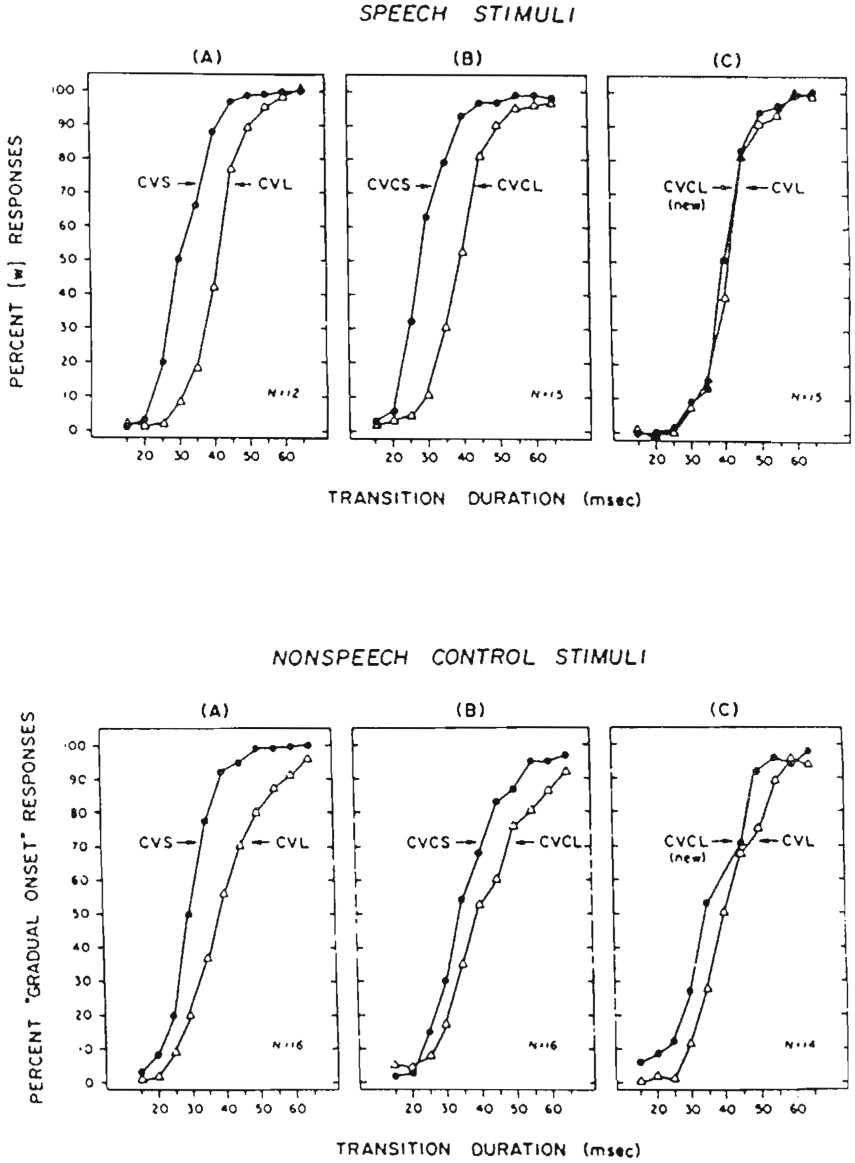


Figure 3. Data from experiments conducted by Pisoni, Carrell, and Gans (1983). The top panel shows categorization functions for synthetic speech stimuli, which replicate the findings reported by Miller and Liberman (1979). The bottom panel shows similar categorization functions for nonspeech stimuli that subjects heard as tones. (Adapted from Pisoni, Carrell, & Gans, 1983, with permission of the publisher.)

subjects, they are typically queried to ensure that they actually thought the stimuli sounded like either speech or tones, depending upon their assigned group.

In an experiment in which subjects were presented with the "sine-wave" sentence, "Where were you a year ago?", Remez, Rubin, Pisoni, and Carrell (1981) found that simply informing subjects that a signal was speech changed subjects perception from a series of whistles and beeps to a correctly transcribed sentence (see also Bailey, Summerfield, & Dorman, 1977; Grunke & Pisoni, 1982; Tomiak, Mullennix, & Sawusch, 1987). Of course, these results may be do to either a qualitative change in perceptual set or any number of changes in response biases. In an experiment that leaves less room for a bias interpretation, Schwab (1981) found substantial backward masking and upward spread of masking for sine-wave stimuli heard as tones. However, all masking was eliminated when subjects heard the stimuli as speech.

The major difference between perception in the "speech mode" versus the "nonspeech mode" for these ambiguous stimuli appears to be the difference between wholistic and componential analysis. It seems that listeners in the speech mode spontaneously forego detailed spectral analysis of the stimuli and make their speech categorizations based on entire, complex configurations of cues.[3] Conversely, subjects in a nonspeech mode behave more analytically, and actually "hear out" the component parts of the stimuli individually. Further evidence for this was provided by Tomiak, Mullennix, & Sawusch (1987) in an experiment using the Garner (1974) speeded classification task. When told they would classify nonspeech patterns, subjects were able to separately process the component dimensions of a set of noise-tone analogs of fricative-vowel syllables. Irrelevant variation in the noise spectra did not affect reaction times for the classification of tones. However, when subjects were told the stimuli were synthetic fricative-vowel syllables, the components were processed in an integral fashion, such that irrelevant variation of either dimension increased reaction times to classify stimuli along the other dimension. Finally, in an experiment reported by Grunke and Pisoni (1982), subjects were asked to identify ambiguous stimuli with either phonetic or acoustic labels, depending on their assignment to conditions. The stimuli were composed of either one, two, or three component tones. In the one- and two-tone conditions, subjects who used acoustic categories ("rising" vs. "falling") performed better than subjects who used phonetic categories. However, when a third tone was added, acoustically-based classifications were greatly reduced and phonetically-based classifications were substantially improved. Apparently, the third tone made the signal more speech-like to those listeners in the speech mode, and more noisy to listeners in the nonspeech mode.

From these and similar findings, it is apparent that the speech and nonspeech modes of perception differ in fundamental, qualitative respects. However, the basis of these differences remains to be explained. Are the differences due to the selective operation of different perceptual modules, response strategies, or attentional capacities? This question carries great theoretical importance, and will certainly merit deeper investigation in future research.

In summary, studies comparing speech and nonspeech perception have repeatedly called into question the strong claims regarding the specialized nature of speech perception. However, the evidence and arguments on either side have been equivocal, and the implications of these studies are often subject to interpretation. To appreciate the degree to which the meaning of these studies is "in the eye of the

[3] Another interpretation may be that subjects in a ''speech mode'' process components of the signal separately, but in accordance with well-learned combinatorial expectations.

beholder," compare the conclusions from two recent reviews of the speech-nonspeech literature on categorical perception:

> The nonspeech studies to this point do more than just refute the view that categorical perception is specific to speech. They demonstrate that there are certain important similarities in the ways certain classes of speech and nonspeech sounds are perceived. (Jusczyk, 1986, page 43).
>
> In summary, despite a few suggestive results, there is no conclusive evidence so far for any significant parallelism in the perception of speech and nonspeech. (Repp, 1983a, page 50).

Duplex Perception. "Duplex perception" refers to a phenomenon, first discovered by Rand (1974), that has recently been cited as strong evidence for a dissociation of phonetic perception from more general auditory perception (Liberman, 1982; Liberman & Mattingly, 1985, 1989; Repp, 1982; Studdert-Kennedy, 1982). The general procedure for eliciting the duplex percept is simple: A listener is presented with two simultaneous, dichotic stimuli. To one ear, an isolated third-formant transition that sounds like a nonspeech chirp is presented. At the same time, a "base" syllable is presented to the other ear. This base syllable consists of the first two formants, complete with transitions, and the third formant without a transition. Typically, the transition presented in isolation completes the syllable to create a /da/ or /ga/, sometimes in graded steps along a continuum. Figure 4 shows typical stimulus materials for such a procedure. When the base and the transition are presented dichotically, the listener's percept is "duplex," such that the completed syllable is perceived categorically, and the nonspeech chirp is heard at the same time. Liberman and Mattingly (1989) argue that separate modules, the phonetic module and another general auditory module, each respond to different aspects of the stimuli, creating the duplex percept.

Insert Figure 4 about here

Beyond the general duplex perception effect, several further findings support the claim that truly segregated modes of processing are responsible for the separate percepts. For example, in one study, Mann, Madden, Russell, and Liberman (1981; reported in Liberman, 1982) presented a series of different third-formant transitions such that, upon fusion, the entire syllables consisted of a continuum from /da/ to /ga/. When subjects attended to the nonspeech side of the percept, continuous discrimination functions typical of nonspeech were obtained. When subjects attended to the speech percept, categorical discrimination functions were obtained, implying that the separate percepts are subserved by separate processing systems. In other experiments, it has been demonstrated that various stimulus or procedural manipulations can affect either the speech or nonspeech percept independently, again implicating separate processors (e.g., Bentin & Mann, 1990; Isenberg & Liberman, 1978; Nygaard & Eimas, 1990).

Taken together, these findings on duplex perception appear to support the claim of an independent and specialized phonetic recognition system. As Repp (1982) concludes:

> Duplex perception phenomena provide evidence for the distinction between auditory and phonetic modes of perception. They show that, in the duplex situation, the auditory mode can gain access to the input from the individual ears, whereas the phonetic mode operates on the combined input from both ears. The "phonological fusion" discovered by Day (1968)-- two dichotic utterances such as "banket" and "lanket" yield the percept

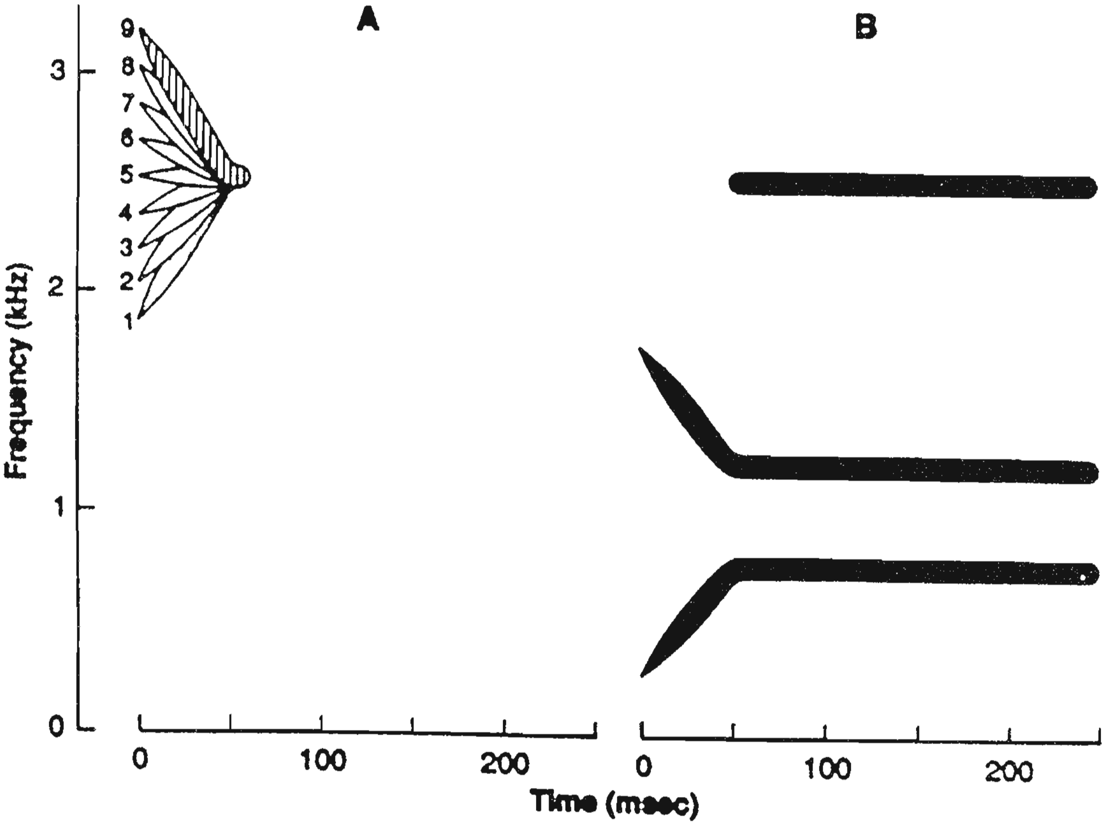


Figure 4. Stimuli presented dichotically to yield duplex perception. (A) shows a series of third-formant transitions, ranging from /ga/ to /da/ when combined with (B), the constant remainder of the syllable. (From Liberman & Mattingly, 1989, with permission of the authors and publisher.)

> "blanket"-- is yet another example of the abstract, nonauditory level of integration that characterizes the phonetic mode. (Repp, 1982, page 102).

Similarly, Whalen and Liberman (1987) describe the phenomenon of duplex perception as evidence for the "preemptiveness" of speech, arguing that the speech module provides the "first crack" at interpreting an auditory signal:

> ...the phonetic mode takes precedence in processing the transitions, using them for its special linguistic purposes until, having appropriated its share, it passes the remainder to be perceived by the nonspeech system as auditory whistles. (Whalen & Liberman, 1987, page 171).

These interpretations of duplex perception, however, are not universal; several lines of counter-evidence have been offered. Pastore, Schmeckler, Rosenblum, and Szczesiul (1983) observed duplex perception for musical chords (two notes in one ear, a third note in the other), casting doubt on the claim that duplex perception is a solely speech-based phenomenon. In addition, Nusbaum, Schwab, and Sawusch (1983) demonstrated that listeners are able to use the information contained in the third-formant transition independently of the base, casting doubt on the claim that the transition in isolation is a true nonspeech signal (see, however, Repp, 1984, and Nusbaum, 1984). This finding by Nusbaum et al. (1983) implies that subjects in the studies reported earlier could have generated their phonetic decisions without any process of auditory fusion between the two ears.

An especially strong challenge to the "speech module" interpretation of duplex perception comes from a recent study conducted by Fowler and Rosenblum (1990). Borrowing language and ideas from Gibson's (1966) event perception, Fowler and Rosenblum argue that duplex perception may not demonstrate the preemptiveness of speech per se, but simply the preemptiveness of *any* meaningful event. The argument is humans' sensory systems have evolved so that we can recognize important objects and events around us ("affordances," in Gibson's, 1979, terminology). As such, our perceptual and cognitive systems are naturally attuned to perceive meaning from any stimulation. Accordingly, Fowler and Rosenblum predicted that duplex perception should occur whenever two acoustic fragments are presented that, when integrated, specify a natural event and when one of the fragments has any unnatural quality. Fowler and Rosenblum (1991) write:

> Under these conditions, the integrated event should be preemptive and the intense fragment should be duplexed *regardless* of the type of natural sound-producing event that is involved, whether it is speech or nonspeech, and whether it is profoundly biologically significant or biologically trivial. (Fowler & Rosenblum, 1991, pages 51-52).

In order to demonstrate duplex perception for a biologically trivial event, Fowler and Rosenblum dichotically presented a low-pass filtered recording of a slamming metal door to one ear, and the remaining high frequency noise to the other ear. Alone, the "base" sounded like a *wooden* door slamming, and the "chirp" sounded to the authors like a can of rice being shaken (recall that we are biased to perceive sounds as events). When the stimuli were played together and the "chirp" was played at a higher amplitude than the "base", most subjects reported the duplex perception of "metal door + chirp." This demonstration of duplex perception for such a completely nonspeech signal calls into question either the relevance of duplex perception to speech research or the specialized nature of speech perception (or both). Findings such as Fowler and Rosenblum's clearly underscore the need for deeper investigation into the duplex perception phenomenon before it is too richly interpreted.

Trading Relations and Integration of Cues. A third class of experimental findings that have been cited as evidence for the specialization of speech perception come from studies of cue trading and cue integration (see review by Repp, 1982). It has been known for many years that the speech signal is replete with cues to phonetic contrasts, and that several different cues may indicate a single contrast (e.g., Delattre, Liberman, Cooper, & Gerstman, 1952; Denes, 1955; Harris, Hoffman, Liberman, Delattre, & Cooper, 1958; Hoffman, 1958; Repp, 1982). This aspect of the speech signal makes cue trading relations possible such that, when the utility of one cue to a phonetic contrast is attenuated, another cue may become the primary cue to the contrast. It is assumed that such trading relations occur because the cues are phonetically equivalent with respect to the contrast in question. The cues may trade in importance when necessary, or integrate to provide robust contrasts when all cues are provided equally. Examples of cue trading have been provided by Denes (1955) and in the studies by Fitch, Halwes, Erickson, and Liberman (1980) that demonstrated the perceptual equivalence of closure durations and first-formant transitions in signalling the contrast between minimal pairs such as "slit"/"split" (see Repp, 1982).

Phonetic trading relations have been cited as evidence for a speech mode of perception primarily for two reasons. First, trading relations can occur between both spectral and temporal cues that are distributed over relatively long intervals. Repp (1982) argues that it is hard to imagine that such cues would be integrated into a single percept unless some speech-specific system were mediating perception. Repp argues further that the knowledge listeners must possess to enable them to integrate such disparate cues is abstract articulatory knowledge (see also Liberman & Mattingly, 1985). Repp (1982) suggests that

> ...trading relations may occur because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic formation by perceiving the most plausible articulatory act. This explanation requires that the listener have at least a general model of human vocal tracts and of their ways of action. (Repp, 1982, page 95).

The second reason that trading relations have been cited as evidence for the speech mode of perception comes again from comparisons of speech and nonspeech perception. Best, Morrongiello, and Robson (1981) reported two experiments using sinewave speech[4] that showed that listeners in a speech mode exhibit cue trading and integration whereas listeners in a nonspeech mode do not. Best et al. considered these findings proof that the integration and perceptual equivalence of multiple cues is specific to speech.

The conclusion that trading relations provide incontrovertible evidence for speech-specific processing has not gone unchallenged. For example, Massaro and Oden (1980; see also Derr & Massaro, 1980; Oden & Massaro, 1978; Massaro, 1972, 1987; 1989; Massaro & Cohen, 1976, 1977) have presented a model of speech perception (see below) that accounts for trading relations while making no assumptions of specialized processing. Massaro and Oden argue that multiple features corresponding to a single phonetic contrast are extracted independently from the speech waveform and are integrated multiplicatively into a unitary percept. The weight given to each feature in this integration is determined by the strength, or certainty, of the feature's presence. By Massaro and Oden's account, speech perception, reduces to a "prototypical instance of pattern recognition" (Massaro & Oden, 1980, pg. 131).

[4] Recall that sinewave speech may be heard as either speech or nonspeech, depending primarily upon the listener's expectation.

Repp later (1983b) arrived at a conclusion similar to that of Massaro and Oden, stating that trading relations

> ... are not special because, once the prototypical patterns are known in any perceptual domain, trading relations follow as the inevitable product of a general pattern matching operation. Thus, speech perception is the application of general perceptual principles to very special patterns. (Repp, 1983b, page 132).

In short, as in the earlier debates regarding speech versus nonspeech perception and duplex perception, the available evidence provided by trading relations is ambiguous with respect to claims of a specialized speech mode of processing.

Cross-Modal Cue Integration (The McGurk Effect). Another recent finding in speech perception that has been attributed to specialized speech-perceiving mechanisms is the phenomenon of cross-modal cue integration, or the "McGurk effect" (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976; Roberts & Summerfield, 1981; Summerfield, 1979). The phenomenon is one of perceptual illusion, and is demonstrated as follows: A subject is presented with a video display of a talker (or synthesized face; see Massaro & Cohen, 1990) articulating simple CV syllables. At the same time, the listener hears spoken syllables that are synchronized with the visual display. The McGurk illusion occurs when the visual and auditory syllables are incongruous. In these cases, the listener typically reports hearing neither the spoken syllable nor the lip-read syllable, but something in between. For example, when presented with a face that articulates /ga/ and an auditory syllable /ba/, most subjects report hearing /da/. According to subjective reports of those who have witnessed the procedure, the effect is quite striking. Liberman (1982) points out that the procedure affects listeners' experience of *hearing* the syllable as an integrated event, to an extent that listeners cannot determine the degree to which their perception of the syllable's identity is due to either source of information. For example, Repp (1982) reports,

> I have experienced this effect myself (together with a number of my colleagues at Haskins) and can confirm that it is a true perceptual phenomenon and not some kind of inference or bias in the face of conflicting information. The observer really believes that he or she hears what, in fact, he or she only sees on the screen; there is little awareness of anything odd happening. (Repp, 1982, page 102, Footnote 8).

The McGurk illusion has been interpreted as particularly strong evidence for a specialized speech perceptual system that makes reference to articulatory gestures. For example, Fowler and Rosenblum (1991) speculate, "Why does integration occur? One answer is that both sources of information, the optical and the acoustic, provide information about the same event of talking, and they do so by providing information about the talker's phonetic gestures (page 104)." However, there are detractors to this position. Massaro and Cohen (1983) have shown that their fuzzy-logical model of perception provides precise accounts of the McGurk and MacDonald data without postulation of any speech-specific mechanisms. Also, the generality of the phenomenon is limited. Easton and Basala (1982) found that the illusion is not invoked if whole words are used instead of syllables. The suggestion made by these findings and Massaro's model is that the illusion is the product of general perceptual biases that are revealed by the highly ambiguous stimulus presentation. Finally, it should be noted that one of the principle tenets of cognitive psychology is that humans routinely perform intricate information processing that may involve any number of stages, computations, heuristics, or biases without any awareness of the operations they perform. Accordingly, despite the impressions that listeners have regarding the illusion, we should note that just because a phenomenon seems "truly perceptual" does not allow us to conclude by fiat that the results cannot be due to biases (Neisser, 1967; Cutting, 1987).

Two further findings related to cross-modal integration, however, do seem to tip the scales back in favor of a specialized-processing account. Miller (1990) cites 4- and 5-month old infants' sensitivity to auditory-articulatory correspondence as strong evidence for innately specified perceptual mechanisms (Kuhl & Meltzoff, 1982; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). Kuhl and Meltzoff (1982), for example, found that infants prefer to watch a display of an articulating face if the accompanying spoken syllables matched the articulation, compared to incongruent audio-visual displays. Finally, in a clever experiment, Roberts and Summerfield (1981) used the McGurk phenomenon in a test of selective adaptation. Roberts and Summerfield presented subjects with an auditory syllable /ba/ and visual syllable /ga/, producing the percept of /da/. However, on a test of adaptation, the perceived audio-visual syllable had the same effects as a purely auditory /ba/ on a /ba/-/da/ series; subjects' phonetic perception of the stimulus as /da/ was not reflected in their adaptation data. Studdert-Kennedy (1982) considers this finding as a powerful indication of the dissociation of general auditory and phonetic perception,

> I take [the procedure of] audio-visual adaptation to demonstrate unequivocally the on-line dissociation of auditory and phonetic perception. Moreover, following Summerfield (1979), I take the results of the audio-visual adaptation study to demonstrate that the support for phonetic perception is information about the common source of acoustic and optical information, namely, articulatory dynamics. (Studdert-Kennedy, 1982, page 7).

Studdert-Kennedy's interpretation of the adaptation data may be correct. Alternatively, we may assume, as in an information-processing model of speech perception (e.g., Cutting & Pisoni, 1978), that the pathway from audition to phonetic perception is composed of processing stages (see also Studdert-Kennedy, 1974, 1976). The locus of the phonetic perception of the McGurk paradigm and the locus of the adaptation effect could be separated, such that the adaptation manipulation affects some stage of processing that precedes the audio-visual integration. This does not seem unlikely. Presumably, the integration of information from vision and audition occurs somewhat late in the speech perception process. As such, the Roberts and Summerfield (1981) data may not imply a strict "auditory versus phonetic" dissociation; the adaptation stimulus could simply affect pre-categorical phonetic perception, an explanation that would be equally compatible with either a motor theory or an auditory theory. Out of all this, perhaps the only firm conclusion that can be drawn is that the McGurk effect, like duplex perception, may eventually constitute compelling evidence for specialized speech perception based on articulatory gestures. For the present, however, more complete investigation of these phenomena is clearly necessary.

Role of Linguistic Experience in Speech Perception. An important, but neglected, issue relevant to the question of specialization concerns the role of linguistic experience on adult speech perception (see Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). It has long been known that infants have the ability to categorically discriminate among not only the set of phonemes that constitute the inventory of their native language, but among many other nonnative phonemes as well. With continued linguistic experience, however, the listener's ability to discriminate between speech sounds that are not phonemically contrastive in his or her native tongue seems to be virtually eliminated (for review, see Aslin, 1985; Aslin & Pisoni, 1980; Logan, Lively, & Pisoni, 1991; Pisoni, Logan, & Lively, in press; Strange & Jenkins, 1978). Maturation appears to "pare down" the set of all possible contrasts (or at least most; see Best, MacRoberts, & Sithole, 1988) that listeners can originally discriminate to only the set required for the native language. Evidence for the language-specific discrimination abilities of adults was first provided by the research of Lisker and Abramson (1964, 1967; see also Abramson & Lisker, 1967). Lisker and Abramson investigated the abilities of speakers of varying languages to perceive three sets of synthetic speech stimuli that formed continua along the dimension of VOT that corresponded to three places of articulation: labial, velar, and palatal. The results of their experiments demonstrated that, in general, subjects from different linguistic backgrounds identified and discriminated the stimuli according to the contrastive phonological categories of their languages. The cross-language identification functions obtained by Lisker and Abramson (1967), which are shown in Figure 5, demonstrate the influence of the native language on perceptual classification.

Insert Figure 5 about here

Beyond the mere influence of the native phonemic repertoire on the *typical* identification of speech sounds, many studies have demonstrated the inflexibility of the adult listener's phonemic categories. It has often been reported that training an adult speaker of one language to discriminate reliably between phonemes of another language is very difficult and requires extensive training to obtain even small and unreliable improvements (Strange, 1972; Strange & Dittmann, 1984; Strange & Jenkins, 1978; Vinegrad, 1972). From data such as these, it was argued that the development of phonetic categories may require a plastic neural substrate that becomes less flexible after a critical period has elapsed (e.g., Eimas, 1975). This view of the nature and development of phonetic categories is clearly compatible with the assumption, recently defended by Liberman and Mattingly (1989), that speech perception is modular. Fodor (1983) describes modules as innately specified, neurally hardwired, and non-modifiable. Fodor's hypothesis is compatible with the view that speech perception is subserved by perceptual/memory systems that are flexible only in infancy, becoming autonomous and impenetrable as early as possible.

More recent research, however, has demonstrated that significant improvements in discrimination of nonnative phonetic contrasts can be obtained using laboratory training procedures. In one experiment, Pisoni, Aslin, Perey, and Hennessy (1982) trained English speaking subjects to perceive three categories along a VOT continuum where only two categories naturally exist. A more recent example comes from a description of training procedures employed by Logan, Lively, and Pisoni (1991; see also Pisoni, Logan, & Lively, in press) to teach Japanese listeners to discriminate /r/ from /l/. Previous research has shown that training listeners to distinguish these phonemes is extremely difficult and usually produces only marginal results (e.g., Goto, 1971; Mochizuki, 1981; Strange & Dittmann, 1984; see also MacKain, Best, & Strange, 1981). However, it could be argued that neither the stimulus materials nor the training procedures employed in most of these studies were ideal for the purposes of teaching listeners the intended contrast. For example, Strange and Dittmann (1984) attempted to use laboratory training procedures to teach Japanese listeners to discriminate /r/ and /l/. Their methods failed and they concluded that training procedures may be ineffective in modifying the phonetic categories of adult listeners. However, several aspects of their methodology call this global conclusion into question. First of all, the stimuli presented to subjects were synthetic tokens of "rock" and "lock," and no other stimuli were used in training. Also, the procedure employed during training was a standard same-different discrimination task with limited feedback provided.

In the more recent experiments conducted by Logan et al., natural tokens of minimal pairs of words contrasting /r/ and /l/ were employed in training and testing. Furthermore, to provide listeners with more robust categories than previous training studies provided, Logan et al. presented tokens produced by a variety of talkers. In Addition, the target phoneme (/r/ or /l/) occurred in a variety of phonetic environments. Clearly, these natural and variable tokens contain far more cues and more ecological validity that the synthetic tokens employed in earlier studies. Using these varied materials and a training procedure that provided extensive feedback, Logan et al. observed substantial improvements in discrimination for all of their subjects. Similar preliminary results have been reported by Pruitt,

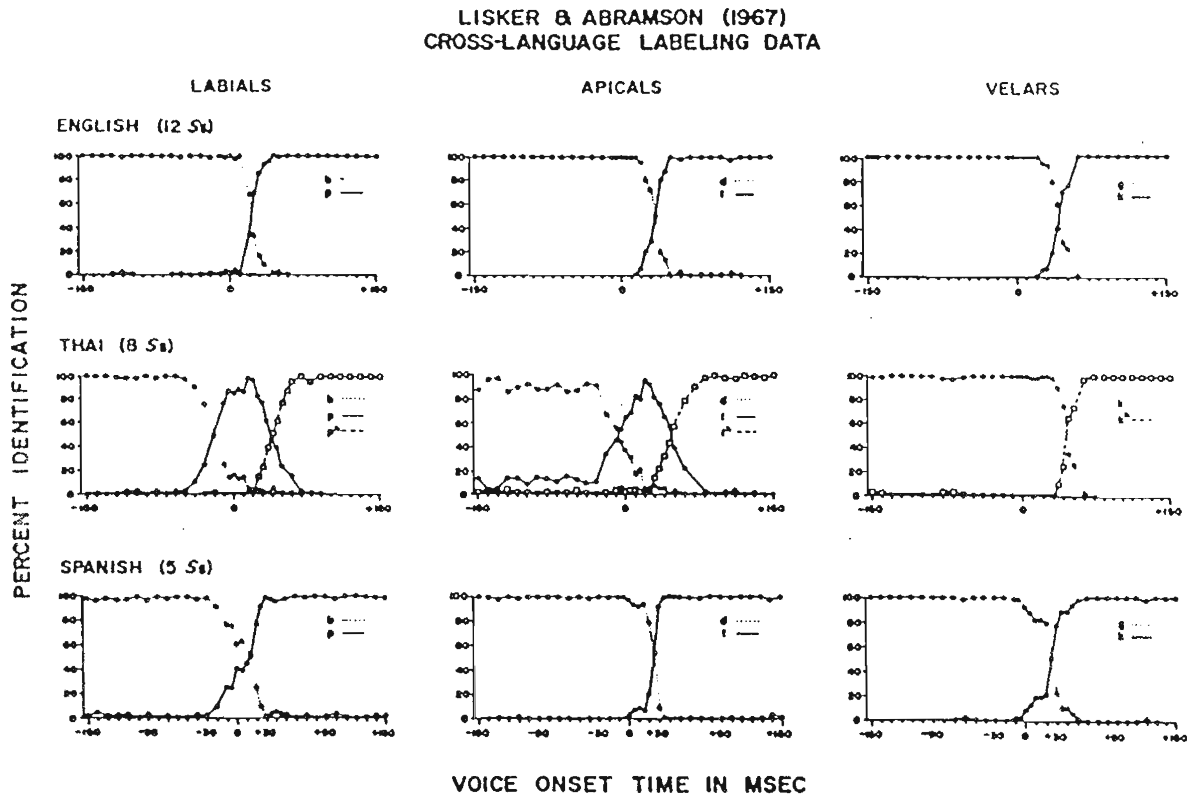


Figure 5. Cross-language identification data reported by Lisker and Abramson (1967) for labial, apical, and velar stops ranging in voice onset time from -150 to +150 ms. (Adapted from Lisker & Abramson, 1967, with permission of the authors.)

Strange, Polka, and Aguilar (1990) in training English listeners to distinguish Hindi retroflex-dental consonants.

Findings such as these, as well as a large body of developmental data (see e.g., Aslin & Pisoni, 1980), have prompted several researchers (e.g., Jusczyk, 1985, 1986) to propose that phonological categories develop and are maintained by general attention and categorization mechanisms. These theories assume that the phonological inventory for any given language can be derived by selectively attending to relevant (contrastive) dimensions while simultaneously selectively ignoring variation along irrelevant dimensions. Nosofsky (1986, 1987) has shown that this kind of selective attention strategy applied in simple category learning tasks can account for a wide variety of findings in the literature on categorization, perceptual identification, and the nature of psychological similarity. Logan et al. (1991) make reference to these attentional mechanisms to explain their learning data. This account also seems to adequately predict the failures of the earlier studies. Moreover, these recent proposals imply that the processes involved in perceiving speech rely on general cognitive principles of pattern recognition, attention, and categorization rather than highly specialized mechanisms that are unique to speech perception. At the same time, however, we cannot be certain whether these training procedures affect early phonetic perceptual processes or some later decisional processes. Clearly, we are still a long way from complete understanding of these issues, especially the developmental aspects of phonetic perception. For the present, however, we can maintain that the adult's phonetic categories do not appear to be completely rigid, as has been suggested, and that the flexibility that has been observed is consistent with a view of speech perception that employs general cognitive mechanisms.

Studies of Speech Perception in Non-Humans. One final area of research that merits consideration in our discussion of the specialization debate is speech perception by nonhuman animals. The logic that motivates such research is simple: When strong claims were made that categorical perception was a speech-specific phenomenon, researchers set out to demonstrate categorical perception of nonspeech signals. Similarly, when claims were made that categorical perception reflected a uniquely human, speech-specific mode of perception, researchers set out to demonstrate that nonhuman animals with auditory systems roughly analogous to the human auditory system could also perceive speech sounds categorically. Clearly, animals do not perceive phonetic content from human speech sounds, so any discrimination or categorization data provided by the animals must reflect general auditory and classification processes.

In studies of speech discrimination by monkeys, Morse and Snowdon (1975) and Waters and Wilson (1976) found preliminary evidence that monkeys perceive place of articulation categorically. More convincing evidence was provided in experiments conducted by Kuhl and Miller (Kuhl & Miller, 1975, 1978) on perception of speech by chinchillas. In experiments that tested categorization (as indicated in an avoidance-conditioning task), it was demonstrated that for stimulus continua that varied in VOT, chinchillas' categorization boundaries were remarkably similar to those for human (English) listeners. Finally, in a recent study, Kluender, Diehl, and Killeen (1987) demonstrated that Japanese quail can learn apparently robust phonetic categories for stop consonants /b/, /d/, and /g/. The quail learned the stops in CV syllables followed by four different vowels, and were later able to discriminate the three stops in the context of eight novel vowels, a generalization implying some form of abstraction of the category.

What are we to make of such demonstrations of speech perception in non-human animals? The results are certainly suggestive; if nothing else, they imply that, given an auditory system similar to the human auditory system and a rudimentary ability to distinguish between stimuli, animals tend to differentially respond to speech signals that correspond well to natural phonetic categories. This, in turn, implies that one need not hypothesize specialized, articulatory-based perceptual mechanisms to account for human speech perception. Unfortunately, the results of the animal studies can be taken *only* as suggestive evidence. There is no reason to assume, for instance, that human languages would have evolved phonetic contrasts that were especially difficult for our auditory systems to discriminate (Stevens, 1972). The animal data may simply illustrate that our current phonetic categories are evolutionarily well-conceived. Furthermore, since we have no access to the animal's experience, we have no basis for assuming that anything speech-like is being perceived at all. In short, examining their behavior is rather like examining the behavior of a Turing machine-- just because it looks the same as ours does not mean it is subserved by the same underlying mechanisms (see also Repp, 1983a).

Finally, what are we to make of the entire debate regarding the specialization of speech perception? As we have seen from our discussions of perception of speech versus nonspeech, duplex perception, cue trading, the McGurk effect, cross-language studies, and animal studies, there are arguments to be made for both sides of the issue. This debate has proven fruitful for the sake of continuing research-- more data has been generated and energy devoted to its resolution than to any other issue in speech perception. At the same time, it remains possible that the specialization hypothesis may be empirically unassailable. This may be especially true now that the specialization mechanism has been described in terms of the modularity hypothesis. Fodor (1985) describes several necessary characteristics that any experiment must have if it is to be considered as a bona fide counter-example to the modularity of a perceptual system: (1) The experiment must demonstrate the influence of "background information" (higher cognitive processes) on the perceptual output. (2) The effect of this information must clearly involve the perceptual system; it cannot reflect post-perceptual processing or a decisional criterion shift. (3) The "cognitively penetrated" system must be the usual system for natural perception in the given domain, not involving some backup systems that are required only in special circumstances, such as in perceiving degraded stimuli. Consider, for example, the finding that mere instructions alter the percept of sinewave speech from a sequence of tones into a sentence (Remez et al., 1981). At first glance, this would appear to violate the impenetrable nature of the phonetic module, whose operations are supposed to be impervious to the listener's various beliefs and expectations. (Fodor's preferred examples are optical illusions, such as the Mueller-Lyer illusion, which persists even when the observer *knows* that the lines are of equal length.) Clearly the sinewave speech demonstration would satisfy condition (1) above, but (2) and (3) are questionable. Furthermore, almost any experiment aimed at demonstrating the non-specialized nature of phonetic perception may fail to satisfy at least one of these criteria. The challenge for future research is to address the relevant issues while circumventing these pitfalls. We have no insights to offer at present, but we remain optimistic that the weight of the evidence will eventually tip the scales unambiguously one way or the other on the specialization issue.

Normalization Problems in Speech Perception

As described above, the problems posed for theories of speech perception by the inherent non-linearity, variability, and non-segmental nature of the speech signal arise from the general assumption that the listener must somehow map distorted information in the speech signal onto linguistic representations in memory. Typically, researchers in speech perception have limited their study of variability to the effects of different phonetic contexts. However, it has long been realized that many factors above and beyond phonetic contexts influence the acoustic realizations of phonetic contrasts. Collectively, the perceptual accommodation of variations in speech patterns to recover canonical linguistic units falls into the category of *perceptual normalization*. Typically, research on normalization has focussed on sources of variation such as talkers' vocal tract differences and speaking rate differences. (Although the "problem of perceptual constancy" is also introduced by a speaker with a mouthful of food, a singing voice, etc.)

Individuals differ in terms of the sizes and shapes of their vocal tracts (Fant, 1973; Joos, 1948; Peterson & Barney, 1952), glottal characteristics (Carr & Trill, 1964; Carrell, 1984; Monsen & Engebretson, 1977), their idiosyncratic articulatory strategies for producing individual phonemes (Ladefoged, 1980), as well as the dialects of their native regions. As such, there is wide variability in the production of the same words and phrases across individuals. Nevertheless, human listeners appear to accurately perceive speech across virtually all (reasonably intelligible) speakers without any apparent difficulty. At present, little is known about the perceptual processes that are responsible for the implied perceptual compensations, nor is it known whether perceptual compensation actually occurs at all.

A second, related aspect of the normalization problem concerns time and rate normalization. Speech is a temporally distributed signal and, as such, the cues to individual phonetic contrasts in speech are distributed in time and are substantially influenced by alterations in speaking rate. Moreover, the acoustic durations of segments are further influenced by the locations of syntactic boundaries in fluent speech, by syllabic stress, and by the component features of adjacent segments (see Gaitenby, 1965; Klatt, 1975, 1976, 1979; Lehiste, 1970). Segmental durations are modified further still by contextual factors in speech. For example, vowels of words spoken in sentences are approximately half the duration of vowels of the same words spoken in isolation (see Luce & Pisoni, 1987, for fuller discussion). In sum, phonetic contrasts in conversational fluent speech is characterized by widespread durational variation. Furthermore, it is well-known that some of this durational variation in speech carries important information about numerous phonetic contrasts, word boundaries, etc. In English, for instance, numerous phonetic contrasts are distinguished by durational cues. Thus, the listener is somehow able to attend to and use durational cues to stress, phonemic contrasts, and pragmatics while simultaneously ignoring irrelevant durational variations due to particular talkers or circumstances (see Miller, 1980; Port, 1977).

Indexical Information in Speech. The human voice conveys information about a speaker's age and gender, as well as more culturally-oriented information such as the speaker's regional origin, temperament, and social group membership. Such aspects of speech, known as *indexical information* (Abercrombie, 1967), do not, in general, relate directly to processes of phonetic perception (other than providing still more variability) but are heavily relied upon in linguistic communication nonetheless. For example, most of us are reasonably expert at discriminating a New England accent from a Japanese accent, just as we are reasonably expert at discriminating the speech patterns of children from those of adults. Finally, indexical information functions to alert the listener to speaker identity and to important changes in the physical or emotional state of the talker (Ladefoged & Broadbent, 1957, refer to these aspects of the voice as "personal information").

Several examples may reveal the pervasive use of indexical information in everyday communication. The example already given is our ability to infer quite extensive information about a speaker's origin and background-- in many societies, speech patterns are commonly associated with social status (Abercrombie, 1967). Aside from cultural speech patterns, speech patterns of any individual speaker are richly informative: We are remarkably sensitive to a speaker's emotional or physical state (within our own cultures), and we can readily identify people we know from their voice alone. We also recognize "signatures" of voice; for instance, most of us could identify even a poor impersonation of W.C. Fields or Porky Pig. Finally, the entire realm of vocal changes we know as "tone of voice" are pervasive in communication and are readily perceived as we recognize anger, depression, or joy in someone's speech. Occasionally, tone of voice may even serve to modify the semantic content of an utterance, such as in a sarcastic comment. Finally, beyond our common experience, research has demonstrated that listeners incidentally store detailed information about speakers' voices and implied connotative states when listening to speech (Geiselman & Bellezza, 1977; Geiselman & Crawley, 1983).

Given these facts, two apparently contradictory questions are introduced in consideration of talker variability. The first concerns the listener's ability to recognize the segments of the language despite the idiosyncratic variability introduced by each new voice. The second concerns the listener's ability to simultaneously exploit such variability to perceive the characteristics of the talker and the communicative situation.

Talker Variability in Speech Perception and Word Recognition. Although the problem of talker variability was described at least as early as 1948 (Joos, 1948), one of the first empirical demonstrations of the effects of talker variability was provided by Ladefoged and Broadbent (1957; although see Peters, 1955a, 1955b). Ladefoged and Broadbent presented listeners with the synthesized sentence, "Please say what this word is:" followed by either *bit*, *bet*, *bat*, or *but*. The carrier phrase was altered in different conditions by raising (by 30%) or lowering (by 25%) either the first or second formants, or both. This manipulation had the effect of changing the perceived dimensions of the talker's vocal tract. Ladefoged and Broadbent observed reliable changes in subjects' identification of the target syllables depending on the perceived talker. The authors concluded that the carrier phrase allowed the listener to "calibrate" the vowel space for each talker, and to adjust their perceptions of the target vowels accordingly (see also Gerstman, 1968). Following this early demonstration, a large number of studies were conducted toward the goal of investigating and explaining the relative constancy of natural vowel perception across talkers (see Johnson, 1990; Shankweiler et al., 1977). The guiding motivation for all such studies was the idea that listeners must somehow extrapolate the entire vowel space of any given talker from a small speech sample (e.g., Joos, 1948; Lieberman, Crelin, & Klatt, 1972).

In further research, however, Verbrugge, Strange, Shankweiler, and Edman (1976; see also Shankweiler et al., 1977) questioned the premise of this approach. First of all, they noted that despite talker variability, listeners' error rates in vowel identification tasks are rather low (e.g., only 4% in Peterson and Barney's, 1952, experiments). Verbrugge et al. re-examined vowel identification across talkers and found that accuracy of vowel identification is generally quite high, despite talker variability. They also found that providing examples of a speaker's point vowels did not improve listener's performance, in contrast to earlier notions of calibration. Finally, they found that listeners adjust their perceptual criteria according to perceived rate of articulation much more than for perceived length of vocal tract. From all these data, Verbrugge et al. (1976) concluded that talker normalization is either a process that requires very little prior information or is not a process that occurs in speech perception at all (see also Strange, Verbrugge, Shankweiler, & Edman, 1976). Instead, they suggested that adjustment to talkers may have more to do with tracking articulatory dynamics than with frequency-based calibration. In later research, Verbrugge and Rakerd (1986) investigated the perception of "silent-center" bVb syllables. Verbrugge and Rakerd presented listeners with /bVd/ syllables spoken by males and females, that had the middle 60% removed, leaving only the beginning and ending transitions with silence in between. Their results showed that considerable vowel identity information is contained in the transitions, and that this information is talker-independent. From their data, Verbrugge and Rakerd conclude,

> This strongly suggests that a dialect's vowels can be characterized by higher-order variables (patterns of articulatory and spectral *change*) that are independent of a specific talker's vocal tract dimensions. (Verbrugge & Rakerd, 1986, page 56).

The fundamental claim of these reports-- that talker normalization involves recovery of underlying articulatory dynamics-- is familiar from the discussions of specialization above. These findings imply that variability introduced from individual talker characteristics may be resolved in the same manner as all other acoustic-phonetic variability. This treatment of perceptual normalization finds support from studies of development as well. Experiments conducted by Kuhl (1979) and by Kuhl and Miller (1982) demonstrated that 6-month-old infants could accurately discriminate vowels produced by three different talkers; the infants did not attend to the changing voice characteristics of the stimuli, but focused on vowel identity instead. Similar results were obtained by Holmberg, Morgan, and Kuhl (1977), who found that infants' discrimination of fricatives was not affected by talker variability (however, see Carrell, Smith, & Pisoni, 1981). Finally, in a recent study, Jusczyk, Pisoni, and Mullennix (1989) examined the effects of talker variability on infants' discrimination of CVC syllables. Jusczyk et al. found that infants could discriminate syllables, such as /bug/ and /dug/, as well in multiple-talker conditions as in single-talker conditions. All of these findings are consistent with a notion that some specialized system, perhaps a phonetic module sensitive to articulatory gestures, are involved in talker normalization, just as has been hypothesized with regard to more general problems of perceptual constancy in speech (e.g., Liberman & Mattingly, 1985, 1989).

The data pertaining to the effects of talker variability are not completely unequivocal, however. Essentially, the claims of non-effects of talker variability come from tasks of perceptual identification or discrimination of stimuli with few additional attentional or time constraints. Despite these demonstrations of the listener's remarkable accuracy in perceiving speech from varying talkers, several experiments have shown reliable effects of talker variability on speech perception and word recognition. In a preliminary study, Creelman (1957) investigated the effects of talker variability on the recognition of phonetically-balanced words that were presented in lists of tokens spoken by either 1, 2, 4, 8, or 16 talkers. Results of perceptual identification of these words in noise showed that words in lists produced by two or more talkers were recognized slightly less accurately (differences on the order of 7% -- 10%) than words in the single-talker list.

Later experiments with larger sets of stimulus materials, however, have shown larger effects of talker variability. Summerfield and Haggard (1973; see also Summerfield, 1975) observed slower reaction times to spoken words in multiple-talker blocks than in single-talker blocks. More recently, Mullennix, Pisoni, & Martin (1989) investigated the effects of talker variability on word recognition using a large sample of CVC monosyllables. Words were presented in lists spoken by either one talker or by fifteen talkers and subjects performed either perceptual identification of words in noise or auditory naming of non-degraded words. Mullennix et al. observed large and reliable effects of talker variability; word recognition was less accurate and slower in multiple-talker conditions than in single-talker conditions. Moreover, talker variability was observed to be *more* robust and less sensitive to changes in task demands than other variables known to affect word recognition, such as word frequency and similarity neighborhood density (see Luce, 1986; Luce, Pisoni, & Goldinger, 1990). Finally, Mullennix et al. also found that talker variability interacts with signal degradation, implying that noise and talker variability affect a common stage of processing. From these data, Mullennix et al. suggested that talker variability affects early stages of speech perception responsible for immediate phonetic perception.

Talker Variability in Memory and Attention. Further insights into the nature of talker variability effects are provided by recent experiments in memory and attention. Martin, Mullennix, Pisoni, and Summers (1989) investigated serial recall of ten-item word lists spoken by either a single talker or by ten different talkers and found that recall for multiple-talker lists was impaired. Specifically, Martin et al. found that recall of multiple-talker lists was less accurate than recall of single-talker lists, but only for items in early list positions. Moreover, they found that recall of visually-presented digits presented *before* the presentation of the spoken lists was less accurate if the subsequent lists were multiple-talker lists than if they were single-talker lists. Finally, they found that the differences in recall between the lists were unaffected by a post-perceptual distractor task (following Peterson & Peterson, 1959). From these converging lines of evidence, Martin et al. suggested that word lists produced by multiple talkers require greater attentional resources for rehearsal in working memory than the same lists produced by a single talker.

Further evidence of the attention-demanding nature of talker variability was provided by a study conducted by Goldinger, Pisoni, and Logan (1991), in which single-talker and multiple-talker lists were presented for serial recall at varying presentation rates. Goldinger et al. found that talker variability interacted strongly with presentation rate whereas other stimulus variables, such as word frequency, did not. At relatively fast presentation rates, recall of single-talker lists was superior to recall of multiple-talker lists, as in the Martin et al. experiments. At very slow presentation rates, however, recall of multiple-talker lists was actually *more accurate* than recall of single-talker lists. The presentation rate manipulation has long been assumed to affect the rehearsal processes of the recall task (Murdock, 1962; Rundis, 1971), so this result suggests that talker variability taxes these attention-demanding stages of processing. A final interesting finding reported by Lightfoot (1989) was that subjects' *familiarity* with the talkers' voices modifies the differences in recall of single- and multiple-talker lists even further. When subjects were trained to recognize the voices of the various talkers and associate them with fictional names (e.g., Brad, Mary, Jane, Sam, etc.), multiple-talker lists were recalled better than single-talker lists, even at relatively fast presentation rates.

The effects of talker variability on memory have been studied in infants by Jusczyk, Pisoni, and Mullennix (1989; see above). As Kuhl and her colleagues had reported earlier, Jusczyk et al. observed that infants recognize phonemic constancy very well despite variation of the stimulus voices. However, Jusczyk et al. also employed a variation of the high-amplitude sucking (HAS) procedure (Eimas et al., 1971) that included a two-minute delay period between the habituation to one syllable and the presentation of a new syllable. This manipulation allowed Jusczyk et al. to assess the effects of talker variability on infants' ability to encode and remember phonetic structure. It was found that infants who heard speech from a single talker were able to detect a phonetic change across the two-minute delay but the infants who heard speech from multiple talkers were not. These results, taken together with the adult data, suggest that maintaining perceptual constancy across talkers requires attentional resources.

A final demonstration of the influence of talker variability on selective attention was provided in a recent study by Mullennix and Pisoni (1990), who employed the Garner (1974) speeded classification procedure to investigate processing dependencies between phonetic variability and talker variability. Subjects classified monosyllabic words according to either the voicing of the initial phoneme (/b/ versus /p/) or the voice of the talker (male versus female). They found that irrelevant variations in neither phonetic constitution nor talker's voice could be ignored; variation along either dimension slowed classifications based on the other dimension. However, a large asymmetry was observed, showing that variability along the voice dimension impaired classification along the phonetic dimension more than vice versa. These data suggest that the processing of voice information and phonetic information are qualitatively different, but dependent on one another in that they share a limited-capacity cognitive system (see Cutting & Pisoni, 1978). Mullennix and Pisoni suggest that both indexical information and phonetic information are processed in a mandatory fashion, following Fodor (1983; for a similar suggestion see Miller, 1987). However, the implied modules may function as a cascade system (McClelland, 1979), such that the output of the phonetic module is affected by the output of the "voice processor."

In summary, the available data on the effects of talker variability in speech perception, word recognition, attention, and memory all appear to indicate that indexical information deserves more thorough consideration in theoretical discussions of speech perception than it has traditionally received. Talker-related information affects perception of speech, affects memory of spoken material, attracts selective attention, and is routinely encoded in parallel with linguistic information (see Geiselman & Bellezza, 1976, 1977). The traditional approach to the study of speech perception has considered only the abstract linguistic units without regard to the effects of the media that carry them. Further investigation into the generality and nature of normalization effects in speech should provide valuable insights into not only speech perception, but perhaps the architecture of general perceptual systems as well.

Prosody and Timing in Speech Perception. Along with the role of indexical information in language perception, another somewhat neglected topic has been the role of prosodic information in language perception. *Prosody* refers to the melody, timing, rhythm, and amplitude of fluent speech, and is typically thought of in terms of changes in the acoustic correlates of stress, such as F_0 and vowel duration (e.g., Lehiste, 1970; Huggins, 1972). Most of the emphasis in speech perception research and theory has been on the segmental analysis of phonemes, whereas the suprasegmental information has received only cursory consideration. Although the role of prosody has been researched more vigorously in recent years, a wide gap remains between the research conducted on the perception of isolated segments and features and on sentences with full prosody and rhythm (see Cohen & Nooteboom, 1975). However, it has become apparent that prosodic factors may serve to link phonetic segments, features, and words to grammatical processes at higher levels of analysis. Moreover, prosody also appears to provide useful information about the lexical, syntactic, and semantic content of the spoken utterance. We briefly review several findings that illustrate the importance of prosodic information in the perception of connected speech (see Darwin, 1975; Huggins, 1972; Nooteboom, Brokx, & de Rooij, 1978; Studdert-Kennedy, 1980; for more extensive reviews).

Empirical evidence suggests that differences in fundamental frequency can provide important cues to the proper parsing of speech into constituents suitable for syntactic analysis. In acoustic analyses of connected speech, Lea (1973) found that a drop in fundamental frequency usually occurred at the end of each major syntactic constituent of a sentence, while a rise in fundamental frequency occurred in the beginning of the following constituent. In more detailed analyses, Cooper and Sorenson (1977) found reliable rise-fall patterns at the boundaries between the main clauses of a sentence, as well as between main and embedded clauses, and between major phrases. Lindblom and Svensson (1973; see also Svensson, 1974) have shown that listeners can parse speech that maintains prosodic integrity but is devoid of segmental cues (see also Nakatani & Schaffer, 1978). These findings and others (e.g., Collier & 't Hart, 1975; Cooper, 1976; Klatt, 1976; Klatt & Cooper, 1975) demonstrate the importance of prosody as a cue to phrasal grouping.

Another function of prosody appears to be the maintenance of perceptual coherence (Studdert-Kennedy, 1980). As an example, Darwin (1975) had listeners shadow a sentence played to one ear while another sentence was presented to the other ear. At some point, the prosodic contours of the sentences were switched, while the lexical, syntactic, and semantic content of the sentences remained unchanged. Shadowing often spontaneously followed the prosodic contour across ears, rather than the syntax or semantics of the message they were originally attending. Nooteboom, Brokx, and de Rooij (1978) suggest that prosodic contours maintain the "perceptual integrity" of the signal, and provide evidence that the continuity of fundamental frequency and formant frequencies seems to underlie this integrity (see also Bregman, 1978, for a discussion of the formation and maintenance of auditory perceptual "streams").

A number of studies conducted by Cutler and her colleagues (Cutler, 1976; Cutler & Darwin, 1981; Cutler & Fodor, 1979; Cutler & Foss, 1977) have demonstrated yet another important function of prosody in speech perception. Cutler's work has shown that prosodic contours enable listeners to predict where sentence stress will fall. Because sentence stress is usually placed on words of primary semantic importance in a sentence, the ability to predict stress placement presumably guides attention to the most important words in the sentence. Thus, prosody appears to guide attention to high-information stretches of fluent speech. To demonstrate that attention follows the prediction of sentence stress provided by prosody, Cutler and her colleagues have demonstrated that phoneme monitoring reaction times are faster to words that are *predicted* by prosodic contour to receive stress, despite the word's actual acoustic realization or form class. A word placed in a sentence position of predicted stress is responded to faster than the same recorded token placed in another sentence position.

These demonstrations of the role of prosody in guiding attention have led Cutler and others to propose accounts of word recognition in which prosody is considered a primary source of information, rather than marginally relevant variability (see Bradley, 1980; Cutler, 1976, 1989; Grosjean & Gee, 1987). These approaches all emphasize the prominence of strong syllables in fluent speech and suggest that such syllables may serve to focus attention and initiate segmental analysis and lexical access. This approach contrasts with more temporally constrained, left-to-right models of speech perception and word recognition, such as cohort theory (Marslen-Wilson, 1987; Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), that assume that word beginnings are necessarily processed first (see below). As Cutler (1989) summarizes:

> ... the major problem for lexical access in natural speech situations is that word starting points are *not* specified. The evidence presented here has shown how prosodic structure, in particular metrical prosodic structure, can offer a way out of this dilemma. Where do we start with lexical access? In the absence of any better information, we can start with any strong syllable. (Cutler, 1989, page 354).

Finally, despite the benefits of prosody, all of these useful prosodic cues make the familiar problems of acoustic-phonetic invariance far more problematic. The durations of phonetic segments vary widely in stressed and unstressed syllables, as well as in varying syntactic environments (Oller, 1973; Klatt, 1974, 1975; Luce & Charles-Luce, 1985). Spoken stress also involves wide spectral variations in formant frequencies as well as fundamental frequency (Lehiste, 1970). The durational variations of speech timing appear to provide useful cues to lexical identity as well as syntactic structure, but at the cost of further removing anything resembling "canonical phonemes" from the signal. This potentially contradictory nature of prosodic information underscores the necessity of some theoretically sound resolution of the "problem" of invariance-- apparently, as more meaningful variation is added to the signal, perception is *improved* rather than impaired.

In the preceding sections of this chapter, we have identified and discussed many of the long-standing issues in speech perception as well as several of the more current issues that researchers have recently explored. We now turn away from these general issues and focus on individual theories and models of speech perception. We briefly introduce and comment only on a few models in the literature, as a comprehensive review is beyond the scope of this chapter (see Klatt, 1989, for a recent and more extensive review). In our brief review, however, we hope to include a representative subset of some of the most important and influential classes of theories that have been proposed, particularly those theories that should figure prominently in future research.

Theoretical Approaches to Speech Perception

The perception of spoken language, encompassing all processes from peripheral auditory processing of the speech signal to comprehension of the speaker's message, is clearly a very complex process. Many sources of knowledge and multiple levels of representation are involved and interact in myriad combinations. The complexity of language has, to date, precluded the formulation of theories of language perception that are at once global and empirically testable. As such, the situation in language perception research is similar to other areas of cognitive science; investigators have typically examined only the details of specific phenomena and paradigms, rather than more complex or integrative issues of their field (see Newell, 1973).

The remaining sections of this chapter illustrate this situation by their unfortunate dichotomy. In the first part we review several models of speech perception and in the second part we review several models of spoken word recognition. This segregation is largely a reflection of the orientation of the models themselves. Although there are a few notable exceptions (e.g., Klatt's LAFS model and the TRACE model), the majority of the models described below were formulated to explain either the identification of phonemes in the speech signal or the mapping of strings of phonemes onto lexical representations in memory. Very few models are specified in enough detail to specify the integrated processes of speech perception *and* word recognition (see Pisoni & Luce, 1987, for further discussion). One trend, especially present in the connectionist movement, has been toward grouping these processes into unitary models. Another trend, however, has been toward justifying the segregation of processes considered within different models by arguing that the processes are segregated in perception. The concept of the phonetic module (Liberman & Mattingly, 1985, 1989) clearly justifies narrow consideration of phonetic perception, without regard to the mapping of speech representations onto lexical representations. Although we believe the former trend will prove more fruitful in the long run, we recognize the value of the earlier models proposed and discuss them next, beginning with the most influential of all models of speech perception, the motor theory of speech perception.

Motor Theory of Speech Perception

The original motor theory described by Liberman et al. (1967) was based on the assumption that "speech is perceived by processes that are also involved in its production (page 452)." This view of speech perception was motivated by the fact that a listener is also a speaker, that a close link exists between the acoustic forms of speech sounds and their underlying articulation. As such, an effective and economical means of perceiving speech is to perceive the articulatory gestures that produce sounds. Advocates of the motor theory argue that a solution to the invariance problem lies in the more reliable nature of articulatory gestures (compared to acoustic phonemes) as units of perception.

Although the original motor theory held a dominant position in accounts of speech perception for many years, the link between the theory and the available data was rarely more than suggestive. As the review of the evidence in the section *Specialization of Speech Perception* showed, the evidence cited in support of motor theory is potentially ambiguous. For example, much of the early support for motor theory came from the finding that synthetic stop consonants were perceived categorically, whereas steady-state vowels were perceived continuously, apparently paralleling their respective articulatory origins. Subsequent research, however, demonstrated that the differences in perception between consonants and vowels were primarily due to their differing demands on auditory short-term memory (Fujisaki & Kawashima, 1969, 1970, 1971; Pisoni, 1971, 1973, 1975). This general cognitive explanation of the continuous-categorical distinction eliminated the need for reference to articulation in perception of stops.

Recently, the motor theory has been revised in two key regards (Liberman & Mattingly, 1985, 1989). First, whereas the original model was based on recognition of observable gestures, the revised model is based on perception of *intended* gestures. Gestures are defined as a set of movements by the articulators that result in a phonetically relevant vocal tract configuration. Each intended gesture of the language has properties that specify it uniquely, and each intended gesture is invariant, such that each segment of the language maps uniquely to a distinctive gesture. The second important modification to the theory is that gestures are perceived directly (following Gibson, 1966) by an innate phonetic module.

The revised motor theory makes four basic claims that were considered in turn by Klatt (1989): The first claim is that speech production and perception are linked psychologically, such that they share common representations and processes. Second, the basic unit of speech perception is the underlying *intended* articulatory gesture associated with a phonetic segment, rather than the actual physical motions implied by the acoustics. Third, perception of the intended gesture is direct, performed by a specialized module. Fourth, the model is supported by the claim that no other model can account for the wide array of phenomena that the motor theory has been applied to over the years.

With respect to the link between production and perception, Klatt (1989) agrees that the processes must be linked in some sense (as inverses, at least), but that there is no simple way to relate the processes that makes articulatory perception any less problematic than acoustic perception. Considering the direct perception of intended gestures, Klatt notes that, while the position is attractive and would solve many problems of variability, there are simply no mechanisms described in the theory that could perform this feat. Furthermore, Klatt argues that our present technology demonstrates the extreme difficulty of determining vocal tract shapes from speech acoustics, but the motor theory is based on faith that this transformation is actually possible. Unlike the premises of Newtonian mechanics, we can not be sure that speech is a "reversible" event. Finally, regarding the uniqueness of the motor theory in accounting for a variety of phenomena, Klatt argues that the revised motor theory is so abstract that it is essentially no different from acoustic-based theories, such as LAFS (see below), and he therefore suggests that the account is no longer unique.

Klatt concludes that "An attractive motor theory *philosophy* has been described by Liberman and Mattingly, but we are far from the specification of a motor-theory *model* of speech perception (page 180)." His point is well-taken; the motor theory and the revised theory are based primarily in logic, parsimony, and intuitive appeal, and a measure of faith, rather than empirical support.

Direct-Realist Approach to Speech Perception

Recently, Fowler (1986; Fowler & Rosenblum, 1990, 1991) has outlined the framework for a *direct-realist* approach to speech perception. The direct-realist approach assumes that, as in Gibson's (1966) view of visual perception, the perception of speech entails the recognition of natural "phonetic events." As in the motor theory, Fowler assumes that the relevant events that are perceived in speech are the speaker's phonetically structured articulations, but the theories are not identical. In the language of event perception, there is a fundamental distinction between the event and the informational medium. For example, an object, such as a chair, is an "event" in the world. When our eyes gaze upon the chair, we perceive it via reflected light that has been provided structure by the edges, contours, and colors of the chair. We do not perceive the *light*, per se. Instead, the light functions merely as the medium by which the chair is perceived. The suggestion for speech perception is very similar to this example--articulatory events lend unique structure to the acoustic waveform just as chairs lend structure to reflected light. Accordingly, Fowler suggests that articulatory events are directly perceived via the acoustic medium.

The direct-realist approach to speech perception is obviously similar to the motor theory in many respects. However, there are important differences. Most notably, the two theories approach the signal in different ways. Motor theory maintains that the acoustic signal is subjected to computations that retrieve underlying gestures. In contrast, the direct-realist approach maintains that no cognitive mediation whatsoever is necessary; the acoustic signal is transparent with respect to the underlying structure of speech (see Liberman & Mattingly, 1985). Following this difference, Fowler and Rosenblum (1991) argue that phonetic perception need not be a modular process, suggesting instead that general perceptual principles can be invoked to perceive the distal events of speech.

The direct-realist approach and motor theory are attractive for many of the same reasons (see Studdert-Kennedy, 1986): It has intuitive appeal, it fares well with many of the data that the motor theory can explain. Moreover, it stems from a respected tradition of event perception theories. However, there are many challenges that the theory must meet as well. Most important, of course, is the need for empirical support for the theory, in which regard it is similar to the motor theory. Forgiving the present lack of critical data, many logical and theoretical challenges can be offered as well (for a review, see commentaries on Fowler's, 1986, target article). For example, as Remez (1986) notes, it is not clear what the proper perceptual objects in linguistic communication actually are. Fowler has adopted a physical perceptual object that is capable of structuring the acoustic media-- the articulatory gesture-- and has made its recognition the central task of speech perception. However, articulations are not ends in themselves. Unlike chairs, articulations are themselves another media because language is symbolic. Strings of articulations are perceived as words, ideas, etc. As such, gesture perception does not fully explain speech perception. Moreover, as noted by Diehl (1986), Porter (1986), and Remez (1986), chairs and gestures are also very different in terms of their perceptual availability. We know unambiguously when we are looking at a chair; we do not have such access to phonetic gestures. A direct-realist theory might claim that our unambiguous recognition of words and sentences implicitly demonstrates our recognition of gestures, but other accounts are clearly available (Massaro, 1986). The resolution of these and other theoretical vagaries, as well as the collection of relevant data, will be important for the direct-realist position to stand the test of time.

Information-Processing Theories of Speech Perception

Perhaps the polar opposite to the direct-realist perspective is the information-processing perspective. The orientation of the theories and models that fall into this category is that of general cognition and perception. All of these models assume distinctive, hierarchically-organized levels of processing. Moreover, all or most of these theories assume that limited-capacity perceptual and memory stores are intimately involved in speech analysis (see Cutting & Pisoni, 1978). This view contrasts sharply with the revised motor theory and the direct-realist framework. A typical information-processing stage model is shown in Figure 6. This model, borrowed from Pisoni and Sawusch (1975), exhibits multiple levels of representation and processes that interact with and depend upon memory stores and control processes. Beyond the stages of processing shown in the figure, once the information has followed the "output" arrow, the linguistic units that have been recognized enter still higher stages of processing that derive syntax, semantics, etc.

Insert Figure 6 about here

Studdert-Kennedy (1974, 1976) was the first to advocate an approach to speech perception based on stages of perceptual processing. He proposed four stages of speech processing: (1) auditory, (2) phonetic, (3) phonological, and (4) lexical, syntactic, and semantic (see review and discussion by Luce & Pisoni, 1987; Pisoni & Luce, 1986, 1987). As the division of Studdert-Kennedy's stages imply, the stages of processing approach to speech perception represents a synthesis of the concepts of information-processing psychology and linguistic theory. An advantage of this sort of framework is its clear division of processes of speech perception; working within such a framework provides a well-defined division of topics for investigation.

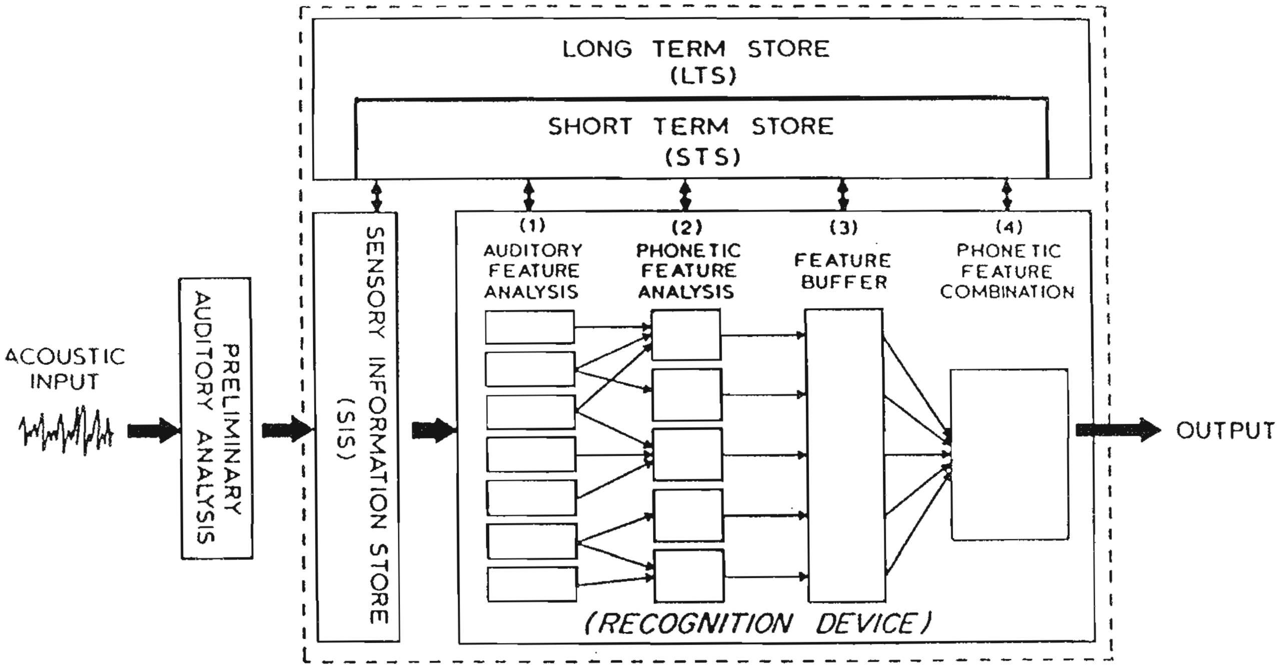


Figure 6. Functional organization of the phonetic perception component of an information-processing model of speech perception. (From Pisoni & Sawusch, 1975, with permission of the publisher.)

The appeal of stage theories has typically been their reliance on more generally accepted mechanisms of cognition and perception. As such, information-processing models introduce certain advantages over modular theories such as motor theory. For example, information-processing models are able to account for the effects of reduced attention or increased memory load on speech perception (Nusbaum & Schwab, 1986). Unfortunately, like many theories of speech perception, information-processing theories have typically been quite vague and not available for direct empirical tests.

Klatt's LAFS Model

Although we introduce Klatt's Lexical Access From Spectra (LAFS) model in this section of the chapter dedicated to models of speech sound perception, it is important to acknowledge that it is a model of spoken word recognition as well. Klatt's LAFS is one of the few models proposed that successfully addresses several of the critical issues of speech perception while simultaneously addressing issues related to accessing the mental lexicon and the nature of lexical representations in long-term memory.

Klatt's model assumes direct, noninteractive access to lexical entries based on context-sensitive spectral sections (Klatt, 1979). The LAFS model assumes that adult listeners have a dictionary of all legal diphone sequences stored in memory. Associated with each diphone sequence is its prototypical spectral representation. These spectral representations are proposed to resolve problems associated with contextual variability of individual segments. In a sense, LAFS resolves the problems of variability by precompiling coarticulatory effects directly into the representations residing in long-term memory. The listener is assumed to compute spectral representations of an input word and to compare these derived spectra to prototypes in memory. Word recognition is accomplished when a best match is found between the input spectra and the diphone representations. In this portion of the model, word recognition is accomplished directly on the basis of spectral representations of the sensory input, with no intermediate levels of computation corresponding to segments or phonemes.

An important aspect of Klatt's LAFS model, then, is its explicit avoidance of any levels of representation corresponding to phonemes. Instead, the model assumes a precompiled, acoustically-based lexicon of all possible words in a network of diphone power spectra. These spectral templates are assumed to be context-sensitive units, similar to "Wickelphones," because they represent the acoustic correlates of phonemes in different phonetic environments (Wickelgren, 1969). Klatt argues that diphone concatenation is sufficient to capture much of the context-dependent variability observed for phonetic segments in spoken words (see also Marcus, 1984). Word recognition in LAFS proceeds similarly to the workings of the HARPY speech recognition system, such that power spectra are computed every 10 msec and compared to the stored representations. When finished, the best "path" through the network represents the optimal phonetic transcription of the signal. Klatt's model is an example of an extreme bottom-up recognition process and may be contrasted to more interactive models of word recognition that we consider below, such as cohort theory or TRACE.

Massaro's Fuzzy Logical Model of Perception

Massaro's Fuzzy Logical Model of Perception (FLMP) (Derr & Massaro, 1980; Massaro, 1972, 1987, 1989; Massaro & Cohen, 1976, 1977; Massaro & Oden, 1980; Oden & Massaro, 1978) was developed to account for feature integration in speech perception, regardless of the nature of the relevant features. For example, FLMP can account for the integration of multiple acoustic cues in the speech waveform, but it can account for audio-visual integration as well. In this brief introduction to the model, however, we restrict our attention to the recovery of phonemes from the speech signal, noting only that integration of information from other sources is possible in the model and is accomplished by processes similar to those described here.

FLMP assumes three operations in phoneme identification. First, *feature evaluation* determines the degree to which any given acoustic-phonetic feature is present in a stretch of sound. Unlike more conventional feature detector theories, FLMP assumes that features are evaluated along a continuous scale, rather than along an absolute feature-present/feature-absent dichotomy. As such, features are assigned continuous, "fuzzy" values ranging from zero to one, indicating the degree of certainty that the feature actually appears in the signal (Zadeh, 1965). The second operation in FLMP is *prototype matching*, in which the feature profiles derived by the earlier operations are compared to prototypes of phonemes stored in memory. Phoneme prototypes are stored as sets of propositions that describe idealized representations of the acoustic correlates of each phoneme. The prototype matching operation specifies the degree of correspondence between these idealized phonemes and the input sets of features. The final operation, *pattern classification*, determines the best match between the candidate phonemes and the input, using "goodness of fit" algorithms. FLMP provides flexibility in pattern classification by using a variety of logical rules for feature integration, such that perfect matches between the input and the prototypes are not required for phoneme identification to proceed.

FLMP is an appealing model for several reasons. First of all, it is a very general framework that demonstrates how acoustic information (as well as other information) can be mapped onto representations in long-term memory without the postulation of specialized procedures or modules. In fact, Massaro (1987, 1989) specifically rejects the notions of specialized or modular processes in speech perception. Furthermore, the model argues that speech perception is not necessarily categorical, but can be explained by integration of continuously evaluated features. This framework is therefore consistent with the data provided by Barclay (1972), Pisoni (1973), and others that continuous information remains available in speech perception, despite the categorical identification and discrimination functions obtained in typical studies (e.g., Liberman et al., 1957). Finally, FLMP is one of the only models of speech perception that has been proposed in terms of a precise mathematical framework (Townsend, 1989). However, the quantification provided by the model has been a source of criticism as well; the model employs large numbers of free parameters to account for patterns of data and the parameter settings do not easily transfer across experimental paradigms (Jenkins, 1989; Warren, 1989). Finally, Massaro's suggestions (1987) that the FLMP framework may be extended beyond speech perception to all forms of perception (e.g., person perception) are attractive, although considerable testing and evaluation are clearly required by these claims.

Theoretical Approaches to Spoken Word Recognition

The theories and models described in the preceding section of the chapter were denoted as models of *speech perception*, meaning that they were primarily concerned with providing accounts of phonetic perception, independent of higher-level lexical or linguistic processes (with the exception of Klatt's LAFS model). In this next section of the chapter, we introduce several models of *spoken word recognition*--models that are primarily concerned with the rapid location of lexical entries in memory once the speech perception system has specified the necessary sublexical components of the input. As has been discussed above and elsewhere (e.g., Pisoni & Luce, 1987), this separation of the focus of theories is unfortunate and appears inappropriate, especially in light of the demonstrations of lexical effects on speech perception (e.g., Ganong, 1980; Samuel, 1986, Samuel & Ressler, 1986). Nevertheless, most of the models considered here assume that some input, perhaps resembling a string of phonemes, is provided by early processes of speech perception and is then compared to the mental lexicon until a best match is found. Very few models of word recognition or lexical access are concerned with the entire range of processes that subserve word recognition (a possible exception is the TRACE model described below).

The myopic nature of theories of word recognition and lexical access is primarily attributable to the origins of most of the theories. Most of the theories in the literature were originally proposed to account for findings of *visual* word recognition, so assumptions of invariance, etc. are easily justified (although most models of visual word recognition certainly allow for some variability). A very general assumption has been that models of visual word recognition can account for spoken word recognition as well, given rudimentary modifications of assumptions to respect the temporal distribution of the speech signal (see Cutler, 1989; Grosjean & Gee, 1987; Marslen-Wilson & Tyler, 1980; Tyler & Frauenfelder, 1987). While the validity of this assumption is subject to debate (Bradley & Forster, 1987), it has served to isolate the processes of word recognition sufficiently to allow for the articulation of precise, albeit simplified, theories. While ignoring questions related to the resolution of problems of early speech perception, models of word recognition have been focused primarily on explaining basic phenomena such as word frequency, context effects, types of knowledge sources brought to bear on word recognition, and the nature of representations in the mental lexicon. Indeed, these considerations largely characterize models of word recognition and have constituted the basis of extensive experimentation and debate. Also, as we found to be the case in our review of the phenomena and theories of speech perception, one of the fundamental debates in discussions of models of word recognition has been the distinction between modular and interactive processes (see Bradley & Forster, 1987; Tanenhaus & Lucas, 1987). In our discussions of the models below, we pay special attention to the models' respective approaches to all of these basic phenomena and theoretical distinctions.

In this section of the chapter, we briefly examine five models of word recognition:[5] the logogen theory, cohort theory, Forster's (1976) search theory, the neighborhood activation model, and the TRACE model. It should be noted that these models represent only a portion of the models that have appeared in the literature, but we hope that review of these selected models will capture and communicate several of the key issues in spoken word recognition. We begin with one of the earliest models of word recognition described in the literature-- the logogen theory.

Logogen Theory. In Morton's (1969, 1979, 1982) logogen theory, passive sensing devices called "logogens" are associated with each word in the mental lexicon. Each logogen contains all of the information about a given word, such as its meaning, possible syntactic functions, and its phonetic and orthographic structure. A logogen monitors discourse for any information indicating that its particular word is present in the signal, and once such information is encountered, the activation level of the logogen is raised. Upon sufficient activation, the logogen crosses a threshold, at which time the information about the word that the logogen represents is made available to the response system, and the word is said to be "recognized."

Several important features of the logogen theory have attracted attention over the years, and have either been strongly debated or incorporated into later models. First is the theory's emphasis on the interaction of multiple knowledge sources in word recognition. One important feature of the logogen theory is that logogens monitor all possible sources of information, including higher-level semantic and syntactic information from the discourse as well as lower-level sensory information. (It is important to note, however, that logogens do not "talk to each other," meaning that any given logogen is oblivious to the activity levels of other logogens.) Thus information from several levels can combine to push the

[5] In the remainder of this chapter, the following terminological distinction is employed. When we refer to *word recognition*, we mean only the recognition of an acoustic-phonetic pattern as a token of a given word represented in memory. When we employ the term *lexical access*, we refer to the moment in time when all information about the recognized word becomes available to working memory (see Morton, 1969; Pisoni & Luce, 1987).

activation level of a logogen toward its threshold. In this sense, logogen theory is a highly interactive model of word recognition, and the effects of context are incorporated into the early stages of word recognition. Words that are readily predicted by the semantic and syntactic context are activated, and therefore recognized, more quickly than words that are not well-predicted by context. A second important feature of the logogen theory is its portrayal of word frequency effects. In the logogen theory, frequency differences among words are represented as adjustments in the recognition thresholds of the words' logogens. Thus, a word of high frequency has a threshold lower than another word of lower frequency, and therefore requires less sensory or contextual input for recognition to occur. The characterization of word frequency as a direct coding in either recognition thresholds, resting activation levels, or activation functions has been adopted in many later models of word recognition (e.g., Marslen-Wilson, 1987).

Taken together, the two major assumptions of the logogen theory place the word recognition stage as the locus of both context and frequency effects. The approach is highly interactive, and its portrayal of context effects has been challenged by theorists who prefer a more modular approach to language processing (e.g., Bradley & Forster, 1987; Forster, 1979, 1990). Likewise, the theory characterizes word frequency as an integral and automatic aspect of word recognition. This characterization of word frequency has been challenged by theorists who argue that word frequency may be better characterized as a form of perceptual or response bias, as demonstrated by the task-dependent magnitude of frequency effects (e.g., Balota & Chumbley, 1984; Luce, 1986).

The specific details of logogen theory have changed somewhat over the years, but the basic mechanisms have remained the same. For example, Morton (1982) divided the logogen system into separate visual and auditory subsystems, but the fundamental notion of the passive threshold device that can monitor information from a variety of sources has remained. Unfortunately, like many of the theories we discuss, logogen theory is extremely vague. At best, the theory helps us conceptualize how an interactive system works and how word frequency can be accounted for, but it says very little about precisely how acoustic-phonetic and higher-level sources of information are integrated, the time-course of word recognition, or the structure of the lexicon.

Cohort Theory. Marslen-Wilson's cohort theory (Marslen-Wilson, 1975, 1980b, 1987; Marslen-Wilson & Tyler, 1975, 1980; Marslen-Wilson & Welsh, 1978) posits two stages in the word recognition process-- one autonomous and one interactive. In the first, autonomous stage of word recognition, acoustic-phonetic information at the beginning of an input word activates all words in memory that have the same word-initial information. For example, if "slave" is presented to the system, all words beginning with /s/ are activated, such as "sight," "save," "sling," and so forth. The words activated on the basis of word-initial information constitute a "cohort." Activation of a cohort is autonomous in the sense that only acoustic-phonetic information can serve to specify the members of a cohort. At this stage of the model (which Marslen-Wilson, 1987, refers to as *access*), word recognition is a completely data-driven or bottom-up process.

Once a cohort is activated, all possible sources of information may come to bear on the selection of the appropriate word from the cohort. Thus further acoustic-phonetic information may eliminate "sight" and "save" from the cohort, leaving only words that begin with /sl/, such as "sling" and "slave." Note that word recognition is based on acoustic-phonetic information and is assumed to operate in a strictly left-to-right temporal fashion. At this stage of word recognition, however, higher-level knowledge sources may also come into play in eliminating candidates from the cohort. Thus, if "sling" is inconsistent with the available semantic or syntactic information, it will be eliminated from the cohort.

At this second stage of word recognition, the theory is highly interactive.[6] Upon isolation of a single word in the cohort, word recognition is accomplished.

An important feature of cohort theory is its sensitivity to the temporal nature of speech. As such, it gives priority to the beginnings of words and assumes strict left-to-right processing of acoustic-phonetic information. Cohort theory also embraces the notion of "optimal efficiency" (Marslen-Wilson, 1980a, 1987; Tyler & Marslen-Wilson, 1982), a principle stating that the word recognition system selects the appropriate word candidate from the cohort at the theoretically earliest possible point (the "recognition point"). This means that the word recognition system will commit to a decision as soon as sufficient acoustic-phonetic and higher-level sources of information are consistent with a single word candidate.

Although earlier discussions of cohort theory made no mention of word frequency, in recent modifications of the theory, Marslen-Wilson (1987) has suggested that frequency in cohort theory operates in a similar fashion to the logogen theory. Specifically, Marslen-Wilson has proposed that word recognition may not require absolute elimination of all members of a cohort, but merely a comparison of relative activation levels among candidates (following Luce, 1986), with the activation levels being modified by the same activation-elimination processes described in earlier discussions of the theory. Word frequency is assumed to modify the individual *rates* of activation of the words constituting the cohort, with higher frequency words becoming more active faster than lower frequency words. Like the logogen theory, then, cohort theory portrays word frequency as an integral aspect of the early phases of word recognition.

Marslen-Wilson's cohort theory has attracted a considerable amount of attention in the last few years for several reasons, including its relatively precise description of the word recognition process, its novel claim that all relevant words in the mental lexicon are activated in the initial stage of the word recognition process, and because of the priority it affords to the beginnings of words, a popular notion in the literature (see also Cole & Jakimik, 1980). The theory is not without its shortcomings, however, both theoretically and empirically. First, Marslen-Wilson (1987, 1989; Warren & Marslen-Wilson, 1987) has argued that the theory requires no conventional linguistic units, such as phonemes, in order to function. In order to maintain optimal efficiency, he proposes that the word recognition system exploits coarticulatory information that crosses phonemic boundaries (e.g., nasalization of a vowel that precedes a nasal consonant) in real-time, thus avoiding unnecessary decisional delays. Unfortunately, the data on this point are somewhat ambiguous, and the argument could be made that nasalization of a vowel is primarily a cue to phonemic, rather than lexical, identity. Affording priority to the lexical cue may be efficient, but it may not be correct. Also, it is not clear that candidates can be efficiently eliminated from the cohort without the use of phonemic dichotomies (see Pisoni & Luce, 1987, for further discussion).

Another theoretical problem with cohort theory is error recovery. For example, if "foundation" is perceived as "thoundation" due to mispronunciation or misperception, the word-initial cohort will not, according to the theory, contain the word candidate "foundation." Although Marslen-Wilson allows for some residual activation of acoustically similar word candidates in the cohort so that a second pass through the cohort structure may at times be possible to attempt a best match, it is still unclear how error recovery is accomplished when the intended word is not a member of the original activated cohort.

[6] At least to a degree. In his revisions of the original cohort theory, Marslen-Wilson (1987) has suggested that the effects of top-down context on the word selection process may be limited, perhaps in such a way that context can only have a facilitatory effect for consistent words, but not an inhibitory effect for inconsistent words.

Finally, several experimental results have challenged some of the cohort theory's stronger assumptions, especially the concept of maximally early decisions in word recognition. For example, although preliminary evidence from the gating task (Grosjean, 1980; Tyler, 1984) supported the notion of early isolation points, later experiments showed that many words are not recognized until well after their acoustic offsets in continuous speech (Bard, Shillcock, & Altmann, 1988; Grosjean, 1985). Also, the cohort theory predicts that the time it takes to decide that an item is a nonword in a lexical decision task should be a function of its "isolation point," meaning the point in the stimulus at which the item could not constitute an English word (e.g., "lotato" should be rejected faster that "potavo"). However, Goodman and Huttenlocher (1988) and Taft and Hambly (1986) have shown that lexical decisions are not reliably predicted by isolation points. Despite these problems, however, cohort theory remains one of the most important theories in spoken word recognition, primarily because it was initially developed to explain spoken word recognition rather than visual word recognition, and it therefore respects the temporal nature of speech.

Forster's Autonomous Search Theory. In contrast to Morton's logogen theory and Marslen-Wilson's cohort theory, Forster's (1976, 1979) theory of word recognition and lexical access is autonomous in the strictest sense. Also, whereas Morton's and Marslen-Wilson's theories allow parallel processing of information at some stage, in Forster's theory, linguistic processing is completely serial. The theory posits three separate linguistic processors: a lexical processor, a syntactic processor, and a message processor. In addition, the latest version of Forster's theory incorporates a third, nonlinguistic processor, the general processing system (GPS). Forster's model may be considered the word-recognition embodiment of several of Fodor's (1983) principles of modularity in perceptual processing (see Forster, 1989, 1990; Tanenhaus & Lucas, 1987), strongly emphasizing algorithmic, non-interactive processing among separate components that are hierarchically related to one another.

In the first stage of Forster's model, information from peripheral perceptual systems is submitted to the lexical processor. The processor then attempts to locate an entry in three peripheral access files: an orthographic file (for visual input), a phonetic file (for auditory input), and a syntactic-semantic file (for either form of input). Search of the peripheral files is assumed to proceed by means of a *frequency-ordered search*, with higher-frequency words being searched before lower-frequency words. Once an entry is located in the peripheral access files, the location of the word in the master lexicon becomes available. Thus word recognition is accomplished at the level of the peripheral access files, where the input pattern is matched to a stored representation. Once an entry is located in these files, lexical access is accomplished by locating the entry in the master lexicon, where other information about the word is stored.

Upon location of an item in the master lexicon, information pointing to the location of that item in the master list is passed on to the syntactic processor, which attempts to build a syntactic structure of the discourse. From the syntactic processor, information is passed to the message processor, which attempts to build a conceptual structure for the intended message. Each of the three processors-- lexical, syntactic, and message-- can pass information to the GPS. However, the GPS cannot influence processing in any of the three dedicated linguistic processors. Rather, it serves to incorporate general conceptual knowledge with the output of information from the linguistic processors in making a decision (or response). In Fodor's terminology, the linguistic processors are *vertically* organized, whereas the GPS is *horizontally* organized, meaning that the GPS, unlike the linguistic modules, integrates information from many disparate domains.

Forster's theory is therefore composed of autonomous, non-penetrable modules. First, the lexical processor is independent of the syntactic and message processors, and the syntactic processor is independent of the message processor. Furthermore, the entire linguistic system is independent of the general cognitive system, as Fodor (1983) suggests. This strictly serial and autonomous characterization of language processing means that word recognition and lexical access are not influenced in any way by higher-level knowledge sources and are exclusively bottom-up or data-driven processes. As such, Forster (1979) attempts to explain all forms of context effects as post-access effects, either in terms of decisional or response biases. Note, however, that Forster's theory, like Morton's and Marslen-Wilson's theories, posits that word frequency exerts an early effect on the word recognition processes themselves (see Forster, 1990, for a defense of this assumption).

Forster's model is attractive because of its relative specificity and the apparently testable claims it makes regarding the autonomy of its processors. The model also attempts to describe word recognition and lexical access in the context of sentence processing. In addition, it incorporates a specific explanation of the word frequency effect-- namely, that entries in the peripheral access files are organized according to frequency and that search proceeds from high- to low-frequency entries. Finally, the notion of the search mechanism lends itself well to empirical testing, although the majority of relevant data that have been reported have come from experiments in visual word recognition (see, e.g., Forster & Bednall, 1976; Andrews, 1989).

Neighborhood Activation Model. The neighborhood activation model of word recognition (Luce, 1986; Luce, Pisoni, & Goldinger, 1990; see also Cluff & Luce, 1990; Goldinger, Luce, & Pisoni, 1989) is based on the notion that word recognition reduces to a process of choosing a "best match" from a pool of activated word candidates, and is thus similar in important respects to both Morton's logogen theory and Marslen-Wilson's cohort theory. However, unlike either of these earlier theories, the neighborhood activation model makes important assumptions about the role of *competition* among activated items for recognition. Central to the predictions of the model is the concept of the *similarity neighborhood* (Luce, 1986; see also Andrews, 1989; Coltheart, Develaar, Jonasson, & Besner, 1976; Landauer & Streeter, 1973). A similarity neighborhood is defined as a collection of words resident in the mental lexicon that are phonetically similar to each other and to any given stimulus word presented for recognition. Similarity neighborhoods are characterized by two main structural characteristics: (1) *neighborhood density*, which refers to the number of words in the neighborhood and their degrees of confusability with the stimulus word, and (2) *neighborhood frequency*, which refers to the frequencies of the words in the neighborhood, relative to the frequency of the stimulus word.

In experiments employing perceptual identification of words presented in noise, auditory lexical decision, and auditory word naming, Luce (1986) observed that these structural characteristics of similarity neighborhoods strongly affected the speed and accuracy of word recognition. Words from sparse neighborhoods were recognized faster and more accurately than words from dense neighborhoods, and words from low frequency neighborhoods were recognized faster and more accurately than words from high frequency neighborhoods. Indeed, it was observed that neighborhood characteristics were more reliable predictors of word recognition than word frequency itself; for example, in the auditory word naming experiment, robust effects of neighborhood density were observed, but no effects of word frequency were evident (see Balota & Chumbley, 1984, 1990, for a discussion of the lability of word frequency effects).

In the neighborhood activation model, word recognition is achieved in a manner that is similar to both the logogen theory and the cohort theory, but with two basic modifications. The model assumes that, upon stimulus input, a set of acoustic-phonetic patterns are activated in memory. The activation levels of these patterns are assumed to be a direct function of their phonetic similarity to the stimulus input. The activated phonetic patterns activate, in turn, a system of *word decision units*, which are conceptually similar to logogens. The word decision units are activated directly and autonomously from the bottom-up information provided by the signal, as in cohort theory. Once the word decision units are activated, they monitor a number of sources of information, especially the fluctuating activation levels of the acoustic-phonetic patterns. Unlike processing in the system of logogens or the cohort, however, the word decision units also monitor the overall level of activity in the decision system itself, in a manner similar to the processing units in the TRACE model of speech perception (Elman & McClelland, 1986; McClelland & Elman, 1986; see below). Finally, the decision units are also sensitive to higher-level lexical information, including word frequency. This information serves to bias the decisions of the units by differentially weighting the activity levels of the words to which they respond. Word recognition occurs in the model when the system of decision units selects a best match from the activated neighborhood, at which time all information about the word is made available to working memory.

The neighborhood activation model places much of the burden of spoken word recognition on the discrimination among similar acoustic-phonetic patterns corresponding to words and the decisions necessary for choosing among these patterns. In this manner, the model can account for the observed effects of similarity between stimulus words and their neighbors in the lexicon. In both Morton's logogen theory and Marslen-Wilson's (1987) cohort theory, it is explicitly assumed that word recognition is independent of the number of activated candidates. As such, these models provide no explanation for neighborhood density or neighborhood frequency effects. In addition, the model accounts for word frequency by assuming that frequency information biases the decisions of the word decision units. By assuming that frequency exerts its effects in the late decision stage of word recognition rather than in the early activation of the word units, the neighborhood activation model accounts for the common observation that word frequency effects vary in magnitude across experimental tasks. As different tasks introduce different decisional requirements, the neighborhood activation model is suited to account for these results. Other models of word recognition, such as logogen theory, cohort theory, and Forster's search model propose that frequency is an integral and early contributor to word candidate activation or search order. As such, these models are not well-suited to account for experiments in which word frequency effects are attenuated or absent (e.g., Balota & Chumbley, 1984; Luce, 1986).

Despite the advantages of the neighborhood activation model over previously proposed models of word recognition, it does introduce several methodological difficulties. First, the concept of phonetic similarity among words in memory is difficult to quantify for empirical tests, and crude estimation methods, such as the *N* metric (Coltheart et al., 1976), are most commonly employed. Also, the concept of similarity is dependent on assumptions of representation. Similarity may be defined with respect to the speaker's phonetic repertoire, or with respect to the listener's idealized phonetic representations (which are unavailable for inspection). Despite these methodological difficulties, however, the concept of similarity is easily handled in theory and the empirical effects of similarity neighborhoods are robust despite crude estimation. A second shortcoming of the model in its current form is the treatment of the temporal characteristics of word recognition. Unlike cohort theory, which provides explicit accounts of the time course of word recognition, the neighborhood activation model presently offers no account of the recognition of longer, multisyllabic words (although see Cluff & Luce, 1990).

Finally, we should mention another model to which the neighborhood activation model bears noteworthy resemblance-- the activation-verification model (Paap, Newsome, McDonald, & Schvaneveldt, 1982; see also Becker, 1976, 1979, 1980, Becker & Killion, 1977). In the activation verification framework, upon presentation of a stimulus word, a pool of similar word candidates are activated, based upon coarse sensory analysis. These candidates are then subjected to a *verification* process, in which each candidate word is compared to the stimulus until a best match is determined. The verification process is similar to the search procedure in Forster's search model; candidates are submitted for verification in descending order of their frequency. By incorporating the concept of the verification set, which is much like a similarity neighborhood, the activation-verification model does account for the effects of set size and similarity among neighbors. However, the model's assumption of the frequency-ordered verification process reduces its flexibility in predictions of word frequency effects across tasks (see Dobbs, Friedman, & Lloyd, 1985).

TRACE and other Connectionist Models. The TRACE model of speech perception[7] (Elman, 1989; Elman & McClelland, 1986; McClelland & Elman, 1986) is an example of a nearly completely interactive system. Coming out of the growing connectionist movement, and based on the interactive-activation model of visual word recognition (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), TRACE advocates multiple levels of representation and rich feed-forward and feed-back connections between processing units. In addition, TRACE incorporates processes for both activation and inhibition of units in the network, as in the earlier interactive-activation framework.

The functional units in TRACE are simple, highly interconnected processing units called nodes. There are three levels of nodes in the model, such that some nodes represent phonetic features, some represent phonemes, and some represent words. When information passes upward through these levels, nodes that collect sufficient confirmatory evidence to surpass a threshold will fire and send activation along weighted links to their related nodes. In this manner, information that is consistent with the "expectations" of the early feature detectors is proliferated upward in the network to encourage recognition of the features' associated phonemes, and then recognition of the phonemes encourages recognition of the phonemes' associated words.

A key property of TRACE is the organization of excitatory and inhibitory links between nodes and levels. All connections from one level to another are excitatory, meaning that activation of a node on one level will increase the activity of all connected nodes on the adjacent level or levels. As an example, if the feature detector node for voicing encounter voicing cues consistent with /k/, then the node for phoneme /k/ will become activated, which will, in turn, activate all words in the lexicon that contain the phoneme /k/. Within levels, however, all nodes are connected by inhibitory links. This forces the model to quickly resolve ambiguity in the signal. For example, if the features for both /k/ and /g/ are encountered simultaneously, the nodes corresponding to the features and phonemes for both possibilities do not only become more activated, but they also produce greater inhibition on their nearest competitors. Computationally, the end result of this process is a "winner take all" form of perceptual decision (Elman & McClelland, 1986), meaning that the node that receives the most positive activation also receives the most "veto power" over its competitors. A final important property of the model is its use of perceptual feedback-- activation and the flow of information in the model proceeds not only from the early feature detection system to the lexicon, but the expectations at the lexical and phonemic levels can bias perception on the levels below.

[7] As we noted in our discussion of Klatt's LAFS model, whereas most contemporary ''models of speech perception'' are clearly concerned with issues in speech perception and most ''models of word recognition'' are concerned with issues in word recognition, several rare models incorporate considerations of both. The LAFS model is one of these, and we opted to include it in the section on models of speech perception above. TRACE is another model that is concerned with accounting for phenomena from both the speech perception and word recognition literature. We recognize the contribution that TRACE and similar connectionist models provide to theories of speech perception-- our decision to discuss the model in this section is simply an acknowledgment of the model's importance as a theory of word recognition.

The interactive nature of TRACE is a conceptual innovation that offers much to theories of speech perception and word recognition. McClelland and Elman (1986) list almost a dozen well-known phenomena from the speech perception literature that the model can simulate, ranging from categorical perception to trading relations, as well as findings from the word recognition literature, such as earliness of word recognition. As a model of speech perception, TRACE does not treat coarticulatory effects of speech as "noise" that is imposed on an idealized string of phonemes. Instead, Elman and McClelland (1984) refer to contextual variability as "lawful variability," which serves as a rich source of information in TRACE. (The authors like to say that "you can tell a phoneme by the company it keeps.") Also, although the model explicitly assumes segmental representations in speech, no explicit segmentation is imposed on the speech signal in the processing of the model. Instead, phones and allophones are simply assumed in the model's architecture, and thus segmentation "falls out" naturally. As a model of word recognition, the inhibitive links among nodes at the lexical level allows TRACE to account for neighborhood effects in a manner similar to the neighborhood activation model. In brief, by virtue of its simple assumptions of interacting units, TRACE demonstrates many of the attributes of theories of speech perception and word recognition in an integrated system, without postulating or proliferating restrictive rules or specialized mechanisms.

However, like all models, TRACE is not without its problems. Many of the model's problems relate to the simplifying assumptions that are made with regard to speech input. Other problems are inherent to the model itself. Among the most serious problems with the model are: 1) The model has no mechanism for predicting word frequency effects (although it is not difficult to imagine how a set of lexical level biases could be instantiated). 2) The model has no obvious way of determining when it has been presented a nonword. The ability to judge lexical status is one of the most important abilities of the human word recognition system (Forster, 1979), and should be included in any model. TRACE could accomplish this by setting criterial "confidence" values for outputs, such that unfamiliar words would be judged as nonwords, but this would confound distinctions between degraded inputs and nonwords (a discrimination human listeners make easily). 3) Perhaps the most important problem with TRACE arises from one of its most attractive features. The model acknowledges, and even exploits, variability and coarticulation in its perceptual decisions, but it does not address other sources of variability common in natural language, such as talker idiosyncracies, changing speaking rates, stress assignments, etc. Even more troubling is the fact that TRACE demands a certain degree of invariance in its variability. The model acknowledges that the cues for phonemes are not localized in specific segments, but at the same time it does require that all cues occur in a pre-determined "time window." While the difficult problems of temporally distributed cues in speech are not easily resolved in the original TRACE model, it is hoped that the new breeds of *recurrent networks* (e.g., Jordan, 1986) may alleviate some of the difficulties of working with time windows in speech.

Summary and Conclusions

In this chapter, we have attempted to identify and elucidate several of the principle issues in research and theory on speech perception and auditory word recognition. Some of the issues constitute long-standing concerns in the field. Despite their long history as empirical and theoretical issues, however, problems such as the lack of acoustic-phonetic invariance and segmentation, the problem of perceptual normalization, and the specialization of speech perception remain vital and controversial areas of research today. And, although new and innovative approaches to these issues have developed both in research and in theory, the fundamental complexity of speech perception continues to pose considerable obstacles to speech researchers. We should not expect that any comprehensive solutions to these problems are in our immediate future, but the current trends are encouraging.

A particularly encouraging trend in the field of language perception is the growing emphasis on considering speech perception and spoken word recognition as interacting stages of a unitary process. The bulk of research on speech perception over the past 40 years has been almost exclusively concerned with the perception of isolated phonetic contrasts or phonemes in brief, meaningless syllables. Although the more modular approaches to speech perception maintain that research and theory should proceed in a vacuum, the major current trend appears to be a move toward interactionism, bridging the gap that has traditionally separated the study of these different stages of spoken language understanding. Already, we have observed the development of several preliminary connectionist approaches to language processing--an approach that emphasizes the value of interaction between levels of processing. Perhaps the major insight afforded by these approaches has been the value of allowing our current models of speech perception and word recognition to constrain each other. As noted by Pisoni and Luce (1987), theorizing about one stage of language processing without regard for related stages is a somewhat myopic approach that may lead to theories that work well only when considered separately. But if our theories about one stage of processing are incompatible with what we understand about another stage, it is not clear what we have learned.

In short, we believe that the growing interest in the perception of spoken language, going beyond the level of the phoneme to the level of the word, reflects a healthy trend toward more comprehensive accounts of the early stages of language perception. Much work, of course, remains to be done on almost every level of spoken language understanding. As such, the problems of speech perception and spoken word recognition, along with all aspects of language perception, promise to provide interesting and challenging research opportunities for at least another 40 years.

References

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

Abramson, A.S., & Lisker, L. (1967). Discriminability along the voicing continuum: Cross language tests. *Proceedings of the 6th international congress of phonetic sciences*. Prague: Academia, 569-573.

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 802-814.

Aslin, R.N. (1985). Effects of experience on sensory and perceptual development: Implications for infant cognition. In Mehler, J., & Fox, R. (Eds.), *Neonate Cognition: Beyond the Blooming, Buzzing Confusion*. Hillsdale, NJ: Erlbaum, 157-183.

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

Bagley, W.C. (1900-1901). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology*, **12**, 80-130.

Bailey, P.J., Summerfield, Q., & Dorman, M. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research, SR-51/52*. New Haven, CT: Haskins Laboratories, 1-25.

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Balota, D.A., & Chumbley, J.I. (1990). Where are the effects of frequency in visual word recognition tasks? Right where we said they were! Comment on Monsell, Doyle, and Haggard. *Journal of Experimental Psychology: General*, **119**, 231-237.

Barclay, J.R. (1972). Noncategorical perception of a voiced stop: A replication. *Perception and Psychophysics*, **11**, 269-273.

Bard, E.G., Shillcock, R.C., & Altmann, G.T.M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics*, **44**, 395-408.

Becker, C.A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 556-566.

Becker, C.A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **5**, 252-259.

Becker, C.A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory and Cognition*, **8**, 493-512.

Becker, C.A., & Killion, T.H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 389-401.

Bentin, S., & Mann, V. (1990). Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes. *Journal of the Acoustical Society of America*, **88**, 64-74.

Best, C.T., MacRoberts, G.W., & Sithole, N.M. (1988). Examination of the perceptual re-organization for speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 245-260.

Best, C.T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, **29**, 191-211.

Bever, T.G., Lackner, J., & Kirk, R. (1969). The underlying structures of sentences are the primary units of immediate speech processing. *Perception and Psychophysics*, **5**, 225-231.

Bradley, D.C. (1980). Lexical representation of derivational relation. In Aronoff, M., & Kean, L. (Eds.), *Juncture*. Saratoga, CA: Anma Libri.

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

Bregman, A.S. (1978). The formation of auditory streams. In Requin, J. (Ed.), *Attention and Performance VII*. Hillsdale, NJ: Erlbaum.

Broadbent, D.E. (1965). Information processing in the nervous system. *Science*, **150**, 457-462.

Carr, P.B., & Trill, D. (1964). Long-term larynx-excitation spectra. *Journal of the Acoustical Society of America*, **36**, 2033-2040.

Carrell, T.D. (1984). Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification. *Doctoral dissertation, Indiana University*.

Chomsky, N., & Miller, G.A. (1963). Introduction to the formal analysis of natural language. In Luce, R.D., Bush, R., & Galanter, E. (Eds.), *Handbook of Mathematical Psychology, Volume II*. New York: John Wiley, 269-321.

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 and Performance*, **16**, 551-563.

Cohen, A., & Nooteboom, S.G. (Eds.) (1975). *Structure and Process in Speech Perception*. Heidelberg: Springer-Verlag.

Cole, R.A., & Jakimik, J. (1980). A model of speech perception. In Cole, R.A. (Ed.), *Perception and Production of Fluent Speech.* Hillsdale, NJ: Erlbau, 133-163.

Cole, R.A., & Rudnicky, A.I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874-1946. *Psychological Review*, **90**, 94-101.

Cole, R.A., & Scott, B. (1974a). The phantom in the phoneme: Invariant cues for stop consonants. *Perception and Psychophysics*, **15**, 101-107.

Cole, R.A., & Scott, B. (1974b). Toward a theory of speech perception. *Psychological Review*, **81**, 348-374.

Collier, R., & t'Hart, J. (1975). The role of intonation in speech perception. In Cohen, A., & Nooteboom, S.G. (Eds.), *Structure and Process in Speech Perception.* Heidelberg: Springer-Verlag.

Coltheart, M., Develaar, E., Jonasson, J.T., & Besner, D. (1976). Access to the internal lexicon. In Dornic, S. (Ed.), *Attention and Performance VI.* Hillsdale, NJ: Erlbaum.

Cooper, F.S., Liberman, A.M., & Borst, J.M. (1951). The interconversion of audible and visible patterns as a basis for research on the perception of speech. *Proceedings of the National Academy of Sciences*, **37**, 318-327.

Cooper, W.E. (1976). Syntactic control of timing in speech production: A study of complement clauses. *Journal of Phonetics*, **4**, 151-171.

Cooper, W.E., & Sorenson, J. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, **62**, 683-692.

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

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55-60.

Cutler, A. (1989). Auditory lexical access: Where do we start? In Marslen-Wilson, W.D. (Ed.), *Lexical Representation and Process.* Cambridge, MA: MIT Press, 342-356.

Cutler, A., & Darwin, C.J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, **29**, 217-224.

Cutler, A., & Fodor, J.A. (1979). Semantic focus and sentence comprehension. *Cognition*, **7**, 49-59.

Cutler, A., & Foss, D.J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, **20**, 1-10.

Cutting, J.E., & Pisoni, D.B. (1978). An information-processing approach to speech perception. In Kavanagh, J.F., & Strange, W. (Eds.), *Speech and Language in the Laboratory, School, and Clinic*. Cambridge, MA: MIT Press.

Cutting, J.E. (1978). There may be nothing peculiar to perceiving in a speech mode. In Requin, J. (Ed.), *Attention and Performance VII*. Hillsdale, NJ: Erlbaum, 229-244.

Cutting, J.E. (1987). Perception and information. *Annual Review of Psychology*, **38**, 61-90.

Darwin, C.J. (1975). On the dynamic use of prosody in speech perception. In Cohen, A., & Nooteboom, S.G. (Eds.), *Structure and Process in Speech Perception*. Heidelberg: Springer-Verlag.

Darwin, C.J. (1976). The perception of speech. In Carterette, E.C., & Friedman, M.P. (Eds.), *Handbook of Perception*. New York: Academic Press.

Day, R.S. (1968). Fusion in dichotic listening. *Doctoral dissertation, Stanford University*.

Delattre, P.C., Liberman, A.M., & Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**, 769-773.

Delattre, P.C., Liberman, A.M., Cooper, F.S., & Gerstman, L.H. (1952). An experimental study of the acoustic determinants of vowel color: Observations of one- and two-formant vowels synthesized from spectrographic patterns. *Word*, **8**, 195-210.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, **27**, 761-764.

Derr, M.A., & Massaro, D.W. (1980). The contribution of vowel duration, F0 contour, and frication duration as cues to the /juz/ - /jus/ distinction. *Perception and Psychophysics*, **27**, 51-59.

Diehl, R.L. (1986). Coproduction and direct perception of phonetic segments: a critique. *Journal of Phonetics*, **14**, 61-66.

Dobbs, A.R., Friedman, A., & Lloyd, J. (1985). Frequency effects in lexical decisions: A test of the verification model. *Journal of Experimental Psychology: Human Perception and Performance*, **11**, 81-92.

Easton, R.D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception and Psychophysics*, **32**, 562-570.

Eimas, P.D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-l] distinction by young infants. *Perception and Psychophysics*, **18**, 341-347.

Eimas, P.D., & Miller, J.L. (1980). Contextual effects in infant speech perception. *Science*, **209**, 1140-1141.

Eimas, P.D., Siqueland, E.R., Jusczyk, P.W., & Vigorito, J. (1971). Speech perception in infants. *Science*, **171**, 303-306.

Elman, J.L. (1989). Connectionist approaches to acoustic/phonetic processing. In Marslen-Wilson, W.D. (Ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press, 227-260.

Elman, J.L., & McClelland, J.L. (1984). An interactive activation model of speech perception. In Lass, N.J. (Ed.), *Language and Speech*. New York: Academic Press.

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

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

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

Fitch, H.L., Hawles, T., Erickson, D.M., & Liberman, A.M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception and Psychophysics*, **27**, 343-350.

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

Fodor, J.A. (1985). Precis of The modularity of mind. *The Behavioral and Brain Sciences*, **8**, 1-42.

Forster, K.I. (1976). Accessing the mental lexicon. In Wales, R.J., & Walker, E.C.T. (Eds.), *New Approaches to Language Mechanisms*. Amsterdam: North Holland, 257-287.

Forster, K.I. (1979). Levels of processing and the structure of the language processor. In Cooper, W.E., & Walker, E.C.T. (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale, NJ: Erlbaum, 27-86.

Forster, K.I. (1989). Basic issues in lexical processing. In Marslen-Wilson, W.D. (Ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press, 75-107.

Forster, K.I. (1990). Lexical processing. In Osherson, D.N., & Lasnik, H. (Eds.), *An Invitation to Cognitive Science (Volume 1)*. Cambridge, MA: MIT Press, 95-131.

Forster, K.I, & Bednall, E.S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, **4**, 53-61.

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

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

Fowler, C.A., & Rosenblum, L.D. (1990). Duplex perception: A comparison of monosyllables and slamming of doors. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 742-754.

Fujisaki, H., & Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual Report of the Engineering Research Institute, Volume 28*. Tokyo, Japan: University of Tokyo, 67-73.

Fujisaki, H., & Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute, Volume 29*. Tokyo, Japan: University of Tokyo, 207-214.

Fujisaki, H., & Kawashima, T. (1971). A model of the mechanisms for speech perception: Quantitative analysis of categorical effects in discrimination. *Annual Report of the Engineering Research Institute, Volume 30*. Tokyo, Japan: University of Tokyo, 59-68.

Gaitenby, J.H. (1965). The elastic word. *Haskins Laboratories Status Report on Speech Research, SR-2*. New Haven, CT: Haskins Laboratories, 3.1-3.12.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, **6**, 110-125.

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

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

Geiselman, R.E., & Bellezza, F.S. (1977). Incidental retention of speaker's voice. *Memory and 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.

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

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton-Mifflin.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifflin.

Goldinger, S.D., Luce, P.A., & Pisoni, D.B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501-518.

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

Goodman, J.C., & Huttenlocher, J. (1988). Do we know how people identify spoken words? *Journal of Memory and Language*, **27**, 684-698.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologica*, **9**, 317-323.

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

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, **38**, 299-310.

Grosjean, F., & Gee, J.P. (1987). Prosodic structure and spoken word recognition. *Cognition*, **25**, 135-155.

Grunke, M.E., & Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception and Psychophysics*, **31**, 210-218.

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1958). Effect of third formant transitions on the perception of the voiced stop consonants. *Journal of the Acoustical Society of America*, **30**, 122-126.

Hawles, T., & Jenkins, J.J. (1971). Problem of serial order in behavior is not resolved by context-sensitive associative memory models. *Psychological Review*, **78**, 122-129.

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

Hoffman, H.S. (1958). Study of some cues in the perception of voiced stop consonants. *Journal of the Acoustical Society of America*, **30**, 1035-1041.

Holmberg, T.L., Morgan, K.A., & Kuhl, P.K. (1977). Speech perception in early infancy: Discrimination of fricative consonants. *Journal of the Acoustical Society of America*, **62**, S76.

Huggins, A.W.F. (1972). On the perception of temporal phenomena in speech. In Requin, J. (Ed.), *Attention and Performance VII.* Hillsdale, NJ: Erlbaum, 279-297.

Isenberg, D., & Liberman, A.M. (1978). Speech and non-speech percepts from the same sound. *Journal of the Acoustical Society of America*, **64**, S20.

Jenkins, J.J. (1989). Is this the way to Camelot? *Contemporary Psychology*, **5**, 451-452.

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**, 1-136.

Jordan, M.I. (1986). Serial order: A parallel distributed processing approach. *Institute for Cognitive Science, Report 8604*. San Diego, CA: University of California.

Jusczyk, P.W. (1985). On characterizing the development of speech perception. In Mehler, J., & Fox, R. (Eds.), *Neonate Cognition: Beyond the Blooming, Buzzing Confusion*. Hillsdale, NJ: Erlbaum, 199-229.

Jusczyk, P.W. (1986). A review of speech perception research. In Kaufman, L., Thomas, J., & Boff, K. (Eds.), *Handbook of Perception and Performance*. New York: Wiley.

Jusczyk, P.W., Pisoni, D.B., & Mullennix, J.W. (1989). Effects of talker variability on speech perception by 2-month old infants. *Research on Speech Perception Progress Report No. 15*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Jusczyk, P.W., Pisoni, D.B., Reed, M.A., Fernald, A., & Myers, M. (1983). Infants' discrimination of the duration of rapid spectrum changes in nonspeech signals. *Science*, **222**, 175-177.

Jusczyk, P.W., Pisoni, D.B., Walley, A.C., & Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *Journal of the Acoustical Society of America*, **67**, 262-270.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, **72**, 379-389.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **73**, 322-335.

Kewley-Port, D., & Luce, P.A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & Psychophysics*, **35**, 353-360.

Klatt, D.H. (1974). The duration of [S] in English words. *Journal of Speech and Hearing Research*, **17**, 51-63.

Klatt, D.H. (1975). Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics*, **3**, 129-140.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**, 1208-1221.

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

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

Klatt, D.H., & Cooper, W.E. (1975). Perception of segment duration in sentence context. In Cohen, A., & Nooteboom, S.G. (Eds.), *Structure and Process in Speech Perception*. Heidelberg: Springer-Verlag.

Kluender, K.R., Diehl, R.L., & Killeen, P.R. (1987). Japanese quail can learn phonetic categories. *Science*, **237**, 1195-1197.

Kuhl, P.K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, **66**, 1668-1679.

Kuhl, P.K., & Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. *Science*, **218**, 1138-1141.

Kuhl, P.K., & Miller, J.D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, **190**, 69-72.

Kuhl, P.K., & Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, **63**, 905-917.

Kuhl, P.K., & Miller, J.D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception and Psychophysics*, **31**, 279-292.

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**, 98-104.

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.

Lea, W.A. (1973). An approach to syntactic recognition without phonemics. *IEEE Transactions on Audio and Electroacoustics*, **Au-21**, 249-258.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman, A.M. (1970a). The grammars of speech and language. *Cognitive Psychology*, **1**, 301-323.

Liberman, A.M. (1970b). Some characteristics of perception in the speech mode. In Hamburg, D.A. (Ed.), *Perception and Its Disorders: Proceedings of ARNMD*. Baltimore: Williams & Wilkins, 238-254.

Liberman, A.M. (1982). On finding that speech is special. *American Psychologist*, **37**, 148-167.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

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

Liberman, A.M., Harris, K.S., Hoffman, H.A., & Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.

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

Liberman, A.M., & Mattingly, I.G. (1989). A specialization for speech perception. *Science*, **243**, 489-494.

Liberman, A.M., Mattingly, I.G., & Turvey, M.T. (1972). Language codes and memory codes. In Melton, A.W., & Martin, E. (Eds.), *Coding Processes in Human Memory*. New York: Winston, 307-334.

Lieberman, P., Crelin, E.S., & Klatt, D.H. (1972). Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. *American Anthropology*, **74**, 287-307.

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

Lindblom, B.E.F., & Svensson, S.G. (1973). Interaction between segmental and non-segmental factors in speech recognition. *IEEE Transactions on Audio and Electroacoustics*, **Au-21**, 536-545.

Lisker, L., & Abramson, A.S. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

Lisker, L., & Abramson, A.S. (1967). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.

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

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., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio. *Journal of the Acoustical Society of America*, **78**, 1949-1957.

Luce, P.A., & Pisoni, D.B. (1987). Speech perception: New directions in research, theory, and applications. In Winitz, H. (Ed.), *Human Communication and Its Disorders*. Norwood, NJ: Ablex, 1-87.

Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In Altmann, G. (Ed.), *Cognitive Representation of Speech*. Cambridge, MA: MIT Press, 122-147.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics*, **24**, 253-257.

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

MacKain, K.S., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, **219**, 1347-1349.

Mann, V.A., Madden, J., Russell, J.M., & Liberman, A.M. (1981). Integration of time-varying cues and the effects of phonetic context. *Unpublished manuscript.* New Haven, CT: Haskins Laboratories.

Marcus, S.M. (1984). Recognizing speech: On mapping from sound to meaning. In Bouma, H., & Bowhuis, D.G. (Eds.), *Attention and Performance X: Control of Language Processes.* Hillsdale, NJ: Erlbaum, 151-164.

Marslen-Wilson, W.D. (1975). Sentence perception as an interactive parallel process. *Science*, **189**, 226-228.

Marslen-Wilson, W.D. (1980a). Optimal efficiency in human speech processing. *Unpublished manuscript.* Cambridge: Cambridge University.

Marslen-Wilson, W.D. (1980b). Speech understanding as a psychological process. In Simon, J.C. (Ed.), *Spoken Language Generation and Understanding*. Dordrecht, Holland: Reidel.

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

Marslen-Wilson, W.D. (1989). Access and integration: Projecting sound onto meaning. In Marslen-Wilson, W.D. (Ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press, 3-24.

Marslen-Wilson, W.D., & Tyler, L.K. (1975). Processing structure of sentence perception. *Nature*, **257**, 784-785.

Marslen-Wilson, W.D., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

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

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

Massaro, D.W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, **79**, 124-145.

Massaro, D.W. (1986). A new perspective and old problems. *Journal of Phonetics*, **14**, 69-74.

Massaro, D.W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D.W. (1989). Multiple book review of Speech perception by ear and eye: A paradigm for psychological inquiry. *The Behavioral and Brain Sciences*, **12**, 741-794.

Massaro, D.W., & Cohen, M.M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/ - /si/ distinction. *Journal of the Acoustical Society of America*, **60**, 704-717.

Massaro, D.W., & Cohen, M.M. (1977). The contribution of voice-onset time and fundamental frequency as cues to the /zi/ - /si/ distinction. *Perception and Psychophysics*, **22**, 373-382.

Massaro, D.W., & Cohen, M.M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753-771.

Massaro, D.W., & Cohen, M.M. (1990). Perception of synthesized audible and visible speech. *Psychological Science*, **1**, 55-63.

Massaro, D.W., & Oden, G.C. (1980). Speech perception: A framework for research and theory. In Lass, N.J. (Ed.), *Speech and Language: Advances in Basic Research and Practice, Volume III*. New York: Academic Press, 129-165.

Mattingly, I.G., & Liberman, A.M. (1988). Specialized perceiving systems for speech and other biologically-significant sounds. In Edelman, G., Gall, W., & Cohen, W. (Eds.), *Auditory Function: The Neurobiological Bases of Hearing*. New York: Wiley, 775-793.

McClelland, J.L. (1979). On the time-relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, **86**, 287-330.

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

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review*, **88**, 375-405.

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

Miller, G.A. (1962). Decision units in the perception of speech. *IRE Transactions on Information Theory*, **IT-8**, 81-83.

Miller, J.D., Wier, C.C., Pastore, R.E., Kelley, W.J., & Dooling, R.J. (1976). Discri-mination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, **60**, 410-417.

Miller, J.L. (1980). The effect of speaking rate on segmental distinctions: Acoustic variation and perceptual compensation. In Eimas, P.D., & Miller, J.L. (Eds.), *Perspectives on the Study of Speech*. Hillsdale, NJ: Erlbaum, 39-74.

Miller, J.L. (1987). Mandatory processing in speech perception. In Garfield, J.L. (Ed.), *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge, MA: MIT Press, 309-322.

Miller, J.L. (1990). Speech perception. In Osherson, D.N., & Lasnik, H. (Eds.), *An Invitation to Cognitive Science, Volume I*. Cambridge, MA: MIT Press, 69-93.

Miller, J.L., & Liberman, A.M. (1979). Some effects of later-occurring information on the perception of stop consonant and semi-vowel. *Perception and Psychophysics*, **25**, 457-465.

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

Monsen, R.B., & Engebretson, A.M. (1977). Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, **62**, 981-993.

Morse, P.A., & Snowdon, C.T. (1975). An investigation of categorical speech discrimination by rhesus monkeys. *Perception and Psychophysics*, **17**, 9-16.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, **76**, 165-178.

Morton, J. (1979). Word recognition. In Morton, J., & Marshall, J.D. (Eds.), *Psycholinguistics 2: Structures and Processes*. Cambridge, MA: MIT Press, 109-156.

Morton, J. (1982). Disintegrating the lexicon: An information processing approach. In Mehler, J., Walker, E.C.T., & Garrett, M. (Eds.), *On Mental Representation*. Hillsdale, NJ: Erlbaum.

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

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

Murdock, B.B., Jr. (1962). The serial position effect in free recall. *Journal of Experimental Psychology*, **64**, 482-488.

Nakatani, L.H., & Schaffer, J.A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, **63**, 234-245.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In Chase, W.G. (Ed.), *Visual Information Processing*. New York: Academic Press, 283-308.

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

Nosofsky, R.M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 700-708.

Nooteboom, S.G., Brokx, J.P.L., & de Rooij, J.J. (1978). Contributions of prosody to speech perception. In Levelt, W.J.M., & Flores d'Arcais, G.B. (Eds.), *Studies in the Perception of Language*. New York: Wiley, 75-107.

Nusbaum, H.C. (1984). Possible mechanisms of duplex perception: "Chirp" identification versus dichotic fusion. *Perception and Psychophysics*, **35**, 94-101.

Nusbaum, H.C., & Schwab, E.C. (1986). The role of attention and active processing in speech perception. In Schwab, E.C., & Nusbaum, H.C. (Eds.), *Perception of Speech and Visual Form: Theoretical Issues, Models, and Research*. New York: Academic Press, 113-157.

Nusbaum, H.C., Schwab, E.C., & Sawusch, J.R. (1983). The role of "chirp" identification in duplex perception. *Perception and Psychophysics*, **33**, 323-332.

Nygaard, L.C., & Eimas, P.D. (1990). A new version of duplex perception: Evidence for phonetic and nonphonetic fusion. *Journal of the Acoustical Society of America*, **88**, 75-86.

Oden, G.C., & Massaro, D.W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Oller, D.K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, **54**, 1235-1247.

Oller, D.K., Eilers, R.E., & Ozdamar, O. (1990). A psychoacoustic model of the ba/wa boundary shift. *Journal of the Acoustical Society of America*, **87**, S38.

Paap, K.R., Newsome, S.L., McDonald, J.E., & Schvaneveldt, R.W. (1982). An activation-verification model for letter and word recognition: The word-superiority effect. *Psychological Review*, **89**, 573-594.

Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception. In Eimas, P.D., & Miller, J.L. (Eds.), *Perspectives on the Study of Speech*. Hillsdale, NJ: Erlbaum.

Pastore, R.E., Schmeckler, M.A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception and Psychophysics*, **33**, 469-474.

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

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

Peterson, G.E., & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Peterson, L.J., & Peterson, M.J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**, 193-198.

Pisoni, D.B. (1971). On the nature of categorical perception of speech sounds. *Supplement to Haskins Laboratories Status Report on Speech Research, SR-27.* New Haven, CT: Haskins Laboratories, 1-101.

Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253-260.

Pisoni, D.B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition*, **3**, 7-18.

Pisoni, D.B. (1977). Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D.B. (1991). Modes of processing speech and nonspeech signals. To appear in Mattingly, I.G., & Studdert-Kennedy, M. (Eds.), *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Erlbaum.

Pisoni, D.B. (1978). Speech perception. In Estes, W.K. (Ed.), *Handbook of Learning and Cognitive Processes, Volume 6.* Hillsdale, NJ: Erlbaum, 167-233.

Pisoni, D.B., Aslin, R.N., Perey, A.J., & Hennessy, B.L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 297-314.

Pisoni, D.B., Carrell, T.D., & Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, **34**, 314-322.

Pisoni, D.B., Logan, J.S., & Lively, S.E. (in press). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In Nusbaum, H.C., & Goodman, J.C. (Eds.), *Development of Speech Perception: The Transition from Recognizing Speech Sounds to Spoken Words*. Cambridge, MA: MIT Press.

Pisoni, D.B., & Luce, P.A. (1986). Speech perception: Research, theory, and the principle issues. In Schwab, E.C., & Nusbaum, H.C. (Eds.), *Perception of Speech and Visual Form: Theoretical Issues, Models, and Research*. New York: Academic Press, 1-50.

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

Pisoni, D.B., & Sawusch, J.R. (1975). Some stages of processing in speech perception. In Cohen, A., & Nooteboom, S.G. (Eds.), *Structure and Process in Speech Perception.* Heidelberg: Springer-Verlag, 16-34.

Port, R.F. (1977). The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochu words. Bloomington, Indiana: Indiana University Linguistics Club.

Porter, R.J., Jr. (1986). Speech messages, modulations, and motions. *Journal of Phonetics*, **14**, 83-88.

Pruitt, J.S., Strange, W., Polka, L., & Aguilar, M.C. (1990). Effects of category knowledge and syllable truncation during auditory training on Americans' discrimination of Hindi retroflex-dental contrasts. *Journal of the Acoustical Society of America*, **87**, S72.

Rand, T.C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

Remez, R.E. (1986) Realism, language, and another barrier. *Journal of Phonetics*, **14**, 89-97.

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

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**, 81-110.

Repp, B.H. (1983a). Categorical perception: Issues, methods, findings. In Lass, N.J. (Ed.), *Speech and Language: Advances in Basic Research and Practice, Volume 10.* New York: Academic Press.

Repp, B.H. (1983b). Trading relations among acoustic cues in speech perception: Speech-specific but not special. *Haskins Laboratories Status Report on Speech Research, SR-76.* New Haven, CT: Haskins Laboratories, 129-132.

Repp, B.H. (1984). Against a role of "chirp" identification in duplex perception. *Perception and Psychophysics*, **35**, 89-93.

Roberts, M., & Summerfield, Q. (1981). Audio-visual adaptation in speech perception. *Perception and Psychophysics*, **30**, 309-314.

Rumelhart, D.E., & McClelland, J.L. (1982). An interactive activation model of context effects in letter perception: Part II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, **89**, 60-94.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, **89**, 43-50.

Samuel, A.G. (1986). The role of the lexicon in speech perception. In Schwab, E.C., & Nusbaum, H.C. (Eds.), *Perception of Speech and Visual Form: Theoretical Issues, Models, and Research.* New York: Academic Press, 89-111.

Samuel, A.G., & Ressler, W.H. (1986). Attention within auditory word perception: Insights from the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance*, **12**, 70-79.

Savin, H.B., & Bever, T.G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, **9**, 295-302.

Schouten, M.E.H. (1980). The case against a speech mode of perception. *Acta Psychologica*, **44**, 71-98.

Schwab, E.C. (1981). Auditory and phonetic processing for tone analogs of speech. *Doctoral dissertation, State University of New York at Buffalo.*

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing. In Bouma, H., & Bouwhis, D.G. (Eds.), *Attention and Performance X.* Hillsdale, NJ: Erlbaum, 165-181.

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

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

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.

Stevens, K.N., & Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic features. In Eimas, P.D., & Miller, J.L. (Eds.), *Perspectives on the Study of Speech.* Hillsdale, NJ: Erlbaum, 1-38.

Strange, W. (1972). The effects of training on the perception of synthetic speech sounds: Voice onset time. *Doctoral dissertation, University of Minnesota.*

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

Strange, W., & Jenkins, J.J. (1978). The role of linguistic experience in the perception of speech. In Pick, H.L., Jr., & Walk, R.D. (Eds.), *Perception and Experience.* New York: Plenum.

Strange, W., Verbrugge, R.R., Shankweiler, D.P., & Edman, T.R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, **60**, 213-221.

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

Studdert-Kennedy, M. (1976). Speech perception. In Lass, N.J. (Ed.), *Contemporary Issues in Experimental Phonetics.* New York: Academic Press, 243-293.

Studdert-Kennedy, M. (1980). Speech perception. *Language and Speech*, **23**, 45-66.

Studdert-Kennedy, M. (1982). On the dissociation of auditory and phonetic perception. In Carlson, R., & Granstrom, B. (Eds.), *The Representation of Speech in the Peripheral Auditory System.* Amsterdam: Elsevier, 3-10.

Studdert-Kennedy, M. (1983). Perceiving phonetic events. *Haskins Laboratories Status Report on Speech Research, SR-74/75*. New Haven, CT: Haskins Laboratories, 53-69.

Studdert-Kennedy, M. (1986). Two cheers for direct realism. *Journal of Phonetics*, **14**, 99-104.

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., & Cooper, F.S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, **77**, 234-249.

Studdert-Kennedy, M., & Shankweiler, D.P. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, **48**, 579-594.

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

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, **36**, 314-331.

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

Svensson, S.G. (1974). Prosody and grammar in speech perception. *Monographs from the Institute of Linguistics, No. 2*. Stockholm, Sweden: Institute of Linguistics, University of Stockholm.

Taft, M., & Hambly, G. (1986). Exploring the Cohort Model of word recognition. *Cognition*, **22**, 259-282.

Tanenhaus, M.K., & Lucas, M.M. (1987). Context effects in lexical processing. *Cognition*, **25**, 213-234.

Tomiak, G.R., Mullennix, J.W., & Sawusch, J.R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America*, **81**, 755-764.

Townsend, J.T. (1989). Winning "20 questions" with mathematical models. *The Behavioral and Brain Sciences*, **12**, 775-776.

Tyler, L.K. (1984). The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics*, **36**, 417-427.

Tyler, L.K., & Frauenfelder, U.H. (1987). The process of spoken word recognition: An introduction. *Cognition*, **25**, 1-20.

Tyler, L.K., & Marslen-Wilson, W.D. (1982). Speech comprehension processes. In Mehler, J., Walker, E.C.T., & Garrett, M. (Eds.), *Perspectives on Mental Representation: Experimental and Theoretical Studies of Cognitive Processes and Capacities*. Hillsdale, NJ: Erlbaum.

Verbrugge, R.R., & Rakerd, B. (1986). Evidence of talker-independent information for vowels. *Language and Speech*, **29**, 39-57.

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.

Vinegrad, M.D. (1972). A direct magnitude scaling method to investigate categorical vs. continuous modes of speech perception. *Language and Speech*, **15**, 114-121.

Walley, A.C., Pisoni, D.B., & Aslin, R.N. (1981). The role of early experience in the development of speech perception. In Aslin, R.N., Alberts, J., & Peterson, M.J. (Eds.), *The Development of Perception: Psychobiological Perspectives*. New York: Academic Press, 219-255.

Warren, P., & Marslen-Wilson, W.D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics*, **41**, 262-275.

Warren, R.M. (1989). The use of mathematical models in perceptual theory. *The Behavioral and Brain Sciences*, **12**, 776.

Waters, R.S., & Wilson, W.A., Jr., (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. *Perception and Psychophysics*, **19**, 285-289.

Whalen, D., & Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, **237**, 169-171.

Wickelgren, W.A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, **76**, 1-15.

Wickelgren, W.A. (1976). Phonetic coding and serial order. In Carterette, E.C., & Friedman, M.P. (Eds.), *Handbook of Perception, Volume 7*. New York: Academic Press.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, **8**, 338-353.

Automaticity in Speech Perception: Some Speech/Nonspeech Comparisons[1]

Keith A. Johnson[2] and James V. Ralston

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

[1] We appreciate the comments of our colleagues Steve Goldinger, John Logan, John Mullennix, David Pisoni, Rich Shiffrin, and Van Summers. Denise Beike helped us conduct the experiments. This research was supported, in part, by NIH Training Grant #DC-00012-12 to Indiana University.

[2] Now at Department of Linguistics, UCLA, Los Angeles, CA 90024.

Abstract

We report the results of three experiments that explored some differences between speech and nonspeech sounds. Our hypothesis is that the perception of speech sounds is automatized, while the perception of less familiar sounds is not. The experiments compared patterns of behavior in categorization and discrimination tasks for listeners reporting either speech or nonspeech percepts. The first experiment replicated the findings of Best et al. (1989) using a synthetic /wa/-/ja/ sinewave analog continuum; listeners who heard the stimuli as speech could consistently categorize the continuum, while nonspeech listeners could not. In the second experiment, both speech and nonspeech listeners could consistently categorize the stimuli (an /aI/-/aU/ sinewave analog continuum), yet there remained differences in identification functions, reaction times and context effects. In the third experiment, nonspeech listeners' discrimination sensitivity was greater than speech listeners, particularly within response categories. The observed pattern of results suggests a degree of automaticity in the categorization processes of speech listeners that is not present in nonspeech listeners. We argue that this automaticity focuses speech listeners' attention on a categorical level of representation, effectively "deafening" them to the auditory properties of sounds. In terms of the "speech is special" debate, then, these results suggest that the perceptual processing of phonemes is distinct from the perceptual processing of other nonspeech events primarily by virtue of listeners' extensive experience in categorizing speech sounds.

Automaticity in Speech Perception: Some Speech/Nonspeech Comparisons

Repetition tends to produce automaticity in many types of human behavior. Of particular interest for the studies reported here is the finding that repetition results in automatic processes of categorization and perceptual attention (LaBerge, 1981; Neumann, 1984; Shiffrin, 1987). These findings are of importance for an understanding of speech perception because speech is a pervasive human activity. It would indeed be remarkable if the speech perception abilities of humans were not affected by such extensive use. Of course, exposure to a language is necessary in order to learn the words of that language, and there is also evidence which suggests that many fine-grained details of speech perception are affected by language experience (Flege, 1988). So, at the level of establishing the categories used in speech perception, language experience plays an obvious role. However, the possibility that the extensive use of speech over the course of the lifetime of the individual results in the automatic perceptual processing of speech has not been adequately investigated. The importance of this topic lies in the relationship between automaticity and some recent claims concerning speech perception. Research on automatic processing has revealed several phenomena which relate to the proposed specialness of speech perception (Liberman & Mattingly, 1985; Whalen & Liberman, 1987; Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989).

Automatic processes have been described in a variety of contexts, but the prototypical example of automatic processing is the visual recognition of letters. Shiffrin & Schneider (1977) found that automatic categorization could be produced with extensive training in a visual search task. Subjects were required to detect any of a subset of letter targets in arrays of similar distractor letters. After thousands of training trials, subjects responded rapidly and showed little influence of the number of distractor letters in a display. This finding, among the many examples which could be cited, demonstrates that automatic categorization can be built through experience.

There is other evidence from the literature on visual perception that language experience may be sufficient for the formation of automatic categorization. Flowers, Polansky and Kerl (1981) had subjects identify random strings of letters in a visual array of letters. They found that the presence of real words inhibited the subjects' performance in the identification task. To explain this effect, Flowers et al. proposed that the words attracted subjects' attention automatically (i.e., words "popped out" of the display) and hence interrupted the search for the target string. Note the similarity between this phenomenon in the visual domain and the "preemptiveness" of speech which has been discovered in the auditory domain (Whalen & Liberman, 1987). However, in the case of visual preemptiveness, one cannot appeal to the operation of a biologically specified processing mechanism, as has been claimed for the auditory phenomenon, in particular, speech perception.

Our hypothesis, then, is that some of the phenomena which have recently been interpreted as evidence for the specialness of speech processing may, in fact, have an interpretation in terms of automaticity and concomitant changes in selective attention. Before turning to the experimental tests of this hypothesis we will first review some previous research which has given rise to the hypotheses (1) that speech perception "preempts" nonspeech perception (Whalen & Liberman, 1987) and (2) that speech perception is guided by "phonetic coherence" (Best et al., 1989).

Duplex Perception

Duplex perception, as an empirical phenomenon and an experimental technique, is historically related to the studies reported by Liberman, Harris, Kinney and Lane (1961) and Mattingly, Liberman, Syrdal and Halwes (1971). These earlier speech/nonspeech comparisons found rather large differences in perception between speech and nonspeech control sounds (as did studies reported by Fujisaki &

Sekimoto, 1975 and Cutting, 1974). However, the comparisons between speech and nonspeech perception were performed using different stimuli. Thus, observed differences that were attributed to a perceptual mode could logically have been attributed to stimulus differences. This methodological weakness is apparently solved in duplex perception.

Rand (1974) investigated release from backward masking in speech by presenting different formants of the same syllable dichotically. He found that when F2 was presented to one ear and F1 and F3 were simultaneously presented to the other ear, listeners were able to use the F2 as if it were integrated with F1 and F3 while at the same time the F2 gave rise to a percept similar to the nonspeech noise (i.e., chirp) that was heard when F2 was presented in isolation (see Mattingly et al., 1971). This type of stimulus presentation has been widely used as a nonspeech control condition for studying the differences in perception between speech and nonspeech sounds (Liberman, Isenberg, & Rakerd, 1981; Mann & Liberman, 1983; Nusbaum, Schwab, & Sawusch, 1983; Bentin & Mann, 1990; Nygaard & Eimas, 1990). An F2 and/or F3 transition of a CV syllable (i.e., the "chirp") is presented to one ear and the remaining formant transitions and vowel steady-states (i.e., the "base") are simultaneously presented to the other ear. This manipulation results in two percepts, a full syllable in one ear and a nonspeech chirp in the other ear. The information in the chirp is perceptually integrated with the base (see Nusbaum, 1984 and Repp, 1984 for a debate about this) and serves as the crucial acoustic information distinguishing different perceptual interpretations of the base, while simultaneously being identified as a separate auditory object with its own perceptual properties. The fact that chirps are perceived differently in the speech and nonspeech percepts in duplex perception has been taken as evidence for a specialized speech mode of perception (Liberman et al., 1981; Mann & Liberman, 1983). And, the fact that chirp perception in duplex perception is different from the perception of isolated chirps has been taken as evidence for the "preemptiveness" of speech perception (Whalen & Liberman, 1987). Preemptiveness suggests that phonetic perception occurs prior to, and inhibits ("preempts") auditory processing. In contrast to this view, we consider duplex perception to be the result of a more general auditory phenomenon, auditory stream segregation.

Presenting a formant transition to one ear while presenting the rest of the syllable to the other ear is one way to make the formant transition stand out as a separate auditory object. This can be accomplished in other ways as well; by manipulating the onset asynchrony of the chirp and the base (Bentin & Mann, 1990; Nygaard & Eimas, 1990), the amplitude of the chirp relative to the base (Whalen & Liberman, 1987), and the relative F0 of the chirp (Darwin, 1981; Gardner, Gaskill & Darwin, 1989). What these manipulations have in common is that they are all parameters influencing auditory stream segregation (Bregman, 1978, 1987, 1990). In the typical duplex perception stimulus (Liberman et al., 1981), there are three stimulus parameters which suggest that the chirp and the base are a single auditory object. These are the formant onset synchrony, plausible amplitude relationships of the formants, and the common F0 of the two portions of the stimulus. Similarly, there are two stimulus parameters which suggest that the chirp and the base are different auditory objects. These are the different ear of presentation and the different formant offset times. With these conflicting cues, it is not surprising that the separate formant transition (i.e., the chirp) in a duplex stimulus would be perceived as a separate perceptual object, while simultaneously being integrated with the dichotically presented phonetic base. Gardner et al. (1989) have demonstrated that, as a formant is made acoustically distinct from a simultaneously presented base form by increasing the acoustic disparity between the formant and the base, it dissociates from the base in two stages. First, it separates from the base (i.e., becomes identifiable as a separate auditory object). Then, as the disparity between the components increases, the phonetic percept disintegrates (i.e., the formant no longer participates in the perceptual identity of the base).

Gardner et al. (1989) investigated these phenomena using four-formant stimuli. When F2 was perceptually integrated with the other formants, the syllable was perceived as /ru/, and when it was not perceptually integrated with the other formants, the syllable was perceived as /li/. By changing the F0 of the second formant in several steps Gardner et al. (1989) found that the F2 perceptually separated from the base at a smaller F0 disparity than was necessary to change the phonetic percept from /ru/ to /li/. Thus, they found a range of F0 differences which resulted in duplex perception. This distinction between perceptual separation and phonetic disintegration is related to Deutsch and Roll's (1976) distinction between "what" and "where" decision mechanisms in processing dichotic stimuli and suggests that separate processes of perceptual grouping and perceptual identification can operate on binaurally presented stimuli (i.e., that localization is a particular instance of an auditory grouping process).

This view of duplex perception correctly predicts that it is possible to create duplex percepts with nonspeech stimuli (Pastore, Harris, & Kaplan, 1982). So, we conclude that duplex perception is the result of general processes of auditory object perception, rather than the product of specialized mechanisms. The fact that perception in these experiments has been duplex rather than monoplex or triplex has to do with the cues for the separability of the auditory objects involved. The fact that perceptual integration occurs has to do with the independence of processes of perceptual grouping from processes of perceptual identification. We should also mention the potential role of primary versus schema-driven processes of stream segregation (Bregman, 1990). The perceptual integration of chirp and base in duplex perception seems to be the product of a schema-driven grouping process, and the segregation of the chirp in duplex perception seems to be the product of a primary stream segregation process.

As we mentioned earlier, the interest in duplex perception as it relates to theories of speech perception has been in the fact that the same signal gives rise to both a speech percept and a nonspeech percept. Thus, duplex perception appears to offer an ideal control condition for studying the differences in perception between speech and nonspeech sounds. Where Mattingly et al. (1971) used isolated F2 transitions (i.e., chirps) or isolated F2 transitions and steady-states (i.e., bleats) as nonspeech controls, Liberman et al. (1981) were able to present exactly the same stimuli for both speech discrimination, and nonspeech discrimination. So, the criticism that the nonspeech control stimuli used by Mattingly et al. (1971) were different from their speech stimuli can be addressed by presenting the same duplex stimuli for speech and nonspeech perceptual judgements.

If, however, duplex perception is the result of the independence of processes of auditory grouping and perceptual identification, it is apparent that although the same signal is presented in duplex perception studies, subjects do not base their responses on the same auditory information. The speech and nonspeech percepts in duplex perception are based on different auditory streams, two streams containing different auditory information. Thus, to find that silent intervals in the speech stream affect the speech percept but not the chirp percept (Liberman et al., 1981), or that a preceding /l/ or /r/ in the speech stream does not affect the chirp percept while it does affect the speech percept (Mann & Liberman, 1983) is irrelevant in a study of the differences in perception between speech and nonspeech sounds, because the chirp stream does not have a silent interval or sonorant in it. At the level of auditory objects, the speech and nonspeech signals contain non-identical information. Thus, we would claim that duplex perception is not an adequate nonspeech control condition.

In summary, we have argued that duplex perception is the result of general principles of auditory grouping, that chirps are integrated in the percept of the base because of a functional independence of processes of auditory grouping and processes of perceptual identification, and that duplex perception is not an adequate control condition for the study of the differences in perception of speech and nonspeech events because the two percepts are based on different auditory objects.

Sinewave Replicas of Speech

It is possible to produce replicas of natural speech sounds by combining sinewaves which obey the regularities of frequency range, amplitude and trajectories of vocal tract resonances. These speech replicas do not contain the acoustic correlates of vocal cord vibration (i.e., fundamental frequency and glottal spectrum), nor do they have the formant bandwidths typical of natural speech, and so are, in these ways, not speech-like. Yet, the preservation of speech-like formant center frequency and amplitude relations coupled with plausible formant trajectories, renders the signals speech-like. The listener's perceptual experience of sinewave replicas of speech reflects the ambiguity of the acoustic signal. For most listeners, these signals are not immediately recognizable as speech, but if listeners are biased to expect to hear speech, many will be able to correctly transcribe the intended utterance (Remez, Rubin, Pisoni, & Carrell, 1981). In addition, many listeners, when required to categorize sinewave replicas of speech, will spontaneously hear them in terms of speech categories (Bailey, Summerfield, & Dorman, 1977), although the signal's unnatural quality remains apparent (Remez et al., 1981). Thus, sinewave replicas of speech provide a control condition for studying differences between perceiving an acoustic signal as speech and perceiving that same signal as a nonspeech auditory event.

Beginning with the studies reported by Cutting & Rosner (1974), there have been a number of studies using simple analogs of speech sounds which have shown that a number of effects observed in speech perception may also be found in the perception of nonspeech auditory objects (e.g., Hillenbrand, 1984; Miller, Wier, Pastore, Kelly, & Dooling, 1976; Parker, Diehl, & Kluender, 1986; Pastore et al., 1982; Pisoni, 1977; Pisoni, Carrell, & Gans, 1983; Pastore, Ahroon, Puleo, Crimmins, Golowner, & Berger, 1976; Pastore, Ahroon, Wolz, Puleo, & Berger, 1975; and Ralston & Sawusch, 1984). On the basis of similarities between speech and nonspeech perception, these studies have concluded that many aspects of speech perception may be the result of general properties of auditory perception rather than a special processing mechanism. However, the studies have been limited in that they can only show qualitative similarities; they have not compared the perception of the same signal in a speech mode versus a nonspeech mode.

The first study to make a comparison between listeners reporting speech and nonspeech percepts was reported by Bailey et al. (1977). In their first experiment, Bailey et al. asked subjects to categorize sinewave replicas of two /b/-/d/ continua in an AXB categorization task. Subjects first categorized the stimuli in a nonspeech mode. Then they were biased to hear the same stimuli as speech by performing the same task with full-formant versions of the same continuum. Finally, they repeated the categorization task with the same stimuli. There was a small difference in the way that the sinewave stimuli were categorized in the two conditions. When the stimuli were heard as speech, the boundary between /b/ and /d/ was at a different point along the continuum than when the stimuli were heard as nonspeech. The boundary in the speech condition was similar to the location of the boundary found for the full-formant versions of the continua. In their second experiment, Bailey et al. sorted the subjects by their written descriptions of a continuum of sinewave replicas from /ba/ to /da/. For speech listeners, the boundary on the continuum in an AXB categorization task was located toward the /da/ endpoint of the continuum as it was for a full formant version of the continuum; while for nonspeech listeners the boundary was located near the middle of the continuum. The identification functions for nonspeech listeners were also somewhat flatter than they were for speech listeners, indicating a less sharp distinction between the categories.

Bailey et al.'s (1977) study is interesting both because of the methodology involved and because of the observed results. First, the authors were able to compare listeners' labeling behavior for the same stimuli (and the same auditory objects) in a speech mode and in a nonspeech mode of perception. Second, they found both similarities and differences between the two modes of perception. If they had only tested subjects in a nonspeech mode they might have claimed that they had found categorical perception of a nonspeech continuum and concluded that speech perception is not different from nonspeech perception. However, the fact that they had data for both speech and nonspeech perception of the same stimuli gave them the opportunity to compare small differences in the listeners' performance under the two modes of perception as well as the basic similarities in response patterns.

There have been several recent studies comparing different modes of perception while listening to sinewave replicas of speech. These studies have found that speech listeners use trading relations among acoustic cues for phonetic categories while nonspeech listeners do not (Best et al., 1981); that speech listeners are able to use acoustic information which seems to be masked for nonspeech listeners (Grunke & Pisoni, 1982; Schwab, 1981); and that speech listeners seem to integrate information over CV-syllable stimuli while nonspeech listeners do not (Tomiak, Mullennix, & Sawusch, 1987). Although most authors have attributed these differences in perception between speech and nonspeech perception to the experience that listeners have with speech, recently Best et al. (1989) have suggested that, in order to categorize speech sounds, listeners must discover the "correct basis for categorization, namely, the articulatory structures that the patterns specify". We will suggest a re-interpretation of this "phonetic coherence" account in the discussion of Experiment 1, below.

Assumptions Underlying Present Experiments

The experiments reported here were designed to investigate the differences in perception between perceiving auditory objects as speech sounds and perceiving those same auditory objects as nonspeech sounds. The design of the experiments has been guided by several assumptions. One of these assumptions is that some of the most interesting differences between speech perception and nonspeech perception are to be found in categorization behavior. It has been demonstrated that speech listeners can perceive speech in terms of fine acoustic differences which do not determine category membership (Samuel, 1977; also see Repp, 1987 for a review), and that even without special training or instructions, listeners are sensitive to small within-category differences when making discrimination judgements (Pisoni & Tash, 1974). But these studies shed light neither on how listeners categorize the sounds of speech nor on whether categorizing speech sounds is accomplished in ways different from categorizing nonspeech sounds. Therefore, rather than attempting to show that speech can be treated in a purely psychophysical way, the experiments reported here were designed to explore differences in the ways speech and nonspeech listeners categorize complex sounds.

We have made two assumptions about the behavioral characteristics of automatic processes. First, we assume that automatic processes are generally faster than nonautomatic processes. For instance, Shiffrin and Schneider (1977) found that subjects could locate visual targets in a search paradigm more quickly in consistent-mapping conditions, in which subjects apparently utilized automatic processing, than in variable-mapping conditions, in which subjects apparently did not utilize automatic processing. Shiffrin (1987) warns against relying exclusively on reaction time data in characterizing a process as automatic. However, in the present experiments, reaction time is an appropriate index of automaticity because our speech and nonspeech listeners were performing identical tasks with identical stimuli. Second, we assume that automatic categorization causes attention to be switched from sensory representations to categorical representations, which results in a functional loss of sensory information. Note that since this is our translation of Whalen & Liberman's (1987) "preemptiveness", any evidence that speech perception involves automatic categorization may also be taken as evidence for the preemptiveness of speech.

However, we prefer the more general automaticity explanation due to its parsimonious description. The hypothesis that attention is drawn to the categorical level in automatic categorization predicts several differences in perception between speech and nonspeech listeners in our experiments beyond the prediction that speech listeners will have shorter reaction times. Assuming that contrast effects arise from auditory processing, we also expect that categorization judgments of nonspeech listeners will be influenced more by context than speech listeners, even if the nonspeech listeners are generally as successful as speech listeners in categorizing the stimuli. Finally, our hypothesis predicts that nonspeech listeners will be better able to detect small acoustic differences among sinewave replicas of speech in a discrimination task.

The experiments were carried out as follows: Experiment 1 replicates the design reported by Best et al. (1989) with a new stimulus continuum. Experiment 2 demonstrates that the predicted differences in perception between speech and nonspeech listeners occur even when nonspeech listeners can successfully label the endpoints of a sinewave replica continuum. Finally, Experiment 3 tests the hypothesis that nonspeech listeners have greater access to sensory information in a discrimination task.

EXPERIMENT 1

One of the most striking differences between speech and nonspeech perception has to do with the perception of CV syllables (Grunke & Pisoni, 1982; Schwab, 1981; Best et al., 1989). Initial consonant transitions seem to be unavailable to nonspeech listeners. When listeners hear sinewave replicas of stop-vowel syllables as nonspeech they are virtually unable to use the initial F2 or F3 transition in a categorization task. For instance, Grunke & Pisoni (1982) presented sinewave replicas of /ba/ and /da/. They found that listeners who were biased to hear the stimuli as speech were able to consistently classify the stimuli, while listeners who were biased to hear the stimuli as nonspeech sounds responded randomly. Schwab (1981) and Best et al. (1989) also found that nonspeech listeners were virtually unable to classify stimuli which differed in terms of their initial F2 or F3 transitions. We attempted to replicate this finding with a /wa/-/ja/ continuum. The stimuli in this continuum have longer transitions than the stop consonants employed by Grunke and Pisoni (1982) and are lower in frequency than the F3 transitions used by Best et al. (1989) and so it might be expected that the nonspeech listeners would be better able to categorize them.

Method

Subjects

Seventeen undergraduate students (7 female, 10 male) at Indiana University, Bloomington, participated in the experiment for partial course credit in introductory psychology. None of the listeners reported a history of speech or hearing disorders at the time of testing.

Materials

An eleven step continuum of sinewave replicas of speech was synthesized. This continuum ranged from /wa/ to /ja/. The onset frequency of the sinewave analog of F2 varied from 900 to 1900 Hz. The changes in F2 onset frequency were calculated as equal intervals in Bark units and then converted into Hertz using the formula published by Schroeder, Atal and Hall (1979). After a transition of 75 ms, the frequency of the F2 analog was 1250 Hz during the 175 ms steady-state vowel. The F2 transition was the only property of the stimuli which varied across the continuum. The F1 analog also had a 75 ms transition followed by a 175 ms steady-state at 850 Hz. The F3 analog was steady-state throughout at 2850 Hz. The parameters for the endpoint stimuli of the continuum are displayed in Table 1.

Insert Table 1 about here

Procedure

Listening sessions were conducted online using a PDP 11/34 computer at the Speech Research Laboratory at Indiana University. Subjects were run in groups of six or fewer.

Because we were interested in exploring differences in categorization behavior for speech and nonspeech, we used a standard identification paradigm in which subjects were asked to label stimuli presented in isolation. During a training phase, only the endpoint stimuli were presented, one per trial. Subjects were told only that they would hear two types of sounds and that they were to learn which button they should press after each sound. After subjects had made a response on a trial, a feedback light was turned on above the button that they should have pressed. Training was continued until all subjects in a group had reached or exceeded 90% correct on a block of 20 trials, which usually only required about 20 to 40 trials total. Next, subjects were asked to write down their subjective impression of the stimuli and their classification criteria. During the subsequent generalization phase of the experiment, all stimuli from the series were presented and the feedback lights continued to mark the correct response for presentations of the endpoint stimuli. The stimuli were ordered in such that each stimulus followed every other stimulus an equal number of times. Finally, subjects completed a written questionnaire that asked for their phenomenological impression of the stimuli and on what basis they classified the stimuli.

Results

The data were sorted into three categories based on the subjects' descriptions of the stimuli before and after the generalization phase. Subjects that reported speech percepts before and after the generalization test were classified as "speech" listeners (n=5). Those that reported nonspeech percepts before and after the generalization test were classified as "nonspeech" listeners (n=7). Finally, subjects whose descriptions were different before and after the generalization were classified as "mixed" listeners (n=5). Because it was not possible to know the precise basis for their decisions during particular trials, these data were excluded from further analyses.

Figure 1 displays the probability that the left response button was pressed after each stimulus was presented. The data for each subject were fit with a cumulative probability function using the method of least squares. The slopes and category boundaries (50% crossover point) obtained from the fitted functions were then submitted to separate analyses of variance with percept treated as a between-subjects variable. The analyses confirmed that nonspeech listeners could not reliably categorize the stimuli while speech listeners could. The slopes of the labeling functions were less steep for nonspeech listeners (-.02) compared to the speech listeners (-.24), a difference that was highly significant [$F(1,10)=65.05$, $p<.001$]. Only four of the seven nonspeech functions had category boundaries that fell within the stimulus range. The difference between speech (6.35) and the four nonspeech (6.33) category boundaries that fell within the continuum was not statistically significant [$F(1,7)<0.01$, $p=.99$]. Therefore, speech listeners identified the stimuli very consistently while the nonspeech listeners were quite inconsistent.

Insert Figure 1 about here

Figure 2 displays the reaction times of labeling responses. Overall, reaction times were slower for nonspeech listeners (785 ms) than for speech listeners (698 ms), but the difference was not significant [$F(1,11)=1.64$, $p=.23$]. Although the reaction time function for nonspeech listeners was relatively flat, the function for speech listeners exhibited a peak near the location of the labeling category boundary.

Table 1.

Frequency and intensity parameters for Stimulus 1 (/wa/ analog) and Stimulus 11 (/ja/ analog) in Experiment 1.

Stimulus 1 (/wa/ analog)	Time (ms)		
	0	130	250
F1 Freq (Hz)	371	723	797
Amp (dB)	60	60	60
F2 Freq (Hz)	**750**	1334	1370
Amp (dB)	55	55	55
F3 Freq (Hz)	2722	2722	2722
Amp (dB)	0	30	30

Stimulus 11 (/ja/ analog)	Time (ms)		
	0	130	250
F1 Freq (Hz)	371	723	797
Amp (dB)	60	60	60
F2 Freq (Hz)	**1850**	1334	1370
Amp (dB)	55	55	55
F3 Freq (Hz)	2722	2722	2722
Amp (dB)	0	30	30

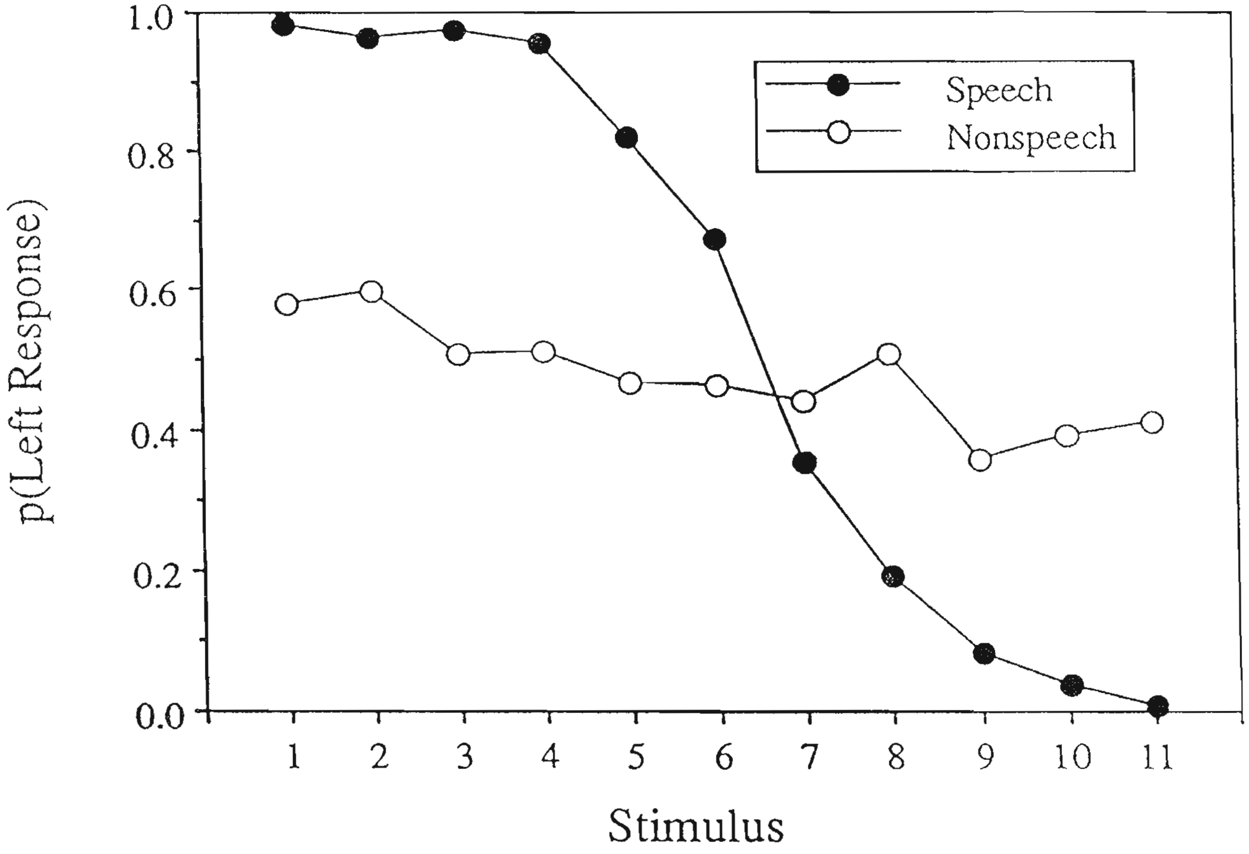


Figure 1. Labeling functions obtained in Experiment 1. Solid dots and connecting lines represent data for subjects reporting speech percepts; open dots and connecting lines represent data for subjects reporting nonspeech percepts.

This difference in the shape of the functions was manifest as a significant interaction between stimulus number and percept [$F(10,110)=2.76$, $p=.004$]. The peaked reaction time data for speech listeners replicates the findings of Pisoni and Tash (1974), who used synthetic speech stimuli differing in VOT.

Insert Figure 2 about here

Figure 3 displays labeling responses as a function of the preceding stimulus. Because context stimuli are known to exert their greatest effects on ambiguous stimuli at the boundary between perceptual categories, context data were calculated only for "focal" Stimuli 4 through 8. Data for individual subjects was modeled with a best fitting linear equation using the standard regression techniques. Positive slopes indicate contrastive effects of context; negative slopes indicate assimilative effects. There was a slight contrast effect of context for nonspeech listeners (.004) and a slight assimilative effect for speech listeners (-.004). However, the difference between these groups was not significant [$F(1,10)=1.49$, $p=.25$].

Insert Figure 3 about here

Discussion

Best et al. (1989) hypothesized that the difference between speech and nonspeech listeners in experiments of this sort indicates that speech listeners have access to the gestures used to produce CV syllables; that is, that speech listeners have discovered the phonetic coherence in the signal. We think that a more parsimonious explanation of the phenomenon can be found in terms of selective attention.

Pisoni, Logan and Lively (1991) have recently argued that selective attention to particular stimulus properties may explain some effects of language experience in speech perception. They argue that language learning involves learning to selectively attend to properties in the speech signal which prove to be useful in maintaining linguistically important categorical distinctions. The difference between speech and nonspeech listeners' identification functions seen in this experiment is reminiscent of the difference between native speakers of English and native speakers of Japanese in the perception of an /r/-/l/ continuum (MacKain, Best, & Strange, 1981). It is exactly this sort of data which prompted Pisoni et al.'s comments on selective attention and speech perception. Yet, this example was also mentioned by Best et al. (1989). They suggested that nonspeech listeners and nonnative speakers in studies such as the one reported by MacKain et al. have not discovered the phonetic coherence in the signal. Earlier we proposed a translation of Whalen & Liberman's (1987) preemptiveness in terms of automaticity. Best et al.'s explanation of the difference between speech and nonspeech listeners and the difference between native and nonnative listeners seems also to have a translation into the more general terms of selective attention. The speech listeners have learned which aspects of the signal are linguistically important and, through selective attention, they have mapped these dimensions onto phonetic categories. Note that an important feature of this translation of Best et al.'s (1989) explanation is that it makes no crucial reference to articulation or perception of phonetic gestures.

Our hypothesis is that speech sound categories are overlearned and thus are automatically categorized and attract automatic attention. However, in the first experiment nonspeech listeners performance was so poor that some interesting and potentially important differences between the two groups may have been obscured. If the automaticity hypothesis is correct we would expect to find predictable differences between speech and nonspeech listeners even when nonspeech listeners can label the endpoints of the continuum as accurately as speech listeners can. We tested this hypothesis in a second experiment.

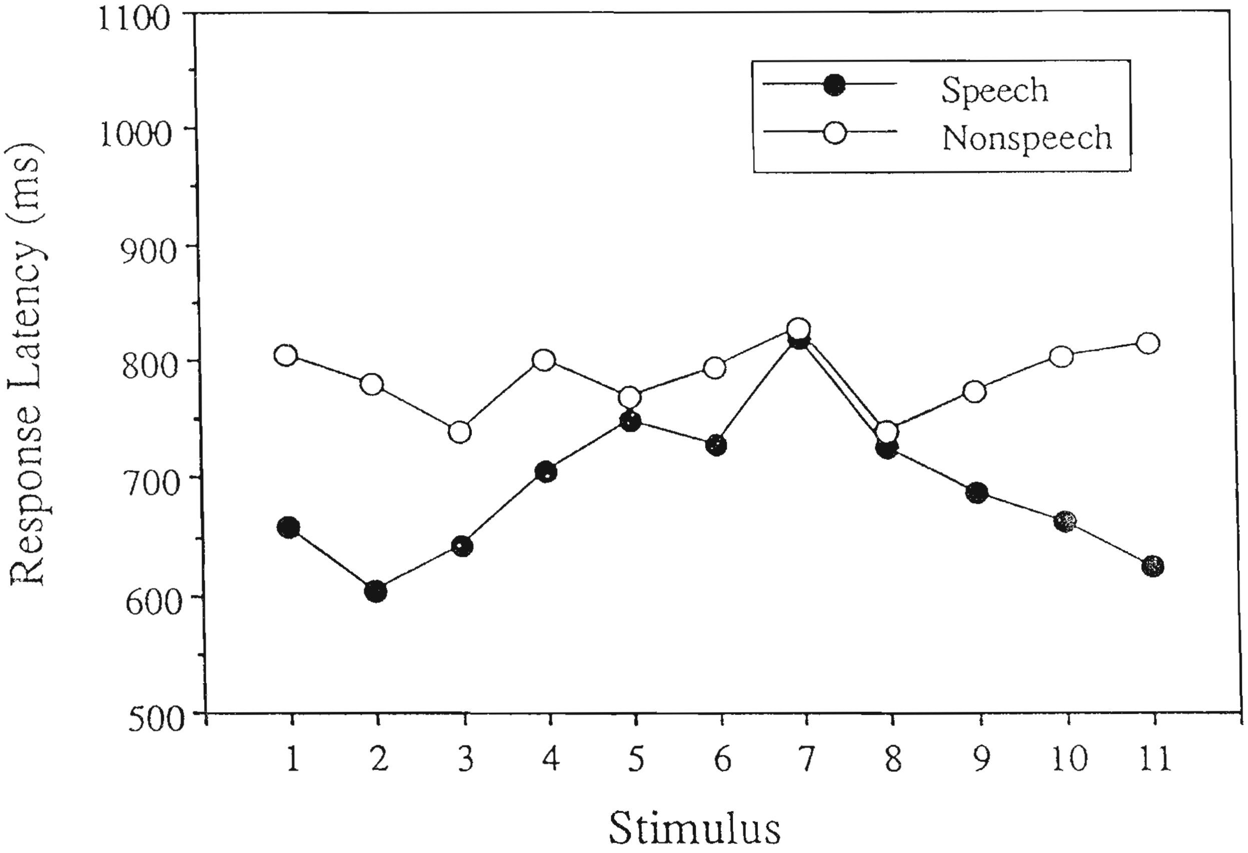


Figure 2. Reaction times for labeling responses in Experiment 1. Solid dots and connecting lines represent data for subjects reporting speech percepts; open dots and connecting lines represent data for subjects reporting nonspeech percepts.

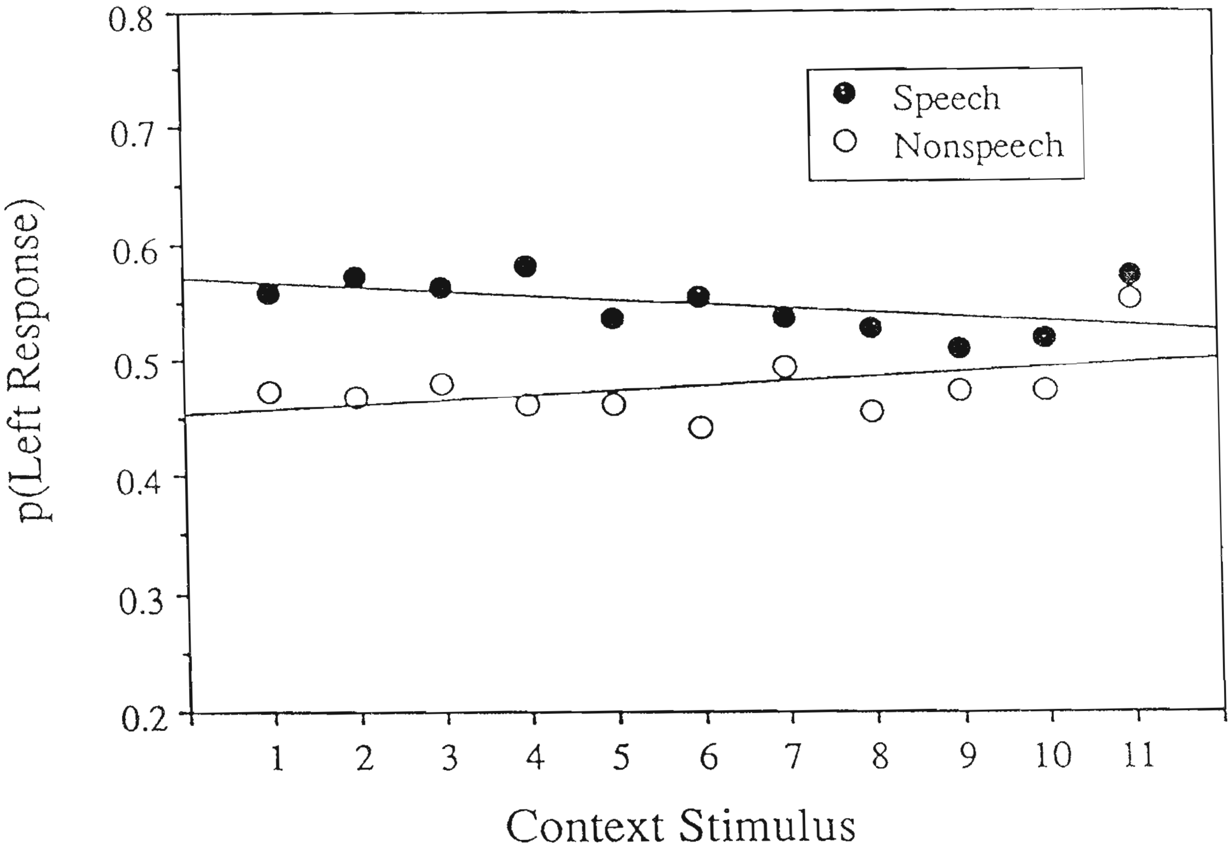


Figure 3. Context effects for labeling responses in Experiment 1. Each point represents the averaged probability that the left response button was pressed after a focal stimulus as a function of the identity of the preceding stimulus. Each set of points has corresponding best-fitting line determined by linear regression.

EXPERIMENT 2

Although there were large differences between speech and nonspeech data in Experiment 1, we suspected that a significant portion of this difference might be due to the inability of several nonspeech subjects to selectively attend to the appropriate stimulus dimensions. Therefore, in Experiment 2 we sought to make stimulus differences more salient and consequently enhance nonspeech performance. Based on the results of Grunke and Pisoni (1982) and Schwab (1981) we expected that nonspeech listeners would be better able to categorize stimuli which are roughly the mirror images of the stimuli used in Experiment 1. That is, by merely reversing the stimuli in time, placing the F2 transition at the end of the stimulus rather than at the beginning, we expected to find that nonspeech listeners would be better able to categorize the stimuli. Therefore, we predicted a closer correspondence between speech and nonspeech labeling probabilities. However, we expected that other interesting differences between groups would remain.

First, based on the presumed automaticity of speech perception, we predicted that speech listeners would categorize the stimuli more quickly than the nonspeech listeners. By itself this difference cannot be taken as proof of automaticity, but it is consistent with the hypothesis that speech perception is an automatic process.

In addition, we predicted that nonspeech listeners would rely more on auditory memory in performing the categorization task than the speech listeners, and thus, their categorization responses would be more affected by context. Note that this prediction is not based on an assumption that auditory memory is somehow less robust for speech listeners. Rather, the prediction is based on the assumption that speech listeners rapidly categorize the stimulus, limiting the effects of auditory contrast. In contrast, the nonspeech listeners' slow categorization is more influenced by auditory contrast.

Method

Subjects

Forty-six undergraduate students (22 female, 24 male) at Indiana University, Bloomington, participated in the experiment for partial course credit in introductory psychology. None of the listeners reported a history of speech or hearing disorders at the time of testing and none were used in the first experiment.

Materials

An eleven step continuum of sinewave replicas of speech which ranged from /aI/ to /aU/ was synthesized. The stimuli were 250 ms in duration and had a steady-state portion (130 ms) and a transition portion (120 ms). The only change across the continuum was in the F2 transition at the end of the stimulus (rising for /aI/ and falling for /aU/). As with the materials for Experiment 1, the F2 offset frequencies were calculated as equal intervals in Bark units and then converted to Hertz. Table 2 displays frequency and amplitude values for the endpoint stimuli of the continuum.

Insert Table 2 about here

Procedure

All aspects of the procedure were identical to Experiment 1.

Table 2

Frequency and intensity parameters for Stimulus 1 (/aI/ analog) and Stimulus 11 (/aU/ analog) in Experiments 2 and 3.

Stimulus 1 (/aI/ analog)	Time (ms)		
	0	130	250
F1 Freq (Hz)	741	741	485
Amp (dB)	60	60	45
F2 Freq (Hz)	1257	1257	**1889**
Amp (dB)	50	50	50
F3 Freq (Hz)	2565	2565	2565
Amp (dB)	43	43	43

Stimulus 11 (/aU/ analog)	Time (ms)		
	0	130	250
F1 Freq (Hz)	741	741	485
Amp (dB)	60	60	45
F2 Freq (Hz)	1257	1257	**778**
Amp (dB)	50	50	50
F3 Freq (Hz)	2565	2565	2565
Amp (dB)	43	43	43

Results

Subjects' data was sorted into three groups according to the same criteria employed in Experiment 1. Fifteen subjects were classified as speech listeners, seven as nonspeech listeners, and 24 as mixed listeners.

Figure 4 displays the probability that the left button was pressed after each stimulus. Labeling data for each subject were again fit with sigmoid functions, and the resulting slope and crossover parameters were entered into separate analyses of variance, treating percept as a between-subjects variable. Averaged speech (5.05) and nonspeech (5.17) crossover points were not significantly different [$F(1,20)=0.15$, $p=.70$]. Although there was a trend for the slope of the labeling functions to be greater for speech (-.31) than nonspeech (-.23), the difference was not significant [$F(1,20)=2.22$, $p=.15$]. Therefore, on the surface, labeling performance appears to be nearly equivalent for the two groups. Both groups were able to partition the stimulus series into two relatively discrete perceptual categories with only a small region of ambiguity.

Insert Figure 4 about here

That nonspeech performance improved dramatically across the two experiments replicates the earlier findings of Grunke and Pisoni (1982) and Schwab (1981). In both of these studies, nonspeech listeners performed consistently poorer when classifying stimuli with final transitions as compared to initial transitions. In the same studies, however, there was no observable effect of the location of transitional information on listeners who perceived the sounds as speech. Our results support the interpretation that nonspeech listeners suffer the effects of backward masking from steady-state information when categorizing stimuli with initial transitions. We suspect that speech listeners' immunity from the same effects is a result of their rapid, automatic classification of the sounds into stable perceptual categories.

Figure 5 shows reaction time data for speech and nonspeech listeners. Nonspeech listeners were slower to identify the stimuli than speech listeners [$F(1,20)=10.09$, $p<.01$]. Based on the observed similarities between the two groups in terms of labeling probabilities, the reaction time difference does not appear to be due solely to the inability of the nonspeech listeners to identify the endpoints of the stimulus continuum. Reaction times were significantly different for the different stimuli of the series [$F(10,200)=22.34$, $p<.01$], reflecting the peakedness of the reaction time functions. The increase in reaction time around the boundary between two categories, presumed to reflect stimulus ambiguity, has been observed before (Pisoni & Tash, 1974). This effect of stimulus ambiguity was present for both the speech and nonspeech listeners, although there was an interaction between percept and token [$F(10,200)=2.41$, $p<.01$]. This interaction appears to reflect the broader peak in the nonspeech listeners' reaction time function, which suggests that more of the stimuli in the middle of the series were ambiguous for the nonspeech listeners.

Insert Figure 5 about here

Figure 6 shows the average influence of context for the speech and nonspeech listeners. As in Experiment 1, context data were calculated only for "focal" Stimuli 4 through 8. Context data was tabulated from each listener's responses and simple regression lines fit to the individual data. The slopes from these regression lines were entered into an analysis of variance that treated percept as a between-subjects variable. The average slope for speech listeners (-0.00007) was less than for nonspeech listeners

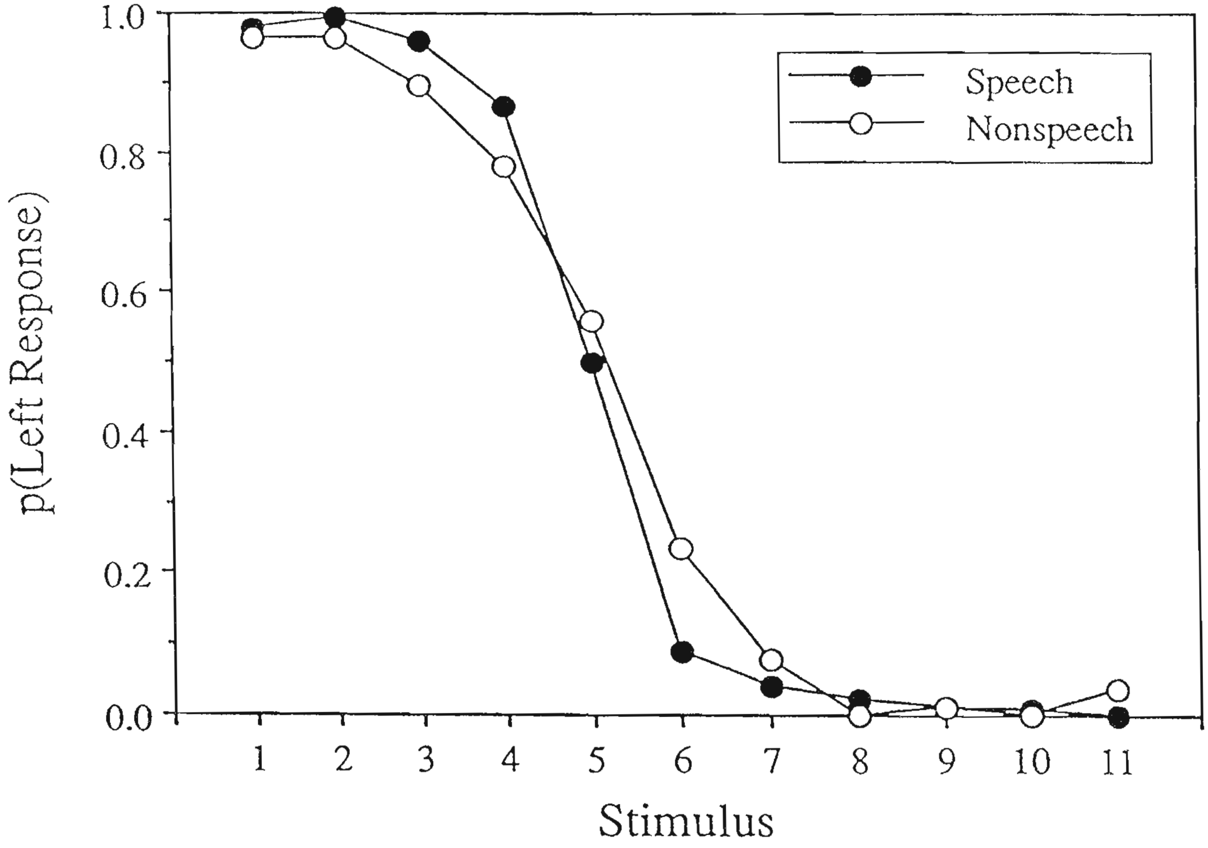


Figure 4. Labeling functions obtained in Experiment 2. Solid dots and connecting lines represent data for subjects reporting speech percepts; open dots and connecting lines represent data for subjects reporting nonspeech percepts.

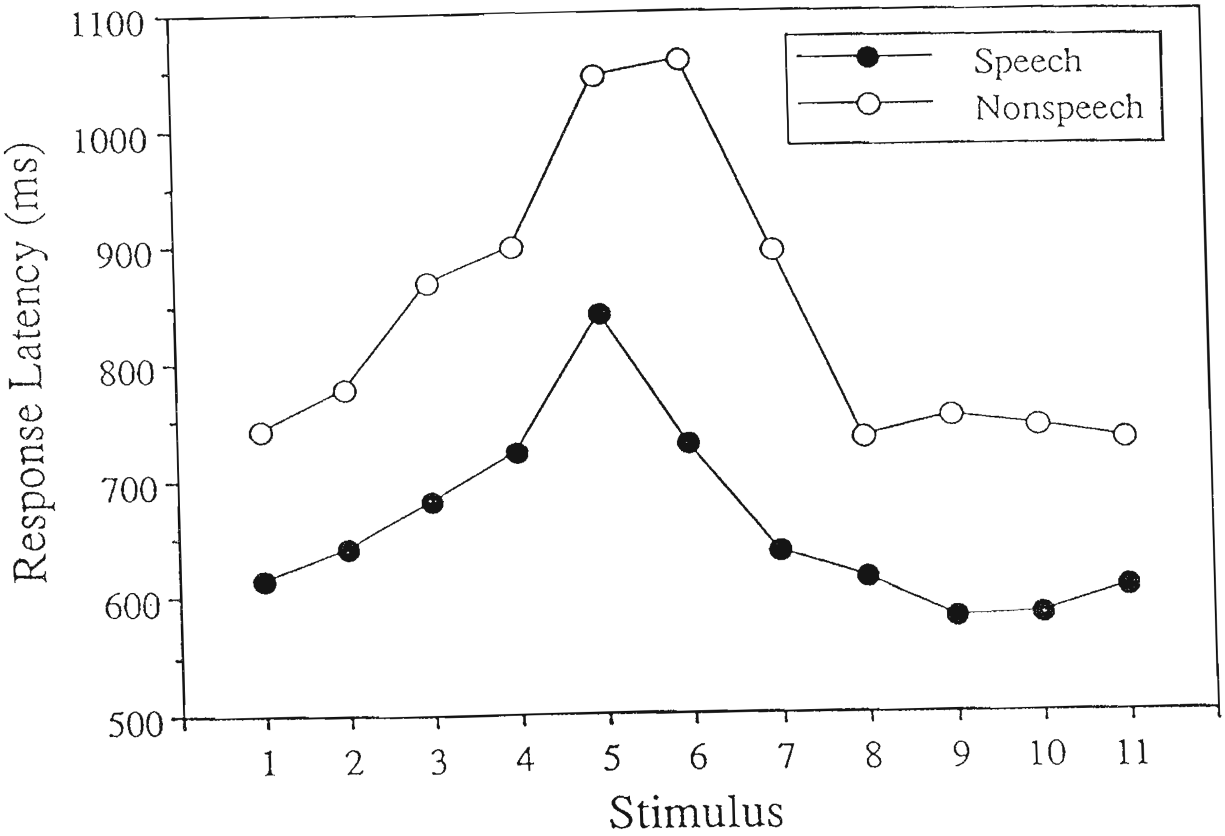


Figure 5. Reaction times for labeling responses in Experiment 2. Solid dots and connecting lines represent data for subjects reporting speech percepts; open dots and connecting lines represent data for subjects reporting nonspeech percepts.

(0.0224), a significant difference [$F(1,20)=8.52$, $p<.01$]. Thus, while speech listeners were not influenced by context, nonspeech listeners exhibited a contrast effect of preceding stimuli. This result is similar to that previously observed for musical stimuli when comparing trained musicians and musically naive listeners (Siegel & Siegel, 1977a).

Insert Figure 6 about here

Discussion

Taken together with the earlier findings, the results of Experiment 2 lend further support to the hypothesis that categorization of speech listeners is more automatic than for nonspeech listeners. As in Experiment 1, the nonspeech listeners in Experiment 2 identified the stimuli more slowly than the speech listeners. Given the almost random categorization performance of the nonspeech listeners in Experiment 1, one could have attributed the reaction time differences in that study to the nonspeech listeners' uncertainty about the judgements they had been asked to make. However, in this experiment, the two groups of listeners were very similar in their labeling probabilities. Thus, a stronger case can be made from the results of Experiment 2 that the reaction time difference reflects a real difference in the speed of perceptual processing.

Consistent with these reaction time differences are other effects of stimulus structure and preceding stimuli. In particular, when comparing across Experiments 1 and 2, nonspeech performance improved while speech performance did not. This effect of stimulus structure has been observed before by other investigators (Grunke & Pisoni, 1982; Schwab, 1981). We believe that the difference in the apparent backward masking effects may also be due to the relative speed of categorization. That is, speech listeners may classify transitional information so rapidly that the deleterious effects of following steady state information is minimized. Likewise, nonspeech listeners in Experiment 2 exhibited contrast effects of preceding stimuli while nonspeech listeners did not. Again, this difference may be related to the speed of categorization. If one assumes that contrast operates on auditory information and that it increases over time, then rapid categorization of focal stimuli should minimize contrastive effects of preceding stimuli.

Therefore, the differences between speech and nonspeech listeners observed in the first two experiments can be accounted for by recourse to the concept of automaticity. As we suggested earlier, one result of an increase in automaticity is that speech listeners suffer a functional loss of auditory information as compared to nonspeech listeners. In an identification task, this phenomenon will tend to produce reduced context effects. This is arguably beneficial in a categorization task and in normal conversation, but a functional loss of auditory information should prove detrimental in a discrimination task where there is a premium on discerning small acoustic differences. This prediction was tested in a third experiment.

EXPERIMENT 3

Method

Subjects

Sixty-six undergraduate students (31 female, 35 male) at Indiana University, Bloomington, participated in the experiment for partial course credit in introductory psychology. None of the listeners

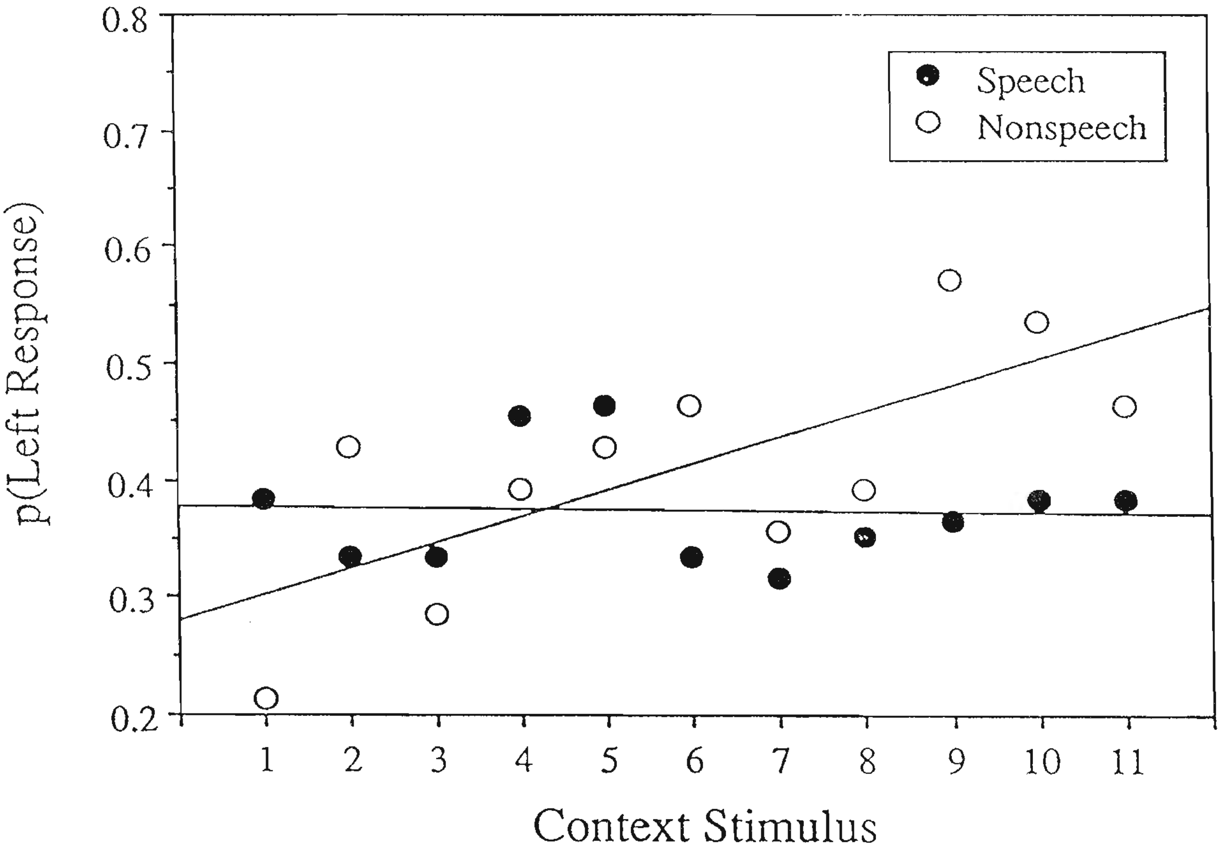


Figure 6. Context effects for labeling responses in Experiment 2. Each point represents the averaged probability that the left response button was pressed after a focal stimulus as a function of the identity of the preceding stimulus. Solid dots represent data for subjects reporting speech percepts; open dots represent data for subjects reporting nonspeech percepts. Each set of points has corresponding best-fitting line determined by linear regression.

reported any history of speech or hearing disorders at the time of testing and none participated in the previous experiments.

Materials

The eleven-step /aI/-/aU/ continuum used in Experiment 2 was again used in this experiment.

Procedure

Each experimental session was composed of the following parts. First, subjects identified the endpoint tokens in a training phase with visual feedback as in the first two experiments. Following this, subjects wrote their descriptions of the stimuli and their classification criteria. Then, the endpoint stimuli were presented with feedback for identification in a paired comparison, or AX, format. On each trial of this familiarization phase, tokens were presented with an interstimulus interval of 500 ms. After the presentation of the pair of stimuli, subjects were required to identify the first and second tokens using the button labels which they had learned in the training phase. During a subsequent identification test, all possible two-step pairs, as well as pairs of stimuli in which the stimuli were the same, were presented in random order for identification responses. Next, the same procedure was repeated, but subjects judged whether the stimuli in each pair were the same or different. This discrimination test was also preceded by a familiarization trial, again with feedback, using just the endpoint stimuli. In both the identification task and the discrimination task, each of the nine "different" pairs was presented 8 times, and each of the 11 "same" pairs was presented four times.

Results

Subjects' data were sorted into three groups based on the written descriptions of the stimuli and their classification criteria. There were 25 speech listeners, 17 nonspeech listeners and 24 mixed listeners. The data of the mixed group were not included in subsequent data analyses. Additionally, the data of seven remaining subjects (five nonspeech, two speech) were excluded from further data analyses because they performed poorly (<90% correct) in either the training or familiarization phases in this experiment. Thus, data from 23 speech listeners and 12 nonspeech listeners were used in the final analysis.

Figure 7 shows the identification data for speech and nonspeech listeners. Although the slopes of the functions are less steep than for the corresponding conditions in Experiment 2, their overall form is similar to the functions derived in that experiment. Although there was a trend for the crossover point to be higher on the stimulus series for the speech group (5.05) than for the nonspeech group (4.78), the difference was not significant [$F(1,34)=2.15$, $p=.15$]. The difference in slope between speech (-.28) and nonspeech listeners was significant [$F(1,34)=8.27$, $p<.01$]. The overall decrease in labeling performance in Experiment 3 as compared to Experiment 2 is obviously related to the format of the stimulus presentation. In Experiment 3, where stimuli were presented in pairs, there was a greater chance for context effects to influence category assignment (Healy & Repp, 1982). In addition, the memory demands of the two-interval identification task are likely higher than those associated with an analogous single-interval task (Pisoni, 1973). We suspect that these two influences interacted with perceptual set, degrading the performance of nonspeech listeners more than the speech listeners. However, the averaged data indicate that subjects in both groups were able to reliably sort the stimulus continuum into two distinct categories.

Insert Figure 7 about here

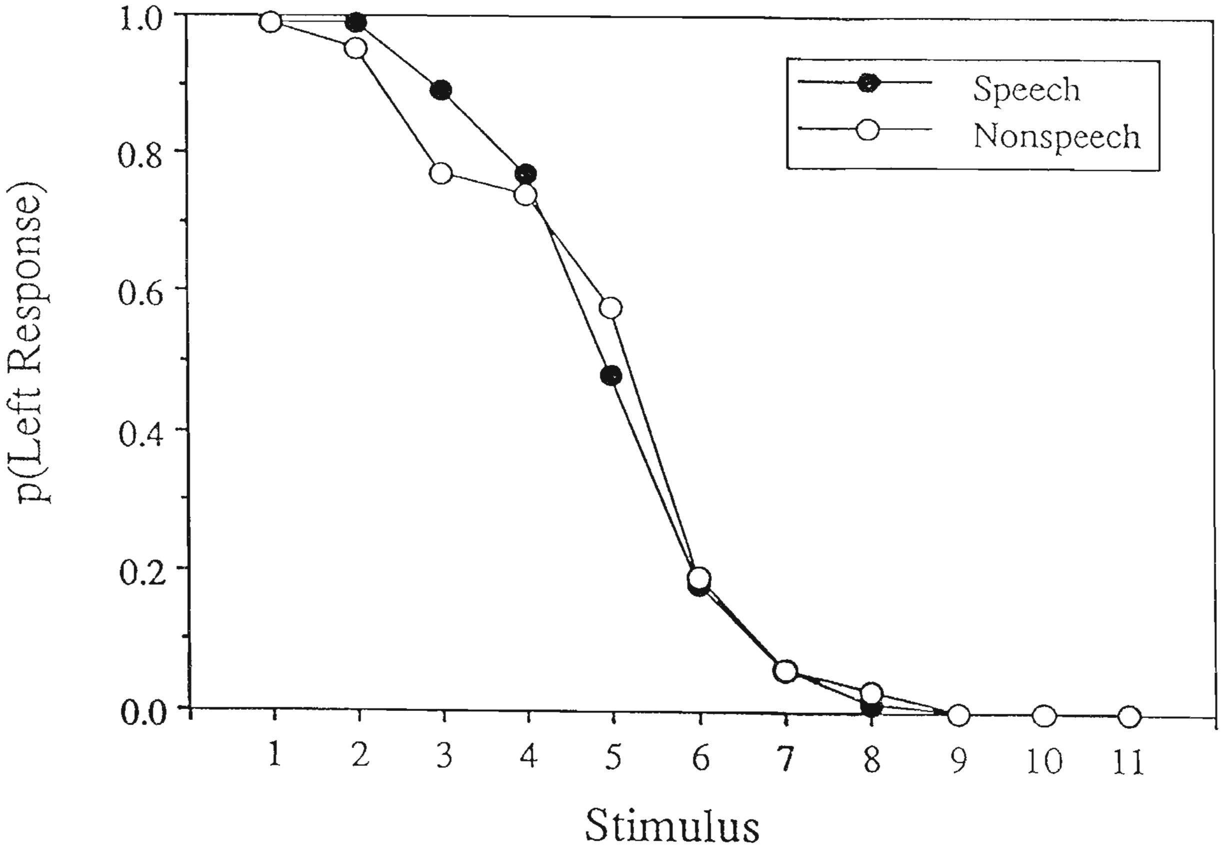


Figure 7. Labeling functions obtained in Experiment 3. Solid dots and connecting lines represent data for subjects reporting speech percepts; open dots and connecting lines represent data for subjects reporting nonspeech percepts.

Figure 8 shows average predicted and obtained d' values collapsed across stimuli for speech and nonspeech listeners. Predicted scores were derived from the labeling data. The main result of this experiment was that nonspeech listeners were more sensitive to stimulus differences than were the speech listeners. An analysis of variance was performed on these data, treating percept as a between-subjects factor and task as a within-subjects variable. The average d' value for speech listeners (2.37) was greater than for nonspeech listeners (2.78), a significant difference [$F(1,33)=4.23$, $p<.05$]. Observed discrimination sensitivity (2.79) was greater than predicted sensitivity (2.23), also a significant difference [$F(1,33)=17.41$, $p<.01$]. This is a very common finding in studies of speech perception using a variety of paradigms (Liberman et al., 1961; Healy & Repp, 1982), and it suggests that listeners make use of auditory information during discrimination. The difference between predicted and observed sensitivity was nearly twice as large for nonspeech listeners (.80) as for speech listeners (.43), suggesting that speech listeners were making greater use of category labels during the discrimination task (Healy & Repp, 1982). However, the interaction between percept and task was not significant [$F(1,33)=1.6$, $p=.21$].

Insert Figure 8 about here

Figure 9 shows A' values for each stimulus pair for both groups of subjects. A' is a nonparametric analog of d' which ranges from 0 to 1. The data in this figure come from the discrimination task. A' functions for each subject were entered into an analysis of variance treating percept as a between-subjects variable and stimulus pair as a within-subjects variable. There was a main effect for stimulus pair [$F(8,264)=23.13$, $p<.001$]. Pairs close to the category boundary were more discriminable than pairs drawn from within categories. This effect is consistent with the results of previous studies and may be due in part to discontinuities in the auditory code for stimuli across the series (Ralston & Sawusch, 1984). Also, there was a main effect for percept [$F(1,33)=6.25$, $p=.018$], as one might expect from the d' analysis. Although the interaction between percept and stimulus token was not statistically significant [$F(8,264)=0.87$, $p=.54$], there was a trend for the differences between speech and nonspeech listeners to be greatest within the labeling categories.

Insert Figure 9 about here

Discussion

The data obtained in Experiment 3 confirm our prediction that speech listeners suffer a functional "deafness" to auditory information. The d' analysis revealed a large main effect of percept. The trend toward an interaction between percept and task also suggests that speech listeners weigh phonetic labels more heavily in discrimination tasks than do nonspeech listeners. The more detailed A' analysis was consistent with the d' analysis. There was a large main effect of percept and this effect appeared somewhat larger within categories. Taken together, these results are consistent with the hypothesis that speech listeners attend automatically to a categorical level of representation and attend less to the subcategorical auditory properties of sinewave stimuli.

Our results appear to conflict with previous discrimination studies utilizing sinewave stimuli that found either no reliable difference between speech and nonspeech listeners (Bailey et al., 1977) or better performance for speech listeners (Best, Morrongiella, & Robson, 1981; Best et al., 1989). This difference in outcome probably reflects some combination of at least three procedural differences: number of observation intervals to be remembered, size of the ISI, and nature of training regimen. First, earlier studies employed paradigms with three observation intervals (AXB or oddity discrimination) which place relatively high demands on memory, whereas we employed a procedure with only two intervals, which

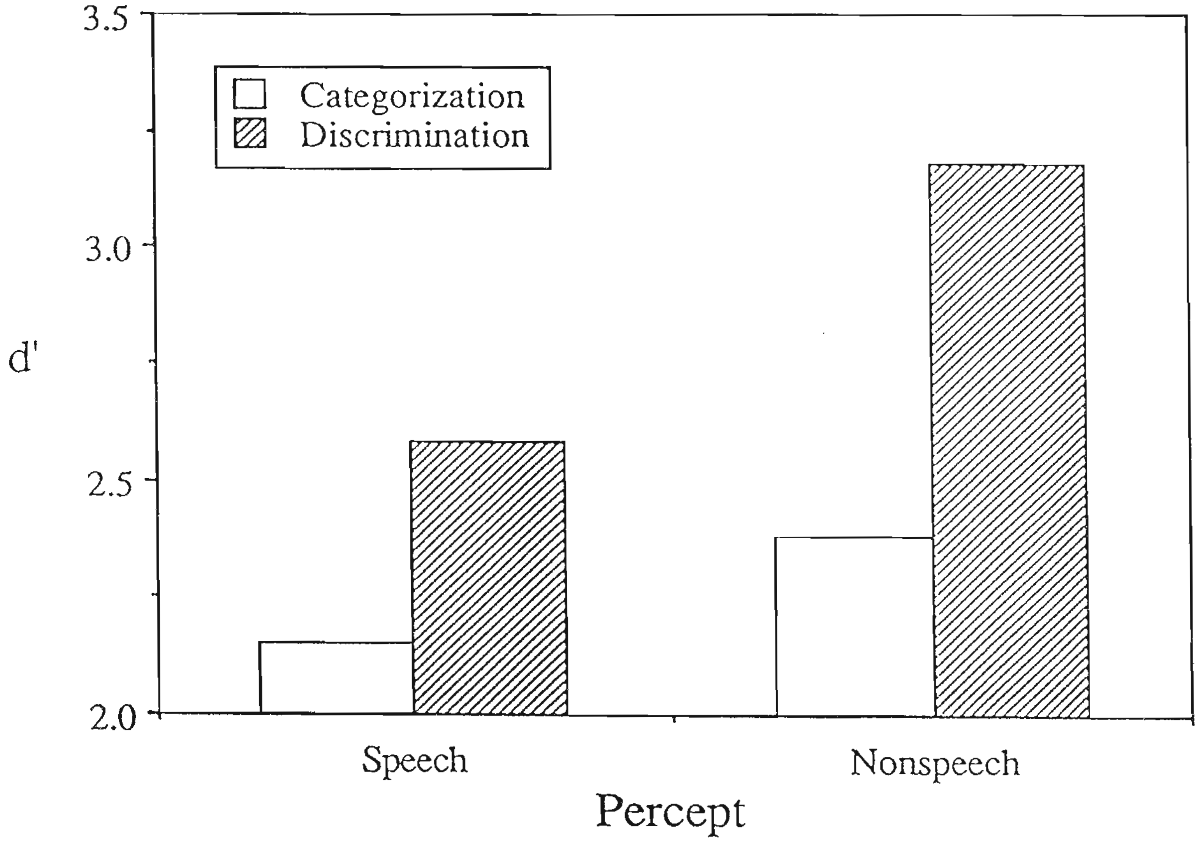


Figure 8. d' data obtained in Experiment 3. Each value averaged across all stimuli. Open bars represent data obtained from identification judgments; striped bars represent data obtained from discrimination judgments.

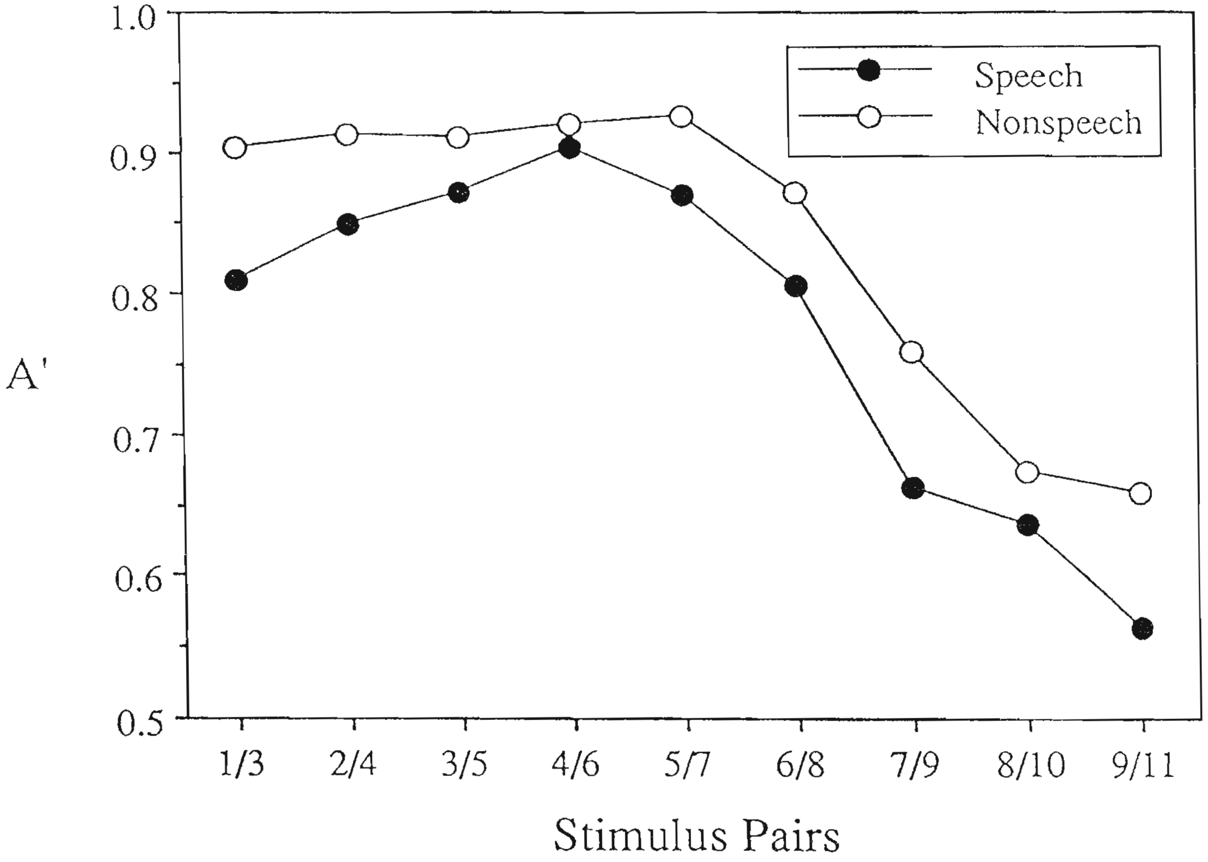


Figure 9. A' data obtained in Experiment 3. Solid dots and connecting lines represent data for subjects reporting speech percepts; open dots and connecting lines represent data for subjects reporting nonspeech percepts.

placed relatively low demands on subjects' memory (Pisoni, 1973). Because subjects in low-demand tasks appear to attend more to fine auditory discriminations than those in high-demand tasks (Pisoni, 1973), the AX task is a more sensitive method of deriving auditory discriminability than paradigms with more observation intervals.

A second related difference between the present methodology and those previously reporting different discrimination results is the size of the ISI. In the present study ISIs were set to 500 ms, whereas in previous studies, ISIs were longer (i.e., 1.5 s in Best et al., 1989). Given that auditory memory decays more rapidly than categorical memory (Pisoni, 1973), and that nonspeech listeners weight auditory information more heavily than speech listeners, it is not surprising that the previous studies utilizing longer ISIs found lower discrimination performance for nonspeech listeners.

Finally, the third difference between the present set of studies and all of the previous studies is the nature of the training regimens. For example, in the Bailey et al. (1977) study, subjects were provided with essentially no training at all. By virtue of listeners' extensive experience listening to and classifying speech sounds, one could reasonably argue that the nonspeech listeners in the Bailey et al. study were at a disadvantage simply in terms of practice. It is likely that with further practice, nonspeech listeners in the Bailey et al. study might have eventually attended to the relevant dimensions of the stimuli, and consequently have performed at levels equal to or better than the speech listeners. A similar criticism may be leveled at the more recent Best et al. (1989) studies. In those experiments, nonspeech listeners were first trained, and in some cases tested, with isolated F3 sinewave analogs, and then tested with three component /ra/ - /la/ analogs without specific training with those more complex stimuli. It is possible that this "perceptual set" treatment of nonspeech listeners may have actually biased them to attend to inappropriate aspects of the three-component sinewave stimuli, further depressing their observed performance.

GENERAL CONCLUSIONS

The hypothesis tested in the present experiments was that speech perception involves the use of automatized, but generic categorization mechanisms. We have noted several studies concerning visual perception which suggest that subjects can develop automatic processing for arbitrary perceptual categories given appropriate experience. Along the way, we noted that it is possible to translate several important concepts associated with the motor theory of speech perception into a more general theoretical framework. One important difference between the more general view and the motor theory view has to do with the role of experience in speech perception. The motor theory and its more recent variants propose that speech perception is the work of an innately specified phonetic module, while the "general" view argues that several characteristics of phoneme perception are the result of experience. Converging evidence for this claim derives from research on music perception. In fact, several similarities between speech perception and music perception suggest that both are mediated by general perceptual mechanisms with similar operational principles.

Music Perception

One fundamental difficulty in studying the role of experience in speech perception is finding naive listeners. The subject pool is composed almost totally of individuals who communicate by means of spoken language. Thus, it is difficult, if not impossible, to perform experiments which contrast expert and naive listeners and control for potential confounds such as maturational level and hearing ability. This problem does not exist in the study of music perception. Both musically sophisticated and musically naive listeners are readily available to participate in perceptual studies.

Several studies have revealed differences between trained musicians and nonmusicians that are remarkably similar to differences observed between speech and nonspeech listeners (Blechner, 1977; Siegel, 1974; Siegel & Siegel, 1977a). For example, Blechner (1977) presented several continua of chords to musicians and nonmusicians for identification and discrimination judgments. Although musicians were able to partition the series in an identification task, nonmusicians were unable to reliably label the endpoint stimuli and the slopes of their labeling functions were generally shallow. Similarly, musicians performed much better than nonmusicians on an oddity discrimination task. These differences parallel those observed between speech and nonspeech subjects in studies utilizing stylized analogs of speech sounds (Bailey et al., 1977; Best et al., 1989; Experiment 1 above).

Likewise, Siegel and Siegel (1977a) observed larger context effects for musicians than nonmusicians. In one experiment, subjects judged the size of musical intervals within a magnitude estimation paradigm. Following six blocks of trials in which subjects estimated the magnitudes of 21 intervals, the stimulus set was changed without the subjects' knowledge. The 11 largest intervals of the original stimulus set were retained and 10 larger intervals were added. Nonmusicians' estimations of the test stimuli were smaller when the stimulus set was altered, but the musicians showed no reliable change. This experiment was repeated using isolated sinewave stimuli in a pitch estimation task and similar results were obtained. The results were similar to those from Experiment 2 that showed that nonspeech listeners were influenced more by context than speech listeners.

Other music perception research has revealed phenomena that once were thought to be unique to speech perception. These include categorical perception (Bachem, 1954; Blechner, 1977; Burns & Ward, 1978; Locke & Kellar, 1973; Pastore, Schmuckler, Rosenblum, & Szczesiul, 1983; Zatorre & Halpern, 1979), critical periods (Sergeant, 1969), maturational changes in sensitivity (Trehub, 1987), cerebral lateralization (Bever, 1980; Bever & Chiarello, 1974; Gordon, 1980; Hirshkowitz, Earle, & Paley, 1978), duplex perception (Pastore et al., 1983), and adaptation effects (Zatorre & Halpern, 1979). For example, Locke & Kellar (1973) found that musicians categorically perceived a continuum of three tone chords ranging from A major to A minor. Observed identification functions were steeply sloped and the discrimination functions exhibited peaks at the category boundary. This finding has been replicated a number of times and with several variations (Blechner, 1977; Siegel & Siegel, 1977b; Burns & Ward, 1978; Zatorre & Halpern, 1979). Siegel (1974), following a study by Bachem (1954), also found that listeners who possessed absolute pitch perceived a frequency continuum of simple sinewave stimuli in a categorical fashion. She also found evidence for dual code processing with these listeners (see Pisoni, 1973 for the speech parallel).

In a similar vein, Pastore et al. (1983) found that musicians perceived three-tone complexes in which the highest and lowest tones were presented to one ear and the middle tone presented to the other ear in a duplex fashion (Rand, 1974). They heard a three tone cord localized at the ear which received two tones (the base) and a single tone localized at the other ear. Labeling and discrimination functions for these two percepts were different although there was some between-subject variability.

These similarities between music perception and speech perception suggest that beyond the phenomenological similarities, there may be functional similarities as well. In particular, music perception by musicians illustrates some properties of a perceptual system which has resulted from extensive exposure to the stimuli. The fact that these properties are very similar to properties found in speech perception suggests that many aspects of speech perception may also be the result of extensive exposure to speech.

SUMMARY

We propose to translate three concepts claimed at various times to be central to the motor theory of speech perception and recent derivatives thereof into more general cognitive/perceptual terms. The "preemptiveness" of speech perception can be expressed in terms of the speed of phonetic categorization relative to less practiced perceptual activities. By virtue of an extensive categorization experience, humans typically attend to or weigh more the categorical phonetic representations of speech sounds, and attend less to the finer-grained auditory properties of stimuli. "Discovering phonetic coherence" in speech signals can be expressed in terms of the focus of selective attention during perceptual categorization. Our nonspeech listeners may initially have attended to aspects of the tonal stimuli which were not correlated with the modeled phonetic distinction. This appeared to be particularly true when the critical auditory information was made less distinctive in the /wa/-/ja/ stimuli. However, with modest amounts of training, the nonspeech listeners attained performance levels comparable to, and in some cases surpassing, the performance of speech listeners. Finally, the relatively "absolute" nature of phonetic categorization, which is essentially a resistance to stimulus order effects, may also be a by-product of its automaticity. These explanations do not make reference to articulation or other specialized mechanisms and instead emphasize the role of more general perceptual and cognitive mechanisms.

In the experiments reported here, we found that nonspeech listeners identified complex acoustic stimuli less quickly and with more influence of context than speech listeners, even when these groups were virtually equivalent in their ability to label items in the continuum. Experiment 3 revealed that nonspeech listeners are better able to make use of auditory information in sinewave stimuli in an AX discrimination task. We argued that these results indicate that, for the speech listeners, categorization was automatic; it was unavoidable even when it caused deteriorated performance in an experimental setting.

Because we are treating the motor theory and information processing accounts of these phenomena as translations of one another, the present data do not eliminate one or the other theory. Any evidence for automatic categorization in speech perception might also be interpreted as evidence for the preemptiveness of speech. However, since it has been claimed that preemptiveness and phonetic coherence are proof that speech perception is special, the fact that these concepts can be translated in a meaningful way into an information processing account invites a serious re-evaluation of the claims surrounding the "specialness of speech". The information processing account views speech perception in terms of a parsimonious theory of perception, and thus, places speech perception within a more general picture of human perception and cognition.

References

Bachem, A. (1954). Time factors and absolute pitch determination. *Journal of the Acoustical Society of America*, **26**, 751-753.

Bailey, P.J., Summerfield, Q., & Dorman, M. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research, SR-51/52.* New Haven, CT: Haskins Laboratories, 1-25.

Bentin, S., & Mann, V.A. (1990). Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes. *Journal of the Acoustical Society of America*, **88**, 64-74.

Best, C.T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, **29**, 191-211.

Best, C.T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception and Psychophysics*, **45**, 237-250.

Bever, T.G. (1980). Broca and Lashley were right, Cerebral dominance is an accident of growth. In Caplan, D. (Ed.), *Biological Studies of Mental Processes*. Cambridge, MA: MIT Press.

Bever, T.G., & Chiarello, R.J. (1974). Cerebral dominance in musicians and nonmusicians. *Science*, **185**, 137-139.

Blechner, M.J. (1977). Musical skill and the categorical perception of harmonic mode. *Haskins Laboratories Status Report on Speech Research, SR-51/52.* New Haven, CT: Haskins Laboratories, 139-174.

Bregman, A.S. (1978). The formation of auditory streams. In Requin, J. (Ed.), *Attention and Performance* VII. Hillsdale, NJ: Erlbaum.

Bregman, A.S. (1987). The meaning of duplex perception, sounds as transparent objects. In Schouten, M.E.H. (Ed.), *The Psychophysics of Speech Perception*. Dordrecht: Martinus Nijhoff Publishers.

Bregman, A.S. (1990). *Auditory Scene Analysis*. Cambridge: MIT Press.

Burns, E.M., & Ward, W.D. (1978). Categorical perception - phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America*, **63**, 456-468.

Cutting, J.E. (1974). Two left-hemisphere mechanisms in speech perception. *Perception and Psychophysics*, **16**, 601-612.

Cutting, J.E., & Rosner, B.S. (1974). Categories and boundaries in speech and music. *Perception and Psychophysics*, **16**, 564-570.

Darwin, C.J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, **33**, 185-207.

Deutsch, D.A., & Roll, P.L. (1976). Separate "what" and "where" decision mechanisms in processing a dichotic tonal sequence. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 23-29.

Flege, J.E. (1988). The production and perception of speech sounds in a foreign language. In Winitz, H. (Ed.), *Human Communication and Its Disorders: A Review.* Norwood, NJ: Ablex.

Flowers, J.H., Polansky, M.L., & Kerl, S. (1981). Familiarity, redundancy, and the spacial control of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 157-166.

Fujisaki, H., & Sekimoto, S. (1975). Perception of time-varying resonance frequencies in speech and non-speech stimuli. In Cohen, A. & Nooteboom, S.G. (Eds.), *Structure and Process in Speech Perception.* New York: Springer-Verlag.

Gardner, R.B., Gaskill, S.A., & Darwin, C.J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America*, **85**, 1329-1337.

Gordon, H.W. (1980). Degree of ear asymmetries for perception of dichotic chords and illusory chord localization in musicians of different levels of competence. *Journal of Experimental Psychology: Human Perception and Performance*, **6**, 516-527.

Grunke, M.E., & Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics*, **31**, 210-218.

Healy A. F., & Repp, B.H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 68-80.

Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. *Journal of the Acoustical Society of America*, **75**, 231-240.

Hirshkowitz, M., Earle, J., & Paley, B. (1978). EEG alpha asymmetry in musicians and nonmusicians: A study of hemispheric specialization. *Neuropsychologia*, **16**, 125-128.

LaBerge, D. (1981). Automatic information processing: A review. In Long, J. & Baddeley, A. (Eds.), *Attention and Performance IX.* Hillsdale, NJ: Erlbaum.

Liberman, A.M., Harris, K.S., Kinney, J.A., & Lane, H. (1961). The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, **61**, 379-388.

Liberman, A.M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics*, **30**, 133-143.

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

Locke, S., & Kellar, L. (1973). Categorical perception in a non-linguistic mode. *Cortex*, **9**, 355-369.

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

Mann, V.A., & Liberman, A.M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, **14**, 211-235.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, **2**, 131-157.

Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., & Dooling, R.J. (1976). Discrimination and labelling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, **60**, 410-417.

Neumann, O. (1984). Automatic processing. In Prinz, W. & Sanders, A.F. (Eds.), *Cognition and Motor Processes*. Berlin: Springer-Verlag.

Nusbaum, H.C. (1984). Possible mechanisms of duplex perception: "Chirp" identification versus dichotic fusion. *Perception and Psychophysics*, **35**, 94-101.

Nusbaum, H.C. Schwab, E.C., & Sawusch, J.R. (1983). The role of "chirp" identification in duplex perception. *Perception and Psychophysics*, **33**, 323-332.

Nygaard, L.C., & Eimas, P.D. (1990). A new version of duplex perception, Evidence for phonetic and nonphonetic fusion. *Journal of the Acoustical Society of America*, **88**, 75-86.

Parker, E.M., Diehl, R.L., & Kluender, K.R. (1986). Trading relations in speech and nonspeech. *Perception and Psychophysics*, **39**, 129-142.

Pastore, R.E., Ahroon, W.A., Wolz, J.P., Puleo, J.S., & Berger, R.S. (1975). Discrimination of intensity differences on formant-like transitions. *Perception and Psychophysics*, **18**, 224-226.

Pastore, R.E., Ahroon, W.A., Puleo, J.S., Crimmins, D.B., Golowner, L., & Berger, R.S. (1976). Processing interaction between two dimensions of nonphonetic auditory signals. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 267-276.

Pastore, R.E., Harris, L.B., & Kaplan, J.K. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America*, **71**, 430-436.

Pastore, R.E., Schmuckler, M.A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception and Psychophysics*, **33**, 469-474.

Pisoni, D.B. (1973). Auditory and phonetic codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253-260.

Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D.B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, **15**, 285-290.

Pisoni, D.B., Carrell, T.D., & Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, *34*, 314-322.

Pisoni, D.B., Logan, J.S., & Lively, S.E. (1991). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In Nusbaum, H.C. & Goodman, J. (Eds.), *Development of Speech Perception: The Transition from Recognizing Speech Sounds to Spoken Words*. Cambridge: MIT Press.

Ralston, J.V., & Sawusch, J.R. (1984). Perception of sine wave analogs of stop consonant place information. *Journal of the Acoustical Society of America*, *76*, *Supplement 1*, M7.

Rand, T.C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

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

Repp, B.H. (1984). Against a role of "chirp" identification in duplex perception. *Perception and Psychophysics*, **35**, 89-93.

Repp, B.H. (1987). The role of psychophysics in understanding speech perception. In Schouten, M.E.H. (Ed.), *The Psychophysics of Speech Perception*. Dordrecht: Martinus Nijhoff Publishers.

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

Schroeder, M.R., Atal, B.S., & Hall, J.L. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In B. Lindblom & S. Ohman (eds.) Frontiers of Speech Communication Research. London: Academic Press.

Schwab, E.C. (1981). Auditory and phonetic processing for tone analogs of speech. *Doctoral dissertation, State University of New York at Buffalo.*

Sergeant, D. (1969). Experimental investigation of absolute pitch. *Journal of Research in Music Education*, **17**, 135-143.

Shiffrin, R.M. (1987). Attention. In Atkinson, R.C., Hernstein, R.J., Lindsey, G. & Luce, R.D. (Eds.), *Stevens' Handbook of Experimental Psychology, Volume 2.* New York: Wiley.

Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic information processing, II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, **84**, 127-190.

Siegel, J.A. (1974). Sensory and verbal coding strategies in subjects with absolute pitch. *Journal of Experimental Psychology*, **103**, 37-44.

Siegel, J.A., & Siegel, W. (1977a). Absolute identification of notes and intervals by musicians. *Perception and Psychophysics*, **21**, 143-152.

Siegel, J.A., & Siegel, W. (1977b). Categorical perception of tonal intervals: Musicians can't tell sharp from flat. *Perception and Psychophysics*, **21**, 399-407.

Tomiak, G.R., Mullennix, J.W., & Sawusch, J.R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America*, **81**, 755-764.

Trehub, S.E. (1987). Infants' perception of musical patterns. *Perception and Psychophysics*, **41**, 635-641.

Whalen, D.H., & Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, **237**, 169-171.

Zatorre, R.J., & Halpern, A.R. (1979). Identification, discrimination, and selective adaptation of simultaneous musical intervals. *Perception and Psychophysics*, **26**, 384-395.

Some Effects of Training Japanese Listeners to Identify English /r/ and /l/[1]

Scott E. Lively, David B. Pisoni and John S. Logan

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

[1] This is a draft of a chapter to appear in Y. Tohkura (Ed.), *Speech Perception, Production and Linguistic Structure*. Tokyo: OHM Publishing Co. Ltd. This research was supported by NIH Research Grant #DC-00111-14 to Indiana University.

Abstract

One of the key issues in cross-language research on speech perception deals with the acquisition of nonnative phonological contrasts in adult listeners. Recent research examining the perception of the English /r/-/l/ contrast in native speakers of Japanese is reviewed in this chapter. The findings that Japanese speakers have difficulty discriminating English /r/ from /l/ are critically evaluated and discussed in light of a number of recent training studies. We conclude that with modest amounts of laboratory training, Japanese subjects can, in fact, improve in their ability to identify English /r/ and /l/. However, successful perceptual learning appears to depend on the presence of stimulus variability in the training materials. Our training results indicate that listeners encode a highly detailed representation of the stimuli and that selective attention to contrastive stimulus dimensions mediates changes in the structure of the developing perceptual categories. We suggest that exemplar-based models of categorization and selective attention, such as those proposed by Nosofsky (1986, 1987) and Kruschke (1990), can account for the development of new phonological categories by nonnative speakers of a language.

Some Effects of Training Japanese Listeners to Identify English /r/ and /l/

Researchers in the field of speech perception have had a longstanding interest in training Japanese listeners to identify the English phonemes /r/ and /l/. There are several reasons for this line of research. First, the /r/-/l/ contrast is primarily spectral in nature. This provides an interesting contrast to the research of Abramson and Lisker (1967) who examined the cross-language perception of a temporally defined cue, voice onset time (VOT) in stop consonants. Second, the /r/-/l/ contrast is extremely difficult for Japanese learners of English to produce and perceive. Third, previous researchers have claimed that training Japanese listeners to identify /r/ and /l/ in the laboratory is an arduous task that is likely to meet with little success (Strange & Jenkins, 1978). The claims about /r/ and /l/ contrast sharply with results showing that native speakers of English can be quickly trained to discriminate new nonnative contrasts varying in VOT (Pisoni, Aslin, Perey, & Hennessy, 1982; Pisoni, McClaskey, & Carrell, 1983). Finally, and most generally, studying the /r/-/l/ contrast in native speakers of Japanese is important because it provides some insight into the malleability of the adult perceptual system. It allows researchers who are interested in issues of speech perception to tie their research to the findings of others who are working on issues of perceptual development, selective attention, and categorization.

In this chapter we review several recent attempts to modify the perception of /r/ and /l/ by Japanese speakers of English. This research can be divided into three lines of inquiry. The earliest work on this problem showed that Japanese listeners have difficulty discriminating English /r/ from /l/ (Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; MacKain, Best, & Strange, 1982; Mochizuki, 1981; Mann, 1985). A second line of research demonstrated the contextual sensitivity of Japanese listeners to the /r/-/l/ contrast (Mochizuki, 1981; Sheldon & Strange, 1983; Dissosway-Huff, Port, & Pisoni, 1984). This work focused on the ability of Japanese listeners to identify /r/ and /l/ produced by several talkers across a variety of phonetic environments. More recently, a third line of research has been concerned with laboratory training studies. In this chapter, we discuss the usefulness of identification and discrimination approaches to training Japanese listeners to perceive English /r/ and /l/ and the importance of stimulus variability in perceptual learning (Gillette, 1980; Strange & Dittmann, 1986; Logan, Lively, & Pisoni, 1991; Lively, Logan, & Pisoni, 1991). Following the discussion of training methods, we examine the role of selective attention in category learning and discuss the implications of training for the types of mental representations that listeners develop while learning to perceive a new phonetic contrast.

General Demonstrations

Early studies on the perception of /r/ and /l/ were designed to show that Japanese listeners could not reliably identify and discriminate between the two sounds in the same manner that native speakers of English did (Miyawaki et al., 1975; MacKain, Best, & Strange, 1982; Mochizuki, 1981; Mann, 1985). Although many studies were dedicated to cataloging the phenomenon, few made any effort to explain the source of the differences between native speakers of English and native speakers of Japanese. Much of the work used synthetic speech stimuli and required Japanese and English listeners to identify and discriminate among stimulus tokens.

In general, the results of the early studies were very similar. Miyawaki et al. (1975) provided one of the earliest experimental demonstrations that adult Japanese listeners were unable to discriminate /r/ and /l/. The researchers showed that Japanese listeners could not, in general, discriminate between two synthetic utterances that spanned the /r/-/l/ category boundary. When only the contrastive /r/-/l/ third formant was presented as a nonspeech control, however, Japanese listeners could easily make the discrimination. Miyawaki et al. concluded that native language experience shaped the way listeners perceived the speech signal, but did not change the basic underlying auditory capabilities of the listener.

MacKain, Best and Strange (1981) and Mochizuki (1983) extended Miyawaki et al.'s initial work by showing that Japanese listeners begin to naturally acquire the /r/-/l/ contrast only after a great deal of experience in an English speaking environment. Additionally, Mochizuki demonstrated that Japanese listeners relied heavily on an intermediate /w/ category when identifying ambiguous synthetic stimuli. Native speakers of English, in contrast, allowed ambiguous stimuli between /r/ and /l/ to remain ambiguous.

The general conclusion to be drawn from these initial experiments is that Japanese listeners do not identify or discriminate between English /r/ and /l/ in the same way that native speakers of English do. Native language experience was given as a nonmechanistic explanation for the differences in performance.[2] Two important findings that must be incorporated into any mechanistic account of the perception of /r/ and /l/ by native speakers of Japanese are the differences in perception across different phonetic environments and the time course of perceptual learning. In the next section, we review several studies that have examined sensitivity across different phonetic environments and then discuss some hypotheses that relate performance measures to properties of the stimulus materials.

Contextual Sensitivity of /r/ and /l/

Early studies demonstrated that Japanese and English listeners differ in how they identify and discriminate between /r/ and /l/. Later studies addressed the questions about whether Japanese listeners' /r/-/l/ judgements are uniformly poor across all phonetic environments and across talkers. This question motivated a number of studies that varied both the phonetic environment and the voice of the talker. Each of the studies that examined the perception of /r/ and /l/ across phonetic environments sampled stimuli from some combination of the environments shown in Table 1.

Insert Table 1 about here

Taken together, the studies all produced similar results. Goto (1971) provided the earliest demonstration of the context sensitive nature of /r/-/l/ perception by native speakers of Japanese. Subjects appeared to be sensitive to the phonetic environment in which the /r/-/l/ contrast occurred. Unfortunately, Goto did not report any statistics on this trend. In addition, he found that listeners were differentially sensitive to the voice of the talker producing the contrast. In some cases, Japanese subjects were more accurate at identifying utterances produced by native speakers of English than they were at identifying their own utterances. Mochizuki (1981), Sheldon and Strange (1982) and Henly and Sheldon (1986) expanded on Goto's results by confirming that Japanese listeners' abilities to discriminate /r/ and /l/ varied as a function of the phonetic environment in which the contrast was produced. Across each of these studies, Japanese listeners were most accurate at identifying /r/ and /l/ in word final position and least accurate for /r/ and /l/ in word initial position.

Several hypotheses have been offered to explain the context sensitivity of /r/-/l/ perception. Dissosway-Huff, Port, and Pisoni (1984) reported that the duration /r/ and /l/ in word final position tends to be longer than in other positions. This increased duration may provide additional perceptual cues for subjects. Sheldon and Strange (1982) made acoustic measurements of the stimulus tokens used in their study and found that /r/ and /l/ in initial consonant clusters had more rapid formant transitions than /r/

[2] It should be noted, however, that with extensive experience or under the proper experimental conditions, Japanese listeners can be shown to be sensitive to the /r/-/l/ contrast. Mann (1985), for example, showed that Japanese listeners were sensitive to the coarticulatory effects of /r/ and /l/ when they preceded /d/ and /g/, despite the fact that they could not consistently identify /al/ or /ar/ syllables in isolation.

Table 1

Examples of phonetic environments.

Environment	Example
r/l v ...	rock-lock
c r/l v ...	cram-clam
...v r/l v...	oreo-oleo
...v r/l c...	mars-malls
...v r/l	fear-feel

and /l/ in other phonetic environments. They suggested that when /r/ and /l/ occur in initial consonant clusters, much of the segment may be devoiced due to a voiceless preceding stop consonant. They also claimed that the third formant transitions may fall short of their normal steady-state target values (see Lehiste, 1964, for a detailed review of the acoustic-phonetic characteristics of /r/ and /l/). The phonotactic rules of Japanese place additional constraints on the Japanese listener when identifying /r/ and /l/ in initial consonant clusters because initial consonant cluster constructions are not found in Japanese. Goto added some additional speculation as to why differences are observed among talkers. He suggested that Japanese listeners are attuned to a narrow set of talkers with whom they have had much listening experience and that when new talkers are introduced, listeners must retune some perceptual or attentional mechanism to the way in which the new talker produces the /r/-/l/ contrast.

In summary, two consistent observations were made in the early studies of /r/-/l/ perception by Japanese listeners. First, Japanese and English listeners do not identify or discriminate between /r/ and /l/ in the same way. Second, Japanese listeners are sensitive to the environment in which the /r/-/l/ contrast occurs. Given these two basic findings, the question arises as to whether Japanese listeners can be trained to improve their identification and discrimination of /r/ and /l/. In the next section, we review several recent training efforts to train Japanese listeners to perceive the /r/ and /l/ contrast.

Laboratory Training Studies on the Perception of /r/ and /l/

Several attempts to train Japanese listeners to identify English /r/ and /l/ have been carried out over the years. Each study differed somewhat in its approach. The assumptions underlying these efforts came from disparate sources, ranging from classical psychophysics to developments in cognitive approaches to categorization. Factors such as task demands, type of speech stimuli, and the role of stimulus variability were critical variables in each of the training experiments. The work carried out to date has provided a mixed set of results. Some studies have shown little generalization to novel tokens after training while others have had moderate success in training Japanese listeners to identify /r/ and /l/ in constrained, but generalizable contexts.

Gillette (1980). Gillette (1980) carried out the earliest study to training Japanese listeners to identify /r/ and /l/. Three native speakers of Japanese were trained in an eclectic paradigm that stressed both production and perception. Gillette claimed that over the course of four training sessions her subjects made modest improvements in their ability to identify /r/ and /l/. She did not provide any statistics on the amount of improvement that was observed. However, she concluded that Japanese listeners could be trained in a relatively short period of time to perceive the /r/-/l/ contrast. Gillette stressed that the use of a wide range of production and perceptual tasks was important in training Japanese listeners to perceive the /r/-/l/ contrast. Her emphasis on variability in stimuli and task is consistent with a cognitively-oriented approach to category learning.

Strange and Dittmann (1984). Strange and Dittmann (1984), in contrast, took a psychophysically-oriented approach to the issue of variability in training by focusing on the minimal cues that differentiate /r/ from /l/. They used a pretest-posttest design and trained subjects in an AX fixed-standard training task with synthetic speech. The AX task requires subjects to make a same-different judgement against a constant standard stimulus. The data Strange and Dittmann reported suggest that while subjects did become quite good at discriminating "rock" to "lock" over the course of training, there was little, if any, generalization to either a synthetically produced "rake"-"lake" continuum or to naturally produced tokens containing /r/ and /l/. Strange and Dittmann's results indicated that subjects engaged in stimulus-specific learning and did not form a single, context invariant representation of the /r/-/l/ contrast. Evidence to support both of these conclusions comes from the finding that subjects were able to discriminate among pairs of stimuli within the same phonetic category. Rather than focusing on the similarities between stimuli of the same phonetic type, subjects learned to direct their attention to the subtle cues that differentiated each of the stimuli within a perceptual category. In our view, their use of an AX fixed standard training task and synthetic speech carries with it four very strong assumptions about the nature of what Japanese subjects learn during training.

First, Strange and Dittmann assumed that discrimination training would facilitate the formation of robust new phonetic categories. In the AX fixed-standard discrimination task, subjects were required to make a same-different judgement about a pair of stimuli from a synthetic speech continuum. The effect of this type of training was that subjects' attention was drawn to very subtle cues differentiating individual stimuli. Subjects trained with this task learned to make very fine, within-category discriminations but had difficulty generalizing to novel tokens.

A second assumption, related to the use of a discrimination task, was that training in a discrimination task would transfer to an identification task. This strategy implicitly required subjects to learn category labels as they made discrimination responses. Thus, subjects must not only discriminate differences between two stimuli, but also must classify each stimulus into higher-level categories. Subjects in an identification task, in contrast, are only asked to make higher-order categorization response (see Jamieson & Morosan, 1986; Jamieson & Morosan, 1989).

A third assumption underlying Strange and Dittmann's study was that subjects would be able to learn abstract phonological units during training, despite being trained with only consonants in initial position, using a single vowel context produced by a single synthetic talker. The authors assumed that the representation developed with a single phonetic environment would be broad and robust enough to transfer to the /r/-/l/ contrast in other environments and to tokens produced by different talkers. This assumption was similar to the template scheme proposed by Henly and Sheldon (1986). Subjects compared the input to be categorized against a template stored in memory. If the input corresponded to the template within some criterial range, then accurate discrimination and identification was expected. Strange and Dittmann's training strategy dictated that acoustic-phonetic variability observed across phonetic environments and across talkers would be incorporated into the abstract phonological units subjects are assumed to learn.

A fourth assumption was also made regarding the use of synthetic speech as training stimuli. Strange and Dittmann assumed that the precise control the experimenter can exercise over the acoustic-phonetic input compensates for the lack of redundant cues in synthetic speech as compared to natural speech (Pisoni, Nusbaum, & Greene, 1985). Formant transitions and stimulus durations can be carefully manipulated and equated using synthetic speech. Additionally, coarticulation and multiple cues can be precisely controlled or eliminated. Thus, presenting a highly constrained, but carefully controlled stimulus set allowed Strange and Dittmann to precisely control the input subjects received. According to their strategy, robust categories would be formed by presenting stimuli that contrasted only the cues that were essential for the discrimination of /r/ from /l/ by native speakers of English.

Variability in Perceptual Learning of /r/ and /l/

As mentioned above, Strange and Dittmann's subjects did not improve significantly in their ability to identify naturally produced stimuli. Given their failure to show generalization to natural speech, we began a series of laboratory training studies to find out why Japanese listeners had so much difficulty learning the /r/-/l/ contrast (Logan, Lively, & Pisoni, 1991). We hypothesized that several of the assumptions made by Strange and Dittmann may have been partially responsible for the poor generalization performance.

Strange and Dittmann (1984) trained subjects using a discrimination procedure. As mentioned earlier, discrimination procedures draw subjects' attention to subtle, within-category differences among the stimuli. Subjects are implicitly encouraged to detect small differences among the stimuli. Strange and Dittmann also forced subjects to shift attentional strategies by requiring them to perform identification tasks in the pretest and the posttest but discrimination tasks during the training phase. In contrast, we used a uniform identification procedure during training and testing. This strategy had two advantages over the previous work. First, subjects in an identification task are required only to make a gross categorization response. Thus, they are encouraged to group similar stimuli into the same perceptual categories. Second, the consistent use of an identification paradigm does not require subjects to change attentional strategies between training and testing: Subjects could apply the same attentional strategy throughout the experiment.

Strange and Dittmann also assumed that subjects would learn an abstract unit or template that was applicable across all phonetic contexts. This assumption ignored the effects of context sensitivity demonstrated by Mochizuki (1981) and Gillette (1980). We adopted a more flexible training strategy based on the visual category learning experiments of Posner and Keele (1968). Posner and Keele demonstrated that robust category learning was facilitated when subjects were trained with a highly variable stimulus set. Subjects in our study were trained with multiple tokens from five phonetic environments produced by five different talkers. Phonetic environments included /r/ and /l/ in all of the environments shown in Table 1. In using this method, we assumed that subjects would develop representations that would be robust across a variety of phonetic environments.

We also took a different approach to the selection of stimulus materials used during training. Strange and Dittmann used synthetic speech because precise control could be exercised over the input signal. In contrast, we used natural speech in both training and testing. Redundant cues that might be eliminated by the use of synthetic speech would be present in natural speech and our assumption was that the redundant cues inherent in naturally produced speech would actually aid listeners in forming robust new phonetic categories (see Elman & McClelland, 1986).

Strange and Dittmann's final assumption was that listeners could learn a new contrast by presenting stimuli that varied only in the contrastive cues relevant to native speakers of English. According to this assumption, variations in the productions of /r/ and /l/ across different talkers should not affect the development of the new perceptual categories. The stimulus materials Strange and Dittmann used were produced by only a single synthetic talker. In contrast, we used multiple talkers during training and tested generalization to talkers who were not used during the training phase. Again, the use of multiple talkers during training increases the amount of variability that subjects were exposed to during learning. As noted above, Goto (1971) claimed that inexperienced Japanese listeners are typically attuned to the way a single talker or small set of talkers produce the /r/-/l/ distinction. By using multiple talkers during training, we hoped to create robust representations incorporating the variability produced by several talkers.

Based on these assumptions, we trained a group of six Japanese speakers of English for 15 sessions in a pretest-posttest design using a two-alternative forced-choice identification task. During the first five days of training, subjects were presented with a new talker each day. Each talker produced the same set of 136 words. These consisted of sixty-eight minimal pairs of English words that contrasted /r/ and /l/ in five phonetic environments. During the next ten days of training, the cycle of five talkers was repeated two more times. During each training trial, subjects were presented with the printed form of a minimal pair of words contrasting /r/ and /l/ on a CRT monitor. Subjects then heard one member of the pair and were asked to press a button corresponding to the word they heard. If a listener made a correct response, the series of training trials continued. If a listener made an incorrect response, the minimal pair remained on the monitor, a light on the response box corresponding to the correct response was illuminated, and the stimulus word was repeated. Accuracy and response time were recorded for each trial during each training session. In addition to training, subjects were also given a pretest and a posttest. These tests were identical to those administered by Strange and Dittmann. We also presented two further tests of generalization which contained new words produced by one of the five training talkers and an additional set of new words produced by a novel talker.

Pretest-Posttest Results. Identification accuracy improved significantly from the pretest to the posttest. Correct identifications increased from a pretest level of 78.1% to a level of 85.9% in the posttest, demonstrating the effectiveness of the training procedure. As in several of the previous studies (Mochizuki, 1981; Sheldon & Strange, 1982), large and reliable effects of phonetic environment were also observed. Subjects were most accurate at identifying /r/ and /l/ in word final position. A significant interaction between the phonetic environment and pretest-posttest variables was also observed. Subjects improved more in initial consonant clusters and in intervocalic position than they did in word initial and word final positions. The lack of improvement for word final position can be accounted for because pretest performance was at ceiling level in this environment (95.8% correct).

Training Results. The training results indicated that subjects' performance improved as a function of week of training. The largest gain in accuracy came from Week 1 to Week 2 of training. The gain from Week 2 to Week 3 was slightly smaller. Each of the six subjects improved, although large differences in absolute levels of performance were observed among the subjects.

Subjects' identification accuracy increased not only as a function of week, but also as a function of the talker producing the contrast. Listeners identified tokens produced by Talkers 4 and 5 more accurately than stimuli produced by Talkers 1 and 2. It may be the case that Talkers 4 and 5 provided a richer set of cues for subjects to respond to. In turn, this may have caused those talkers to be more intelligible to Japanese listeners. Talkers 1 and 2 may present an impoverished set of cues which subjects find difficult to attend to. Alternatively, these talkers may present an inconsistent set of cues that mislead listeners as to the identity of the target segment.

Data collected from each of the five phonetic environments used in training replicated the previous findings (Gillette, 1980; Mochizuki, 1981; Sheldon & Strange, 1982). Targets appearing in final position and in final consonant clusters were identified at near ceiling levels of accuracy. Accuracy in initial position, initial consonant clusters and intervocalic position ranged from 70-80% correct. In addition to the differential sensitivity across phonetic environments, an interaction between the phonetic environment and talker was also observed. Subjects were at near ceiling levels of performance for targets occurring in word-final position for words produced by all talkers. In the three poorly identified environments, however, performance varied widely as a function of talker. Performance was relatively good for items produced by Talkers 4 and 5. However, subjects were less accurate when identifying words produced by Talkers 1 and 2. Figure 1 displays the percentage of correct responses in each phonetic environment as a function of talker.

Insert Figure 1 about here

A similar pattern of data was obtained in the analysis of the response latencies. Two distinct patterns of latencies across week of training were observed. These patterns depended on the phonetic environment of the contrast. Subjects' response times decreased steadily in the environments in which initial performance was good (e.g., final position and final consonant clusters). For environments in

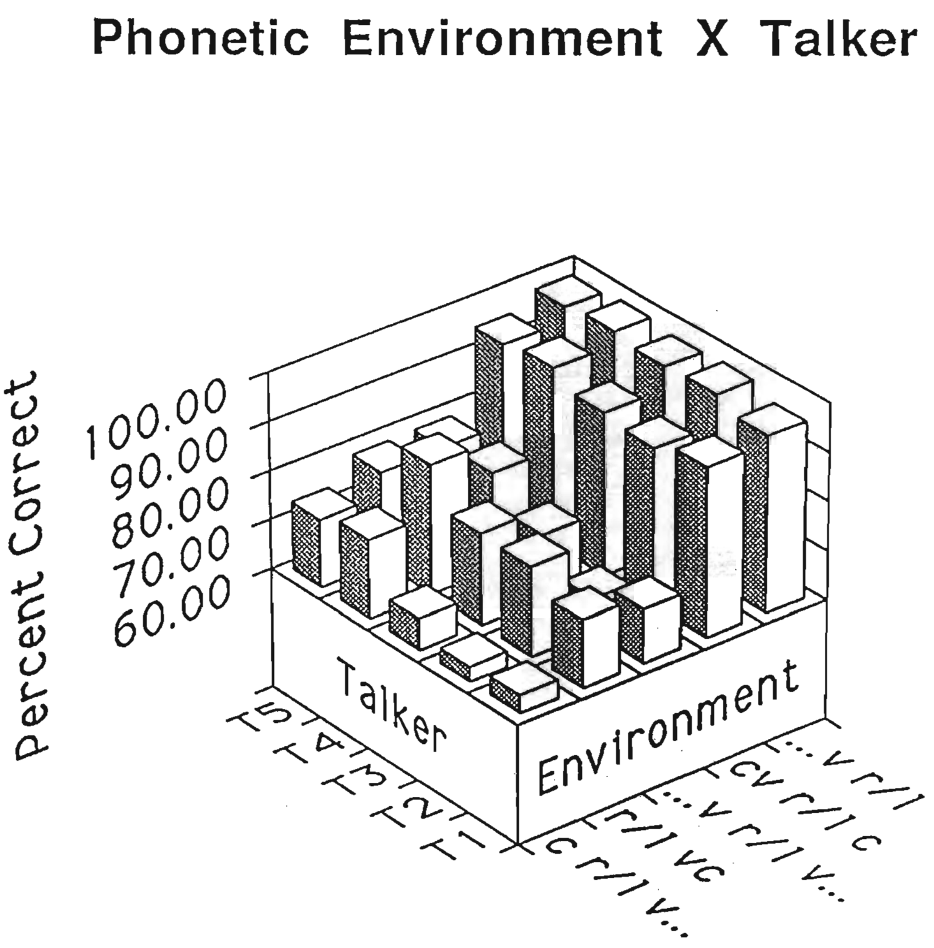


Figure 1. The interaction between phonetic environment and talker observed in Logan, Lively and Pisoni (1991).

which initial accuracy was low, response times increased from week 1 to week 2, but then decreased from week 2 to week 3. Figure 2 presents the interaction of week by phonetic environment observed in the response time data. This pattern of results suggests that subjects required more training trials to determine the appropriate contrastive cues in initial position, initial consonant clusters and intervocalic position. Cues to the identity of targets occurring in final positions, in contrast, were quickly and accurately detected from the outset of training.

Insert Figure 2 about here

Test of Generalization Results. The tests of generalization provide an additional way of assessing the effectiveness of the training procedure. Subjects were presented with new words spoken by training Talker 4 and with new words spoken by a novel talker. The /r/-/l/ contrast occurred in all five phonetic environments and listeners were required to perform the same two-alternative forced choice identification task. Accuracy was marginally greater for words produced by the training talker, compared to the novel talker (83.7% vs. 79.5%). No significant difference was observed between the novel talker and Talker 4 in the response time data. The results of the test of generalization suggest the high degree of context sensitivity inherent in the perceptual learning of these contrasts: Listeners are sensitive both to the voice of the talker producing the contrast and to the phonetic environment in which the contrast occurs.

We concluded from this first study that training with multiple talkers in multiple phonetic environments could be an effective means for improving Japanese subjects' identification of /r/ and /l/. Subjects improved over the course of training, but were differentially sensitive to the voice of the talker producing the contrast. These results led us to examine further the relative importance of stimulus variability in perceptual learning.

Effects of Phonetic Environment

More recently, we have replicated and extended our earlier findings to address the issues of effects of phonetic environment and talker variability. In one of our experiments, we reduced the number of training environments to the three most difficult environments (initial position, initial consonant clusters and intervocalic position) and we increased the number of training trials. The rationale behind these modifications was twofold. First, subjects were at asymptotic levels of performance for stimuli containing /r/ and /l/ in word final positions in the previous study. Thus, little benefit could be derived from increased training in these environments. Second, Atkinson (1972) demonstrated that subjects show increased amounts of learning if they are forced to practice harder items more often. We reasoned that if more trials were dedicated to the difficult environments, greater improvements in overall performance might be observed.

The design of our first experiment was identical to the earlier study of Logan et al. (1991), except that training occurred with a reduced set of exemplars. Accuracy measures and response times were collected during all phases of testing and training.

Pretest-Posttest Results. As in the previous experiment, subjects' accuracy increased from the pretest (79.96%) to the posttest (85.57%). In addition, a decrease in response times from the pretest (1018.46 ms) and the posttest (931.47 ms) was observed. Furthermore, subjects demonstrated contextual sensitivity to the environment in which /r/-/l/ occurred. Their identification performance was at near ceiling levels for targets occurring in word-final position and was least accurate for targets occurring in initial consonant clusters.

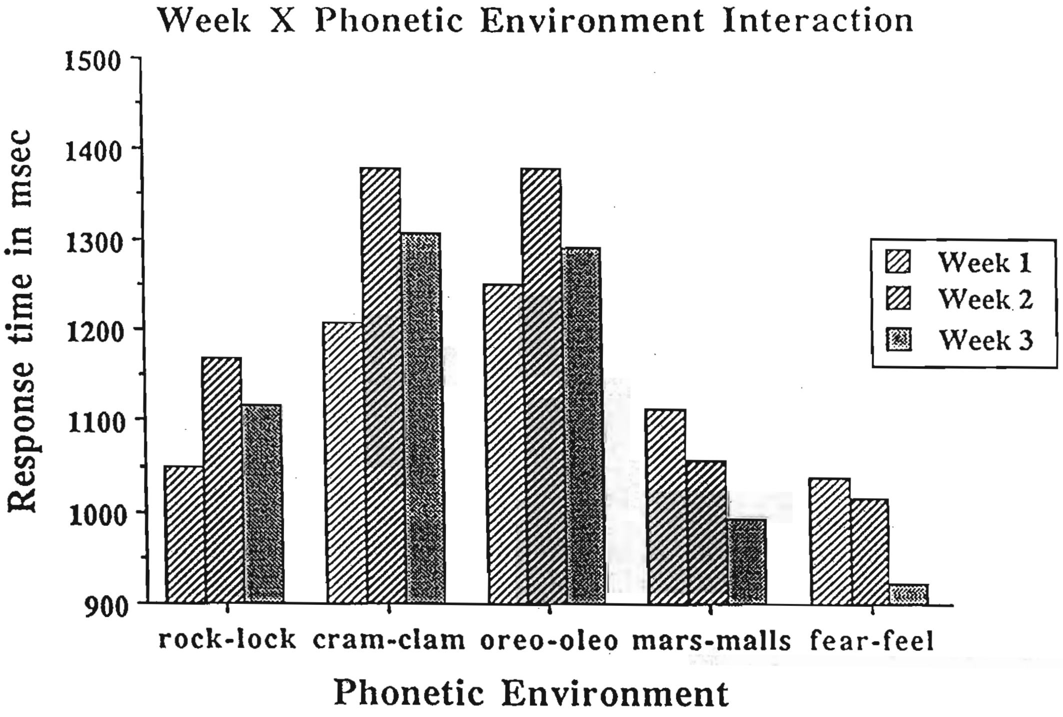


Figure 2. The interaction between phonetic environment and week of training, observed in Logan, Lively and Pisoni (1991).

Training Results. The data collected during the training phase also replicated our earlier results. Subjects improved as a function of week of training. The magnitude of change was equivalent to that observed by Logan et al. Subjects improved the most from Week 1 to Week 2. Smaller gains were observed from Week 2 to Week 3. Response time also decreased steadily as function of week across all phonetic environments. Subjects became more accurate and faster at identifying the target segments over the course of training. As in the Logan et al. (1991) study, subjects were differentially sensitive to the talker who produced the contrast. The rank ordering of talkers replicated the pattern found in the earlier study, although response time patterns differed slightly between the two studies. Talkers producing highly intelligible contrasts were responded to rapidly, while less intelligible talkers were responded to more slowly. Again, this indicates that listeners may attune perceptual or attentional mechanisms more easily to some talkers than to others.

Despite the increased number of training trials incorporating stimuli from difficult phonetic environments, subjects still had difficulty identifying /r/ and /l/ in initial consonant clusters. Average performance did not exceed 75% correct even after three weeks of training. In addition, response times varied as a function of phonetic environment, with fastest responses for targets occurring in word initial position and slowest responses for targets in initial consonant clusters. We observed a steady decrease in response times across the three weeks of training in each phonetic environment. Small increases in accuracy were also observed. Figure 3 displays accuracy and latency measures as a function of week and phonetic environment. The data suggest that a great deal more training would be needed to allow subjects to reach ceiling levels of performance.

Insert Figure 3 about here

Test of Generalization Results. Results from the test of generalization differed from those obtained by Logan et al. (1991). No significant difference in identification accuracy was observed between new words produced by Talker 4 and new words produced by a novel talker. Accuracy for both talkers was comparable to the average performance across all training sessions using Talker 4. However, response times were slower when responding to the new talker, however. This pattern indicates that subjects generalized to a limited degree to new tokens produced by a familiar talker.

In summary, the data from our second experiment indicate that training with a reduced set of phonetic environments and multiple talkers can be effective in training Japanese listeners to identify /r/ and /l/. Subjects' accuracy and response times improved from the pretest to the posttest and during training. The perceptual learning appears to have generalized to a limited extent to new words and to a new talker. Despite the increased number of training trials in difficult environments, gains were still very modest. The results indicate that a much longer training phase might be needed in order to demonstrate additional improvements in the identification accuracy of /r/ and /l/ in difficult phonetic environments.

Training with a Single Talker

Goto's experiment and our previous studies raise several interesting questions about what Japanese listeners incorporate in their developing representations of /r/ and /l/. Strong arguments in favor a context sensitive representation could be made on the basis of the numerous reports of how phonetic environment affects identification performance. In addition to the context sensitivity of phonetic environment, there is also evidence that subjects encode highly specific information about the talker producing the stimulus word. To assess the effects of talker variability on the formation of robust new phonetic categories, we carried out a third experiment in which subjects were trained using the voice of only a single talker. If subjects become attuned to a particular talker when learning to perceive the new

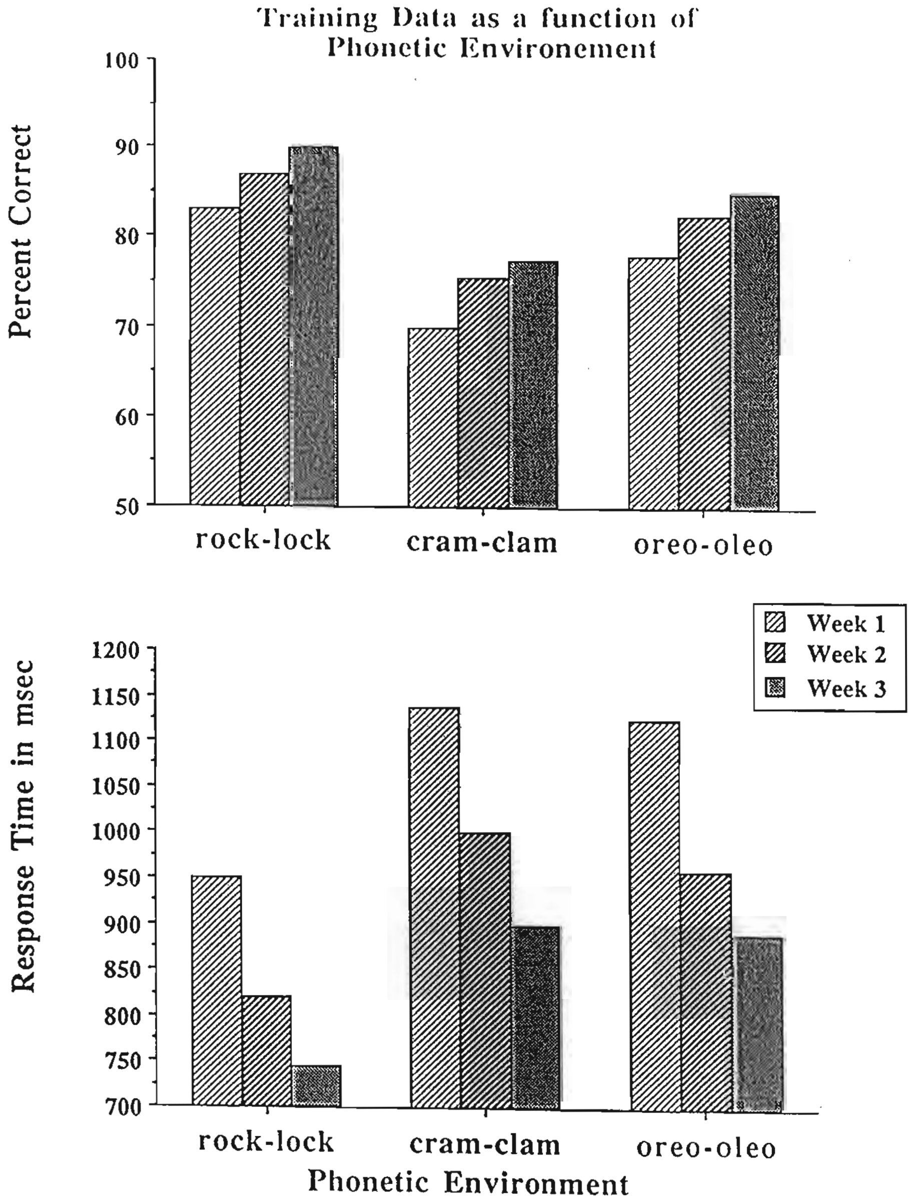


Figure 3. The interaction between phonetic environment and week of training. The upper panel shows accuracy data and the lower panel shows response time data.

contrast, as Goto (1971) claimed, then listeners should become very proficient at identifying targets produced by the training talker. Performance with new talkers, however, should suffer due to the lack of experience with these talkers.

Listeners were trained with a single talker (Talker 4) who produced the /r/-/l/ contrast in five phonetic environments. All other aspects of the training and testing phases remained the same.

Pretest-Posttest Results. An examination of the data comparing the pretest and the posttest performance showed that subjects improved in their ability to identify /r/ and /l/ over the course of training. Average performance increased from 70% correct to 79% correct. The 9% improvement was comparable to improvements observed in the two previous experiments. A large decrease of over 500 ms in average response time was also observed from the pretest to the posttest.

Training Results. Data collected during training replicated the previous results. Subjects made the greatest amount of improvement from Week 1 to Week 2 and a small gain from Week 2 to Week 3. Corresponding decreases in response times were also observed. Changes in both accuracy and response times are comparable to those obtained in our previous studies. Differential sensitivity to the phonetic environment of the /r/-/l/ contrast were also observed. Subjects were most accurate for word final contrasts and least accurate for initial consonant clusters. A speed accuracy tradeoff was suggested in the response time data because listeners were fastest when responding to contrasts in initial consonant clusters and slower for contrasts in word final positions.

Test of Generalization Results. The test of generalization revealed a pattern of results that was consistent with our predictions. Subjects were more accurate at identifying targets produced by the training talker than they were for words produced by a new talker. Response times were also faster for the training talker than to the new talker. This pattern is not surprising considering the extensive amount of training with the same talker. Subjects were exposed to more than 2000 utterances from this talker over the course of training. Listeners became highly familiar with the manner in which the training talker contrasted /r/ and /l/. As Goto (1971) suggested, listeners may incur a severe penalty when they are forced to realign perceptual or attentional mechanisms in order to perceive the manner in which a new talker makes the /r/-/l/ contrast.

In summary, we concluded that the representation of the new contrast that Japanese listeners developed during training was sensitive both to the phonetic environment in which the contrast occurred and to the voice of the talker producing the contrast. Subjects benefited from a training program that emphasized stimulus variability on both dimensions. Generalization performance was impaired by a reduction in variability, although further experimentation is necessary to test this claim. When listeners are presented with only a single talker during training, they may become too finely tuned to the manner in which that talker produces the contrast. Retuning perceptual mechanisms to new talkers may require considerable effort on the part of the listener. Alternatively, subjects trained with only a single talker may lack a strong base of exemplars to compare new inputs against.

Implications for Laboratory Training Studies. Based on our recent training studies, several general conclusions about training Japanese listeners to perceive the /r/-/l/ contrast can be made. First, our results reveal the importance of three types of stimulus variability in the training procedure. Multiple talkers produced each of the stimulus words in three to five phonetic environments. Variability was also increased by presenting a minimum of ten words in each of the phonetic environments. A minimum of 136 unique training stimuli were presented to subjects over the course of the three training experiments. We used more than ten times as many training stimuli as Strange and Dittmann (1984) did in their study.

at the beginning of training than do representations of the contrast in initial consonant clusters. A priori, categorization of /r/ and /l/ in final positions would be predicted to be more accurate than categorization of initial consonant clusters. As outlined above, there are several reasons why the cues in initial consonant clusters may be difficult to identify. These include the short duration of the segment, its coarticulation with preceding segments and its failure to reach its normal steady state. Because of the less salient cues in initial consonant clusters, subjects may find it more difficult to stretch apart representations of the /r/-/l/ contrast in this environment.

A second issue that must be considered in category formation is the talker producing the contrast. Subjects in our experiments appear to have encoded highly specific information about the voice of the talker producing the stimulus. Large and consistent differences were observed among the talkers. Listeners attended more accurately to the dimensions contrasting /r/ and /l/ when the stimuli were produced by Talkers 4 and 5. Highly salient, redundant cues may facilitate a perceptual reorganization of the psychological space for the /r/-/l/ contrast.

The specific type of representation that subjects develop over the course of training must take into account the fact that the perception of /r/ and /l/ is highly context dependent. Strange and Dittmann (1984) assumed that subjects would develop an abstract unit which would suffice as a template for /r/ or /l/ in any phonetic environment. We demonstrated context dependency for the talker producing the stimulus and for the phonetic environment in which the contrast occurs. Our results support an exemplar storage model (e.g., Hintzman, 1986; Gillund & Shiffrin, 1984) which would also be consistent with the assumptions of Nosofsky's and Kruschke's models. Within the framework of these models, an object is categorized by comparing how similar it is to each exemplar of the category. Exemplars that are highly similar to other members of the category are good category members.

Categorization of /r/ and /l/ by Japanese listeners may occur in an analogous manner. At the outset of training, subjects have a very sparse set of exemplars to compare against new inputs. Additionally, these exemplars may not be well differentiated within the underlying psychological space. The lack of clearly differentiated exemplars would lead to many confusions. As training proceeds, subjects stretch the perceptual space containing /r/ and /l/ according to the contrastive dimensions. The selective attention mechanism may be more difficult to tune for talkers who produce highly similar /r/s and /l/s or for environments which are differentiated by a less salient set of cues. In both cases, this would result in high within-category similarity and high between-category similarity. Thus, the psychological distance between two categories would be very small and no simple decision boundaries could be drawn.

Two distinct advantages are provided by the exemplar-based category learning strategy. First, if a template matching strategy is invoked, the template must serve as an adequate comparison for all incoming stimuli. It must be sensitive to variability in both phonetic environments and talker. The template is destined to fail if it is strictly matched against an incoming stimulus. The high variability of the training set would seem to make it impossible to form an abstract representation of the new contrast which is independent of the voice producing the stimulus and the environment in which it is produced.[3]

[3] If the subject is allowed to form multiple templates, each of which is sensitive to a talker's voice, a phonetic environment, or both, then a revised view of the template matching model might be preserved. The revised template view may offer an economization of storage over the exemplar model.

The second advantage the exemplar model offers is that it can act as a template model at the time of retrieval. When subjects are asked to identify a stimulus as an /r/ or /l/, the exemplar-storage model predicts that subjects compute the similarity of the incoming item to all other items in memory. A response would be based on the degree of similarity between the input representation and its similarity to all other items in that space. A group of objects in memory that is highly similar to the input stimulus would serve as an ad hoc template against which the stimulus could be matched. If the stimulus is similar to a set of exemplars, then it will be correctly identified. However, if the stimulus is not similar to any group of exemplars in memory or is similar to several groups, then it will be difficult to identify on an absolute basis.

The exemplar view extends the template model in its ability to elegantly reorganize categories through the similarity computation process. Category reorganizations occur by reweighing the importance of the features within a category. Similarities among some features may be accentuated, while similarities among other features may be attenuated, allowing categories to be stretched along a new set of dimensions (see Pisoni, 1991).

General Conclusions

The studies we have reviewed suggest that training Japanese subjects to identify /r/ and /l/ in the laboratory appears to be a difficult, but not impossible, task. Several factors were shown to be critical to the success of the training program. First, a highly variable stimulus set appears to be crucial for the development of robust new phonetic categories. This variability includes the use of multiple talkers, multiple phonetic environments, and multiple tokens within each phonetic environment. Second, tasks which emphasize similarities among stimuli, rather than minute differences between stimuli, should be used to encourage categorization responses. Third, task demands across training and evaluation phases of the experiment should be held constant. Finally, the training program should employ a natural vocal source that is rich with cues to the identity of the target segment. The research described in this chapter supports the view that context sensitivity to both the phonetic environment in which the contrast occurs and to the voice of the talker producing the contrast play a central role in perceptual learning that may have been underestimated in earlier studies. These forms of context sensitivity indicate that listeners do not form a single, context-independent abstract, phonetic representation. Rather, they seem to form exemplar-based categories which exhibit sensitivity to both the environment in which the contrast occurs and to the voice of the talker producing the stimulus. Selective attention to the relevant contrastive cues may serve as a mechanism to stretch the newly developed representations apart within the psychological space containing the contrast. The operation of selective attention may be best described as a warping of the underlying psychological space. As the space becomes warped by attention, sounds that were previously close to together are spread apart and made more distinctive. Thus, the process of perceptual learning of the sound structure of a language entails the reorganization of sensory input into a set of stable categories in long term memory.

References

Abramson, A.S., & Lisker, L. (1967). Discrimination along the voicing continuum: Cross-language tests. *Proceedings of the 6th International Congress of Phonetic Sciences*, 569-573.

Atkinson, R.C. (1972). Optimizing the learning of a second language vocabulary. *Journal of Experimental Psychology*, **96**, 124-129.

Best, C.T. (in press). The emergence of language-specific phonemic influences in infant speech perception. In Nusbaum, H.C., & Goodman, J. (Eds.), *Development of Speech Perception: The Transition from Recognizing Speech Sounds to Spoken Words*. Cambridge, MA: MIT Press.

Dissosway-Huff, P., Port, R., & Pisoni, D.B. (1982). Context effects in perception of /r/ and /l/ by Japanese. *Research on Speech Perception Progress Report No. 8*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech waveform. In Perkell, J., & Klatt, D. (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum, 360-380.

Flege, J. (1989). The production and perception of foreign language speech sounds. InWinitz, H. (Ed.), *Human Communication and Its Disorders: A Review*. Norwood, NJ: Ablex, 224-401.

Flege, J. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training. *Journal of the Acoustical Society of America*, **86**, 1684-1697.

Gillund, G., & Shiffrin, R. (1984). A retrieval model for both recognition and recall.*Psychological Review*, **91**, 1-67.

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minnesota Papers on Linguistics and the Philosophy of Language, 6*. Minneapolis, MN: University of Minnesota, 59-72.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds ’L’ and’R’. *Neuropsychologia*, **9**, 317-323.

Henly, E., & Sheldon, A. (1986). Duration and context effects on the perception of English /r/ and /l/: A comparison of Cantonese and Japanese speakers. *Language Learning*, **36**, 505-521.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**(4), 411-428.

Jamieson, D., & Morosan, D. (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., & Morosan, D. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, **43**, 88-96.

Jusczyk, P. (1990). Developing phonological categories from the speech signal. Paper presented at The International Conference on Phonological Development. Stanford, CA: Stanford University.

Kruschke, J.K. (1990). ALCOVE: A connections model of category learning. *Cognitive Science Research Report 19.* Bloomington, IN: Indiana University.

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, **7**, 275-309.

Lane, H. (1969). A behavioral basis for the polarity principle in linguistics. In Salzinger, K., & Salzinger, S. (Eds.), *Research in Verbal Behavior and Some Neurological Implications*. New York: Academic Press, 9-98.

Lehiste, I. (1964). Acoustic characteristics of selected English consonants. *International Journal of American Linguistics*, **30**, 10-115.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify /r/ and /l/. *Journal of the Acoustical Society of America*, **89**(2), 874-886.

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

Mann, V.A. (1985). Distinguishing universal and language-dependent levels of speech perception: Evidence for Japanese listeners' perception of "l" and "r". *Cognition*, **24**, 169-196.

McClaskey, C., Pisoni, D., & Carrell, T. (1983). Transfer of training to a new linguistic contrast in voicing. *Perception and Psychophysics*, **34**, 323-330.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J., & 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.

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

Nosofsky, R.M. (1986). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **15**, 700-708.

Pisoni, D.B. (1991). Modes of processing speech and nonspeech signals. In Mattingly, I., & Studdert-Kennedy, M. (Eds.), *Modularity and the Motor Theory of Speech Perception: Proceedings of a Conference to Honor Alvin M. Liberman*. Hillsdale, NJ: Erlbaum. 225-238.

Pisoni, D.B., Aslin, R.N., Perey, A.J., & Hennessy, B.L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, **8**(2), 297-314.

Pisoni, D., Nusbaum, H., & Greene, B. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **73**(11), 1665-1676.

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

Sheldon, A., & 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., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-/l/ by Japanese adults learning English. *Perception and Psychophysics*, **36**(2), 131-145.

Strange, W., & Jenkins, J. (1978). Role of linguistic experience in perception of speech. In Walk, R.D. & Pick, H.L. (Eds.), *Perception and Experience*. New York: Plenum Press, 125-169.

Terbeek, D. (1977). A cross-language multidimensional scaling study of vowel perception. *Working Papers in Phonetics, 37.* Los Angeles, CA: University of California, Los Angeles.

Comprehension of Synthetic Speech Produced by Rule: Word Monitoring and Sentence-by-Sentence Listening Times[1]

James V. Ralston, David B. Pisoni, Scott E. Lively,
Beth G. Greene and John W. Mullennix[2]

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

[1] The authors thank Susan Rivera for help in running subjects and collecting data in Experiment 2. The authors also thank Paul Luce and Bob Peterson for their comments on an earlier draft of this paper. This research was supported by National Science Foundation Grant #IRI-86-17847 to Indiana University.

[2] Now at Department of Psychology, Wayne State University, Detroit, MI 48202

Abstract

Previous comprehension studies using post-perceptual memory tests have often reported negligible differences in performance between natural and synthetic speech despite large differences in segmental intelligibility. The present experiments investigated the comprehension of natural and synthetic speech using two different on-line tasks: word monitoring and sentence-by-sentence listening. On-line task performance was slower and less accurate for passages of synthetic speech than natural speech. Recognition memory performance in both experiments was less accurate following passages of synthetic speech compared to natural speech. Monitoring performance, sentence listening times and recognition memory accuracy all showed moderate correlations with intelligibility scores obtained using the Modified Rhyme Test. The results suggest that the poorer comprehension of passages of synthetic speech is due, in part, to increased encoding demands relative to natural speech. Compared to earlier studies, the present results demonstrate that on-line tasks can be used to measure differences in comprehension performance between natural and synthetic speech.

Comprehension of Synthetic Speech Produced by Rule: Word Monitoring and Sentence-By-Sentence Listening Times

The recent proliferation of voice output technology has extended its practical range of applications as a useful human-machine interface. Limited-vocabulary applications, such as warning and feedback systems make only modest demands on computer storage capacity by using pre-stored utterances (McCauley, 1984). However, other important applications, such as prostheses for the language-impaired, reading machines for the blind, computer-aided instruction and information retrieval systems require the production of a potentially unlimited vocabulary of unrestricted text (Massey, 1988). In these situations, it is more desirable to make use of speech synthesizers controlled by rule-based systems (McCauley, 1984). Devices that automatically convert text into fluent speech are referred to as text-to-speech (TTS) systems (Klatt, 1987). A wide variety of TTS systems are currently available on the commercial market. These systems range from inexpensive hobbyist products that typically produce speech of poor quality, to relatively expensive systems that produce highly intelligible and natural sounding speech (see Logan, Greene, & Pisoni, 1989).

Due to the increasing use of voice response systems in many applications, it becomes important to study how synthetic speech is perceived and understood by listeners (Pisoni, Nusbaum, & Greene, 1985). The bulk of research on this problem has been concerned primarily with assessing the intelligibility of the segmental properties of synthetic speech (Hoover, Reichle, van Tasell, & Cole, 1987; Logan et al., 1989). For example, Logan et al. (1989) recently reported intelligibility scores obtained using the Modified Rhyme Test (MRT) with synthetic speech produced by eight different TTS devices and a natural speech control (Nye & Gaitenby, 1973). On each trial, subjects heard an isolated monosyllabic consonant-vowel-consonant word and were required to select the correct orthographic label from six response alternatives.

Logan et al. (1989) observed wide variations in the segmental error rates across the various types of speech, from 0.53% for natural speech to 35.56% for the Echo TTS. The observed error rates on the MRT, presumably reflecting the amount and complexity of acoustic-phonetic knowledge embodied in the different systems, were used to define four broad classes of output devices. These classes included natural speech with error rates of less than 1%, high-quality synthetic systems with error rates from 3.25% to 6.72%, moderate-quality synthetic systems with error rates from 12.5% to 17.75% and low-quality synthetic systems with error rates greater than 25%. Thus, at the segmental level, the quality of speech synthesis varies over a wide range. In the best case, the segmental intelligibility comes close to natural speech; in the worst case, the error rates are substantially higher. Low intelligibility scores on the MRT for a particular synthesis device suggest the possibility that listeners may have serious problems in understanding longer messages, particularly if the materials are novel and/or linguistically complex.

Although segmental intelligibility tests such as the MRT provide important information about speech perception using simplified stimuli in highly constrained tasks, it is likely that one may not be able to accurately predict listening performance when the stimulus materials are longer passages of fluent, connected speech. Subjects in forced-choice intelligibility tests such as the MRT are not required to access the meanings of isolated words, but only to recognize different sound patterns. In more natural communication situations where fluent speech is used, listeners must not only recognize words but must also derive meaning from successive strings of words in sentences. Understanding the semantic content of an utterance relies heavily on internalized linguistic knowledge and information in the listener's long-term memory. Listeners in intelligibility tests are also not able to use redundant contextual cues that characterize fluent speech (Miller, Heise, & Lichten, 1951). Given the paucity of research on the perception of fluent connected speech, it is important to develop new comprehension measures for synthetic speech produced by rule and relate them to traditional measures of segmental intelligibility that are typically discussed in the literature. In this report, we are interested in the extent to which comprehension measures can be predicted from segmental intelligibility scores obtained with the MRT. Given this goal, it is useful to review some of the previous work on the comprehension of synthetic speech.

Two distinct lines of research have been used to study the comprehension of synthetic speech. One set of studies has examined isolated sentences using a sentence verification paradigm. Other studies have used memory techniques or traditional multiple-choice questions with much longer passages of connected speech. Results of both methods of measuring listening comprehension have raised several fundamental issues about the comprehension process and the relationship between comprehension and segmental intelligibility of synthetic speech produced by rule.

In the sentence verification paradigm, subjects determine the validity of short sentences vis-a-vis some prior information, such as a picture or general world knowledge (Gough, 1965; Larkey & Danly, 1983). Because the verification task itself is easy, error rates are typically low. Therefore, response latency is the primary dependent variable of interest. Manous, Pisoni, Dedina and Nusbaum (1985) used the sentence verification technique (SVT) in combination with a transcription task to study the comprehension of natural speech and synthetic speech produced by five TTS systems. The verification latency data yielded a rank ordering of the performance of the systems similar to that found in other studies of segmental intelligibility (see Logan et al., 1989). Also, moderate to high correlations were observed between transcription error rate and verification measures. These correlations were $r=-.86$ between transcription accuracy and verification accuracy for correct "true" verification responses, and $r=+.75$ between transcription error rate and verification latency for correct "false" verification responses. Other studies using the SVT task with synthetic speech, natural speech, and coded speech have found similar results (Pisoni & Dedina, 1986; Pisoni, Manous, & Dedina, 1987; Schmidt-Nielson & Kallman, 1987).

The close relationship between segmental intelligibility and sentence comprehension suggests that comprehension of isolated sentences may depend, to a large extent, on phoneme and word recognition. However, it is unclear whether the results obtained with the SVT can be generalized to the comprehension of passages of fluent speech. In the latter case, listeners must integrate information across several sentences to derive the semantic content of an entire text. How well global comprehension performance can be predicted by traditional measures of segmental intelligibility is unknown at this time.

A second class of comprehension studies has examined differences in performance between passages of natural and synthetic speech using multiple-choice questions. Typically, the passages and their corresponding comprehension questions have been drawn from published sources, such as standardized reading comprehension tests (Farr & Carey, 1986). These experiments have produced conflicting results. Oddly, those studies reporting the largest comprehension differences have used the highest quality synthetic speech, whereas those studies reporting negligible to no differences have generally used poor-quality synthetic speech.

Several studies using passages of natural and synthetic speech have found either no differences in comprehension or differences that disappeared with moderate exposure or training (Jenkins & Franklin, 1981; McHugh, 1976; Pisoni, & Hunnicutt, 1980). For example, Pisoni and Hunnicutt (1980) compared the comprehension of passages produced by a human talker and by the MITalk-79 TTS system (Allen, Klatt, & Hunnicutt, 1987) with a reading control. After each passage was presented, subjects answered a series of multiple-choice questions. Comprehension accuracy increased from the first half to the second half of the test session for the passages of synthetic speech. Accuracy for the synthetic speech group was significantly worse than the reading control during the first half of testing, but was equivalent to that for the reading and natural speech controls during the second half of testing.

Other comprehension studies have reported reliable differences in comprehension accuracy between natural and high-quality synthetic speech (Luce, 1981; Hersch & Tartarglia, 1983; Moody & Joost, 1986). In one study, Luce (1981) compared the comprehension of natural speech and synthetic speech produced by the MITalk system with a recognition memory task. Subjects judged the truth of short sentences with respect to information presented in the preceding passage. The sentences tested memory for information that ranged from the occurrence of specific words, to factual information not related to exact wording ("gist"), to information that was available only from semantic inferences based on understanding textual information. Recognition accuracy was uniformly higher for the passages of natural speech. However, Luce also found an interesting crossover interaction between voice and sentence type. Subjects listening to synthetic speech had more accurate memory for words, whereas subjects listening to natural speech had more accurate memory for more abstract propositional information.

Luce suggested that the interaction was evidence for a resource sharing model of spoken language comprehension (LaBerge & Samuels, 1974; Kintsch & van Dijk, 1978). According to this model, the various comprehension processes, from relatively superficial analyses of phonology to relatively deep analyses of semantic information, all share a limited pool of processing resources. Luce argued that subjects listening to synthetic speech allocate a greater proportion of their cognitive resources to analyzing the initial acoustic-phonetic structure of the signal, leaving fewer resources for comprehending the semantic content of the passages. He also proposed that this difference in encoding strategies could explain the observed differences in memory performance for natural and synthetic speech. Other experiments also have found evidence suggesting that perceptual encoding of synthetic speech requires more processing resources (Hersch & Tartarglia, 1983; Lee & Nusbaum, 1989; Luce, Feustel, & Pisoni, 1983; Mack, 1987, 1989).

Taken together, the previous results of intelligibility and comprehension studies of synthetic speech provide a mixed picture. Studies of segmental intelligibility have consistently demonstrated reliable differences in performance between natural and synthetic speech. Similarly, the SVT studies have reported comparable differences with isolated sentences. However, comprehension studies utilizing passages of connected speech have produced conflicting evidence, some reporting sizeable differences between natural and synthetic speech (Hersch & Tartarglia, 1983; Luce, 1981; Moody & Joost, 1986), and others reporting no differences at all (Jenkins & Franklin, 1981; McHugh, 1976; Pisoni & Hunnicutt, 1980; Schwab, Nusbaum, & Pisoni, 1985). How can these findings be accounted for in some principled way? We believe that the discrepancies observed across these various comprehension studies may be the result of the specific experimental methods used to measure comprehension performance and the selection of stimulus materials.

The existence of comprehension differences for fluent connected speech has important practical ramifications. First, a failure to find comprehension differences between natural and synthetic speech might be thought of as being fortuitous from an applications perspective. These results would suggest that a less expensive TTS system may be adequate to output lengthy texts. Second, if real differences do exist and previous experiments have simply been unable to measure these differences reliably, then a low quality TTS system might be less desirable in a variety of applications and may have serious consequences for users.

In the present study, we investigated the second possibility. We hypothesized that previous research using long passages of connected speech may have used relatively crude and insensitive experimental techniques to measure comprehension. In particular, all of the previous studies that have reported negative results used "successive measures" (Levelt, 1978). That is, comprehension was assessed by measuring responses to questions presented *after* listeners heard the test passages. Although successive measures are useful because they examine the contents of memory and the products of the comprehension process, they are known to be highly vulnerable to processes such as inferencing and reconstruction that may selectively modify the memory representation for the stimulus materials (Bartlett, 1932; Kintsch & Van Dijk, 1978; Levelt, 1978).

In contrast, "simultaneous" or "on-line" measures provide an experimental methodology that can be used to assess comprehension as it proceeds in real-time (Levelt, 1978). Because on-line measures are taken during the process of comprehension, they are assumed to be less susceptible to post-perceptual cognitive processes (Levelt, 1978). In the present set of experiments, we have used two on-line tasks -- a word monitoring task and a new sentence-by-sentence listening time task -- with a post-perceptual recognition memory task to examine the comprehension of passages of fluent connected speech. Using these two on-line measures, we hoped to address a number of unresolved issues surrounding the comprehension of passages of synthetic speech.

First, we attempted to determine whether more subtle differences can be detected using a simultaneous measure of comprehension, even if successive measures fail to reveal differences. Previous studies have demonstrated that the word monitoring task is sensitive to a number of linguistic variables (Foss, 1969; Cutler, 1976). The interpretation of monitoring data assumes a limited-capacity model of cognitive processing (Kahneman, 1973). That is, both comprehension processes and the simultaneous monitoring task place competing demands on a common pool of processing resources. If these task demands exceed available resources, then performance on one or both tasks may become degraded. Therefore, as comprehension becomes more difficult, less resources are available for other concurrent tasks. Increases in secondary task speed or decreases in secondary task accuracy are typically interpreted as signs that the difficulty or processing load of the primary comprehension task has increased (Posner & Boies, 1971; Wickens, 1987).

Second, we attempted to determine whether reliable differences in comprehension performance between natural and synthetic speech could be detected using a recognition memory test with more carefully controlled items. The test questions in several of the earlier studies could have been answered without listening to the passages (Jenkins & Franklin, 1981; Moody & Joost, 1986). Therefore, we prescreened all test sentences used in the verification task to eliminate items that could be answered without knowledge of the contents of the passage.

Third, we attempted to determine whether the increased demands associated with the encoding of synthetic speech interact with comprehension load and/or memory load. This issue is important for both theoretical and practical reasons. First, it is of theoretical interest to determine whether perceptual encoding shares common resources with other comprehension processes. Other studies have suggested that comprehension shares resources with other mental processes, such as STM storage (Baddeley & Hitch, 1974; Britton, Holdridge, Westbrook, & Curry, 1978), and that perceptual encoding of synthetic speech shares resources with STM storage (Lee & Nusbaum, 1989; Luce et al., 1983). However, it is not known whether perceptual encoding and other comprehension processes share resources. Second, it is of practical interest to determine the extent to which performance decrements associated with synthetic speech are exacerbated by other cognitive demands. Therefore, difficulty of comprehension was manipulated in the present experiments by employing two sets of passages that differed in terms of their "readability" or "text difficulty" (Kintsch & Vipond, 1979). Memory load was manipulated in the first experiment by requiring subjects to remember and monitor for either zero, two, or four words.

We predicted that if perceptual encoding competes for limited processing resources with higher-level comprehension processes and STM storage, then voice would interact with the other two experimental variables. Specifically, we expected that comprehension performance would be poorer for synthetic speech than natural speech, that performance would be poorer for more difficult texts (and larger target-set sizes) than for easier texts (and smaller target-set sizes), and that the performance decrease as a function of voice would be more pronounced for difficult texts (and larger target-set sizes) compared to easier texts (and smaller target-set sizes).

Finally, we attempted to assess the relationship between segmental intelligibility and comprehension measures. Manous et al. (1985) found a strong relation between transcription accuracy and verification performance using isolated sentences in an SVT (e.g., $r=-.86$ between transcription accuracy and verification accuracy for correct "true" verification responses). However, the relationship between segmental intelligibility and comprehension of fluent speech has never been studied, so far as we know. Because discourse processing relies heavily on linguistic knowledge, one might expect that the correlation between intelligibility and comprehension would be low compared to that observed by Manous et al. (1985).

EXPERIMENT 1

Method

Subjects

Subjects were 112 volunteers enrolled in introductory psychology classes at Indiana University in Bloomington. They received either course credit or were paid four dollars for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Materials

Modified Rhyme Test. A subset of 50 words from the MRT (House, Williams, Hecker, & Kryter, 1965) was selected to assess segmental intelligibility. The MRT measures the intelligibility of syllable-initial and syllable-final singleton consonants using a six-alternative forced-choice procedure.

Insert Tables 1 and 2 about here

Passages. Tables 1 and 2 list the passages used in the different text difficulty conditions. The passages listed in Table 1 were taken from Pauk's Six-Way Paragraphs (Pauk, 1983). Each was designed for fourth-grade readers. The passages listed in Table 2 were selected from a variety of written comprehension tests designed for college-level readers. Tables 1 and 2 also include the results of analyses conducted on the written form of the texts, including standard readability indices. Results of the analyses indicate that the passages in Table 1 are more readable, or less difficult, than the passages in Table 2. The passages were written in narrative or expository styles and generally concerned scientific and cultural topics. We refer to the difference between these sets of passages as "text difficulty", reflecting differences in readability (Haberlandt, 1984; Kintsch & Vipond, 1979).

Table 1

Text statistics for fourth-grade passages.

Measure	Passage					
	Bears	Bees	Birds	Firsts	Turtle	Average
Readability Indices						
Flesch[3]	84	84	89	88	85	86
Gunning's[4]	7	7	5	7	7	6.6
Flesch-Kincaid[5]	5	5	4	3	5	4.4
Paragraph-Level						
No. Paragraphs	7	3	3	3	5	4.2
Avg. No. Sentences	3.5	6.6	4.6	7.6	4.2	5.3
Sentence-Level						
No. Sentences	25	20	14	23	21	20.6
Avg. No. Words	11.9	11.3	11.2	9.5	12	11.2
Word-Level						
Total Number	298	226	157	220	253	230.8
Avg. No. Syllables	1.3	1.3	1.3	1.3	1.3	1.3

[3] Flesch Reading Ease (Flesch, 1949). Varies from 0 (very difficult to read) to 100 (very easy to read).

[4] Gunning's Fog Index (Gunning, 1968). Corresponds roughly to the grade level needed to understand the text.

[5] Flesch-Kincaid (Flesch, 1949). Value represents the average number of school years needed to understand the text.

Table 2

Text statistics and sources for college-level passages. (See Table 1 for an explanation of readability indices.)

Measure	Passage					
	Iberia[6]	Build[7]	Lens[8]	Radio[9]	Speech[10]	Average
Readability Indices						
Flesch	61	68	40	32	43	48.8
Gunning's	14	11	19	22	16	16.4
Flesch-Kincaid	10	9	15	17	13	12.8
Paragraph-Level						
No. Paragraphs	2	2	1	2	3	2
Avg. No. Sentences	6	6.5	7	3	5	5.5
Sentence-Level						
No. Sentences	12	13	7	6	15	10.6
Avg. No. Words	21	18.4	29.5	33	21.7	24.7
Word-Level						
Total Number	252	240	207	198	326	244.6
Avg. No. Syllables	1.5	1.4	1.6	1.7	1.7	1.6

[6] Cooperative English Tests: Reading Comprehension (1960)

[7] Iowa Silent Reading Tests (1972)

[8] The Nelson-Denny Reading Test (1973)

[9] Stanford Test of Academic Skills: Reading (1972)

[10] Stanford Test of Academic Skills: Reading (1972)

Word targets. Target words for each passage were selected on the basis of several criteria, including word-initial phoneme, number of syllables, and grammatical class. Because we used existing texts, we were unable to completely control the properties of the target words. However, target words were selected according to a set of priorities. The most common target words were monosyllabic nouns with word-initial stop consonants. However, if all words of this type had been selected from a passage, other words were used, such as monosyllabic nouns with word-initial fricatives or affricates. The target words were balanced across different experimental conditions.

Recognition memory test. Eight sentences were selected for each passage to test comprehension of the preceding material. These sentences were presented visually on a CRT monitor immediately following the presentation of each passage. The sentences ranged from five to ten words in length and were stated in a declarative form. Four of the sentences assessed listeners' memory for specific words that occurred in the passage (referred to here as "word-recognition" sentences), and four of the sentences assessed memory for propositions of the passage (referred to here as "proposition-recognition" sentences). Based on information in the preceding passage, half of all sentences were true and half were false. Word-recognition sentences all had the frame "The word XXX appeared in the passage." None of the target words in the word-recognition sentences appeared as word monitoring targets in the same passage. True proposition-recognition sentences described relationships between concepts that occurred in the passage. False proposition-recognition sentences contained concepts and relationships from the passage that were recombined in semantically meaningful ways.

A large set of sentences were initially generated and then tested in written form with a separate group of introductory psychology students. The pretest identified proposition-recognition sentences that could be correctly answered without knowledge of the passage and proposition-recognition sentences that could not be correctly answered even after reading the passage. The pretest also identified word-recognition sentences for which verification accuracy was unacceptably low even after subjects had read the passages. In the pretest, half of the subjects were given copies of the passages with the full set of verification sentences, and the other half were given only the proposition-recognition sentences without the passages. Proposition-recognition sentences were discarded if accuracy was less than 80% correct for both test groups of subjects, or accuracy was greater than 80% correct for both test groups. This left a set of sentences for which accuracy was relatively low (50-75% correct) for subjects who had not read the passages and relatively high (75-100% correct) for subjects who had read the passages. Word-recognition sentences were selected from the pretest that were recognized with 85% or better accuracy after subjects had read the passages.

Recording and playback techniques. A male native speaker of American English with a midwestern accent read both the MRT list and the eleven passages. The materials were recorded in an IAC sound-attenuated booth with an Electro-Voice D054 microphone. An identical set of MRT words and passages was produced by a Votrax Type-N-Talk speech synthesizer controlled by a VAX 11/750 computer. No special stress or pronunciation correction was applied to the synthetic speech. Both natural and synthetic speech were recorded onto one channel of an audio tape at 15 ips using an Ampex AG500 tape recorder.

All stimuli were low-pass filtered at 4.8 kHz, sampled and digitized using a 12 bit A/D converter running at 10 kHz. With a digital waveform editing program, flags were set at the onset of each target word using the upper four bits of each sixteen-bit computer word (Luce & Carrell, 1981). All stimuli were equated for RMS amplitude. The speech stimuli were then recorded back onto one channel of an audio tape using a D/A converter. At this time, the bit flags were used to trigger a tone burst generator that output a timing tone onto the second channel of the audio tape. Each audio tape contained a calibration vowel, 50 MRT words, and 11 comprehension passages. Target-set size and voice of the passages were between-subjects variables. Thus, an experimental tape contained either natural or synthetic speech and had either 0, 2, or 4 word targets. Speech signals were mixed with 55 dB broadband noise to mask tape hiss and ambient room noise and were presented to listeners binaurally over TDH-39 matched and calibrated headphones at 80 dB SPL.

Procedure

Subjects were run in groups of five or fewer in a quiet room. Each subject sat at a sound-treated booth equipped with a set of headphones, a video display monitor (GBC Standard CRT Model MV-10A), and a two-button response box that was interfaced to the computer. Each subject was given a booklet containing instructions and response forms for the MRT test, instructions for the comprehension task, and a post-experimental questionnaire that was designed to record subjective reactions to the experimental stimuli and tasks. After reading the MRT instructions, subjects were presented with a block of 50 trials. Subjects were required to circle one of six response alternatives in their answer booklets after each trial.

After completing the MRT, subjects were asked to read the instructions for the comprehension portion of the experiment. These instructions were also thoroughly explained by the research assistant. Subjects were told that they would be required to simultaneously detect all word targets and comprehend the passage, and that they would be required to respond to a series of test sentences at the end of each passage. The first passage served as practice to acquaint the subjects with the experimental procedures and the quality of the voice used in each condition.

Each trial of the comprehension test consisted of three parts. In the first part, either 0, 2, or 4 word targets were presented visually on the CRT display for a 30 second memorization period. When the memorization period ended, the target words disappeared from the screen and the passage was presented over headphones. Subjects were instructed to press a button on the response box each time they detected one of the target words in the memory set. At the end of each passage, eight sentences were presented visually on the CRT monitor for four seconds each. Subjects were instructed to determine whether the sentence was "true" or "false" and to record their response by pressing one of two buttons on the response box while the sentences were presented. No feedback was provided to the subjects about the accuracy of their responses. Button position (left or right) was counterbalanced. Subjects received a five minute break after the fifth passage. The comprehension portion of the experiment was controlled in real-time by a PDP 11/34 minicomputer which presented the target words and test sentences and recorded subjects' responses. An experimental session lasted approximately one hour.

Results and Discussion

Two criteria were used to screen subjects. First, subjects' data were excluded from further analyses if they failed to make any word monitoring responses on at least five of the ten test passages. Second, subjects' data were excluded if they failed to respond to at least eight of the eighty comprehension sentences that were presented after the test passages. The data of 20 subjects failed to meet one or both of these criteria and were eliminated from further analysis. The absolute rate of rejection per condition was relatively low; the largest number of subjects was rejected in the Votrax/4 target condition (4.5% of all subjects). A chi-squared test confirmed that there was no differential rejection rate as a function of experimental condition [$X^2=0.013$, $p > .05$]. The results will be presented in three major sections corresponding to the three tasks: segmental intelligibility, word monitoring, and recognition memory. These data were subjected to analyses of variance, treating voice (natural or Votrax) and target-set size (0, 2 or 4) as between-subjects variables, and text difficulty (fourth-grade or college) and recognition sentence type (word or proposition recognition) as within-subjects variables. All post-hoc statistical analyses were done using Newman-Keuls tests.

Modified Rhyme Test

MRT accuracy refers to the percent correct responses out of 50 trials. Analysis of variance revealed a significant main effect of voice [$F(1,85)=743.65, p<.001$]. None of the other main effects or interactions were significant. Mean accuracy in the MRT was 94% correct for natural speech and 60% correct for synthetic speech. Although large differences in overall error rate existed for the two voices, a higher proportion of errors occurred for word-initial consonants as compared to word-final consonants. The overall level of performance, as well as the pattern of errors noted above, is consistent with previously reported data based on more extensive testing with the full set of MRT words (see Logan et al., 1989).

Word Monitoring

Mean accuracy and latency data were examined in separate analyses of variance. Monitoring hits were defined as those responses that occurred from 100 ms to 1200 ms after the onset of a target word. Responses outside these limits were presumed to reflect either anticipation or inattention by the subjects. Because the probability of false alarms was exceedingly low (the mean p(FA) was less than .01), accuracy is presented in terms of the probability of a correct detection (i.e., a hit).

Insert Figure 1 about here

Accuracy. Word monitoring accuracy data are displayed in Figure 1. Word monitoring accuracy was higher for natural speech than synthetic speech [$F(1,54)=17.39, p<.001$]. Monitoring accuracy was also higher for two targets compared to four targets [$F(1,54)=55.87, p<.001$]. Although the interaction between text difficulty and target-set size was significant in the ANOVA [$F(1,54)=4.32$, $p=.042$], subsequent post-hoc analyses failed to reveal the precise locus of the significant difference between the means underlying the interaction.

Insert Figure 2 about here

Latency. Word monitoring latencies are shown in Figure 2. Overall, monitoring responses were faster for target words in natural speech compared to synthetic speech [$F(1,54)=11.44$, $p<.001$]. Responses were faster for the two-target word condition than the four-target word condition [$F(1,54)=8.76, p=.004$]. Monitoring responses were also slower for the college-level passages than the fourth-grade passages [$F(1,54)=75.36$, $p<.001$]. Finally, the interaction between voice and text difficulty was significant [$F(1,54)=10.53$, $p=.002$]. Post-hoc analyses revealed that the increase in response latency from fourth-grade to college-level was greater for the passages of synthetic speech than natural speech. This result is consistent with the account suggested earlier that perceptual encoding of synthetic speech imposes greater processing demands compared to natural speech, that difficult texts impose greater processing demands compared to easier texts, and that both of these activities compete for the same limited pool of processing resources.

Insert Figure 3 about here

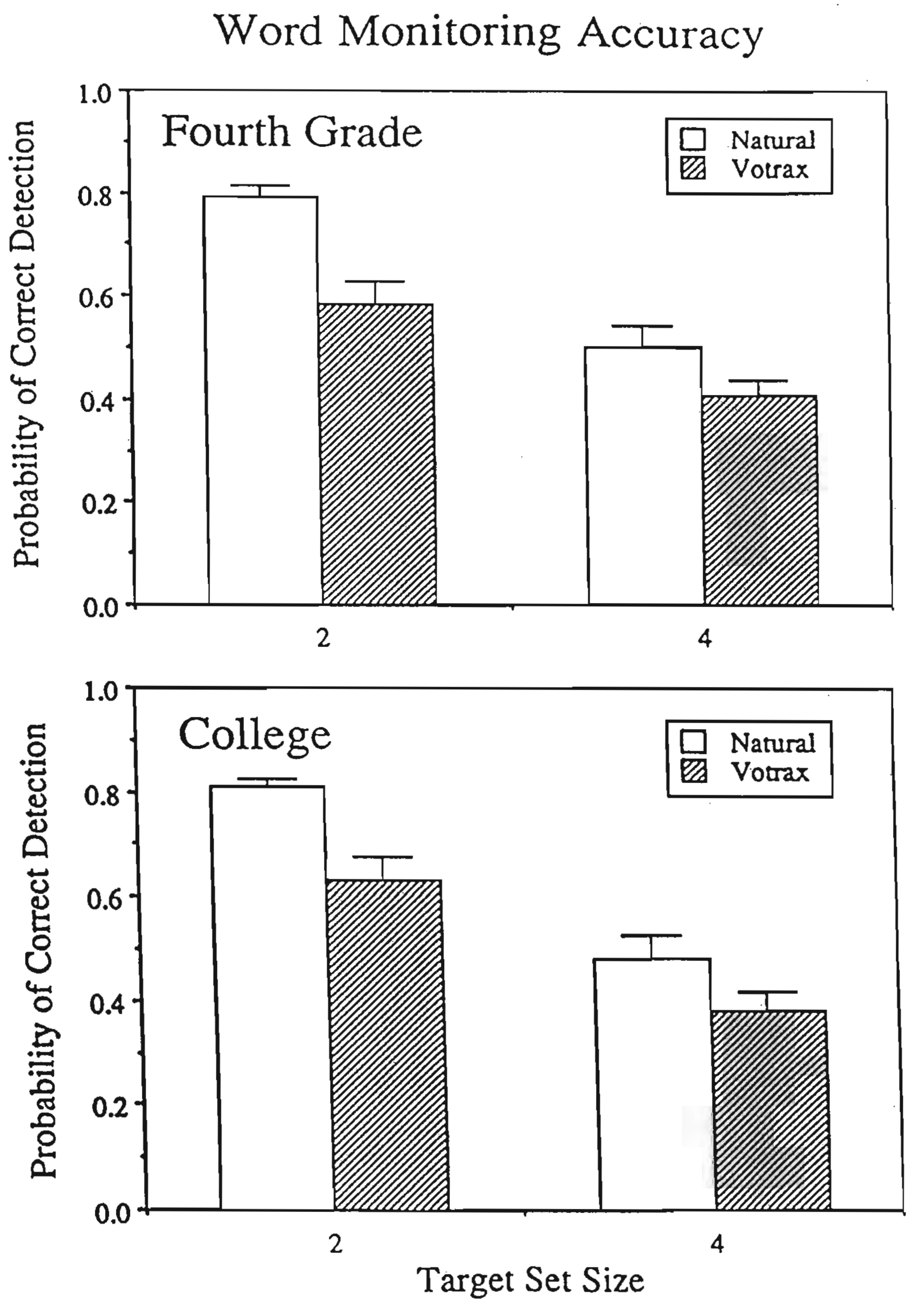


Figure 1. Word monitoring accuracy (probability of a correct detection) as a function of target-set size in Experiment 1. The upper panel shows data for fourth-grade passages and the lower panel shows data for college-level passages. Open bars represent accuracy for passages of natural speech, and striped bars represent accuracy for passages of Votrax synthetic speech. Error bars represent one standard error of the sample means.

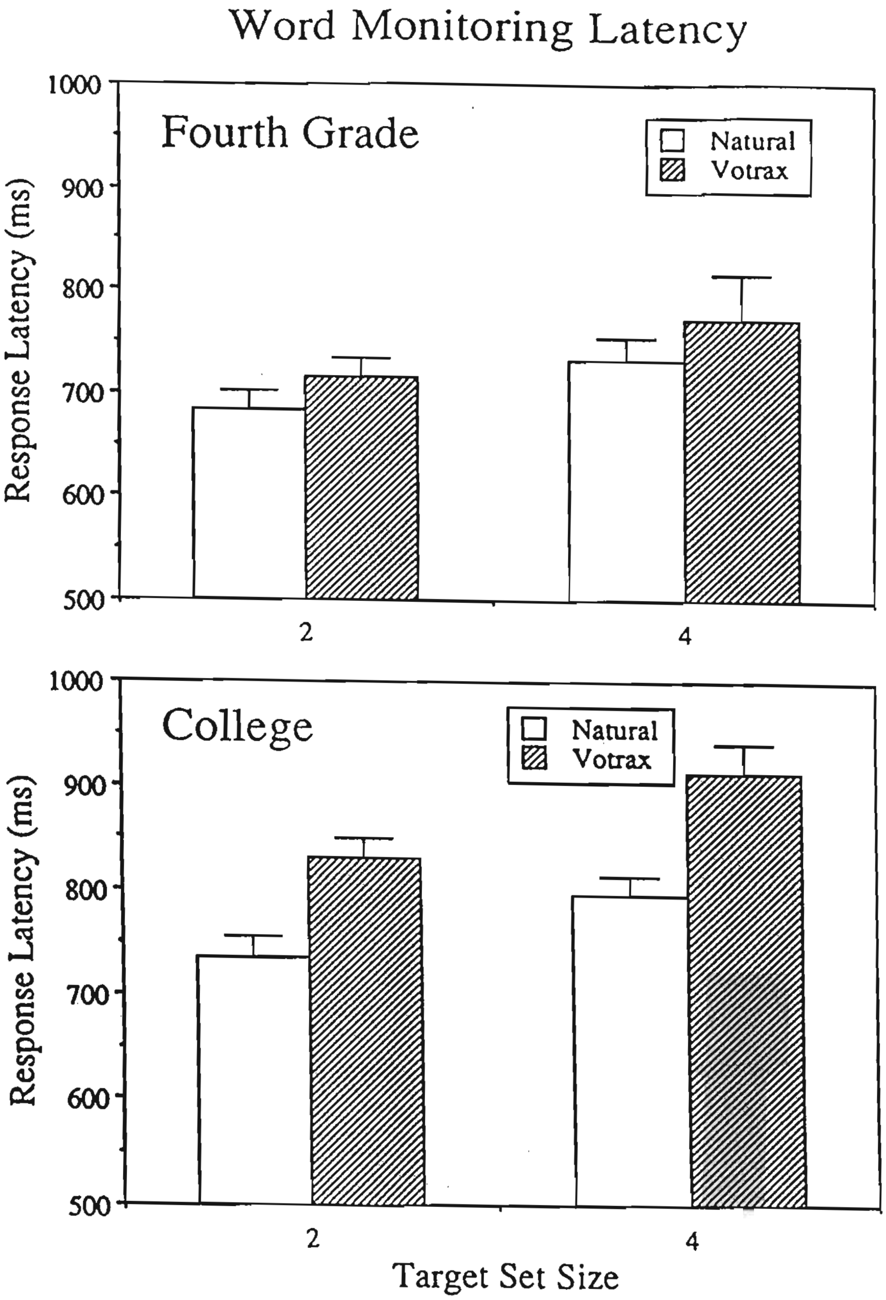


Figure 2. Word monitoring response latency (in ms) as a function of target-set size in Experiment 1. The upper panel shows data for fourth-grade passages, and the lower panel shows data for college-level passages. Open bars represent latencies for passages of natural speech, and striped bars represent latencies for passages of Votrax synthetic speech. Error bars represent one standard error of the sample means.

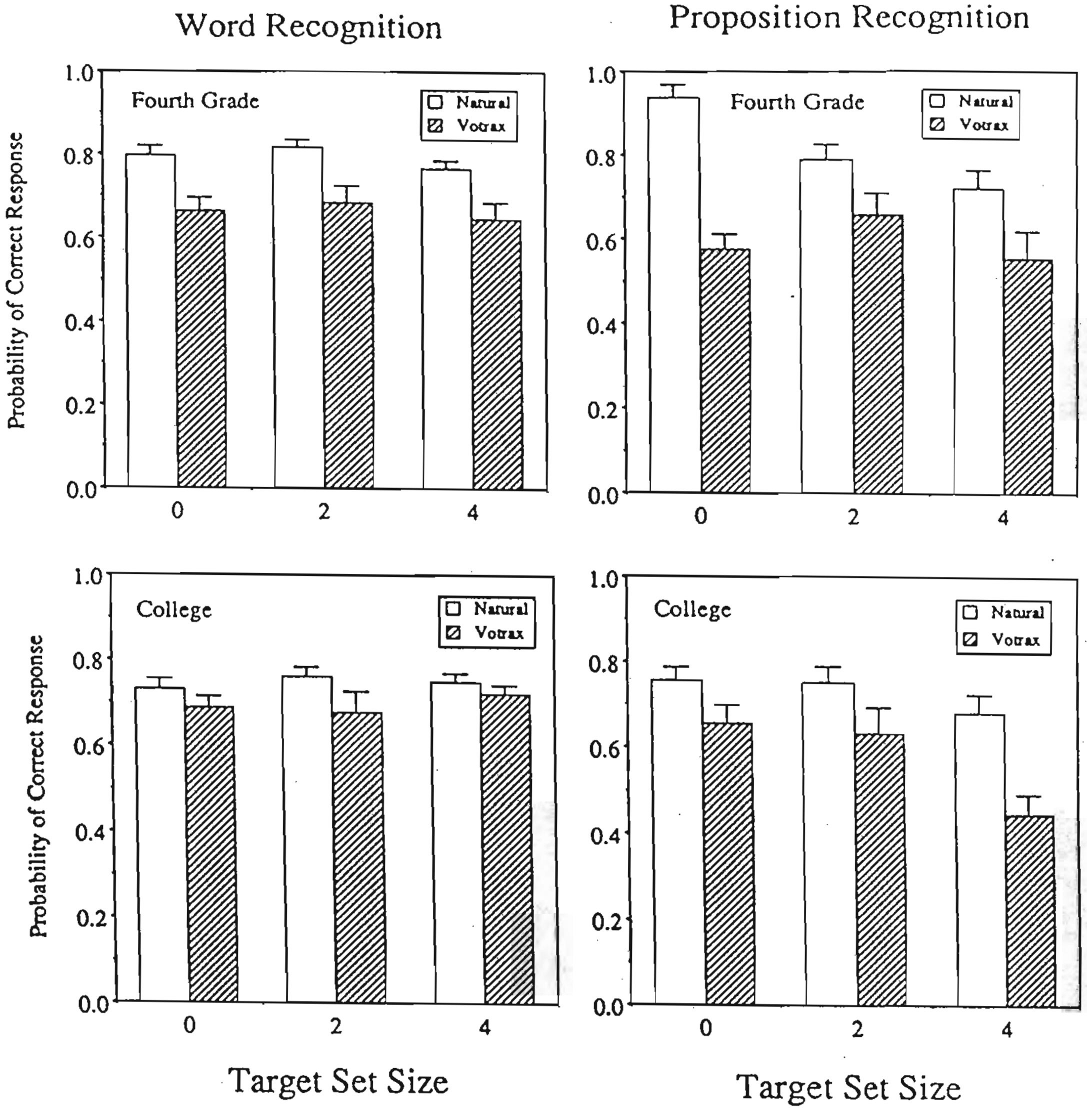


Figure 3. Recognition memory accuracy (probability correct) as a function of target-set size in Experiment 1. The panels on the left show data for word-recognition sentences, and the panels on the right show data for proposition-recognition sentences. The upper panels show data for fourth-grade passages, and the lower panels show data for college-level passages. Open bars represent accuracy for sentences after passages of natural speech, and striped bars represent accuracy for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

Recognition Memory

Recognition accuracy data were analyzed with ANOVA, treating voice and target-set size as between-subjects variables, and treating sentence type and text difficulty as within-subjects variables. The recognition accuracy data are displayed in Figure 3. The panels on the left present data for word-recognition sentences, and the panels on the right present data for proposition-recognition sentences. The top panels present data for fourth-grade passages, and the bottom panels present data for college-level passages.

Subjects recognized sentences more accurately after listening to passages of natural speech than synthetic speech [$F(1,85)=37.21$, $p<.001$]. Accuracy was also higher for word-recognition sentences than proposition-recognition sentences [$F(1,85)=9.94$, $p=.002$]. Apparently, natural speech provides a more robust representation of information in the passages. This finding is consistent with the results of Luce (1981) and Moody and Joost (1986) who reported that passages of natural speech were comprehended better than passages of synthetic speech.

The interaction between voice and sentence type was also highly significant [$F(1,85)=10.95$, $p=.001$]. Post-hoc analyses revealed that, although no difference was observed as a function of sentence type in the natural speech condition, accuracy was significantly lower for the proposition-recognition sentences in the synthetic speech condition. This result is similar to findings reported by Luce (1981). Overall, accuracy decreased with increasing target-set size [$F(2,85)=3.19$, $p=.046$]. Post-hoc analysis revealed that accuracy for the four-target condition was significantly lower than accuracy for the zero- and two-target conditions. However, the latter two conditions were not significantly different from one another. This result replicates findings reported by Bruner and Pisoni (1982) who found that recognition performance after monitoring for a single word target was the same as verification performance after listening to a passage without monitoring for any word targets.

Recognition accuracy was significantly lower for sentences following college-level passages than fourth-grade level passages [$F(1,85)=4.18$, $p=.044$]. In addition, the interaction between sentence type and target-set size was significant [$F(2,85)=6.79$, $p=.002$]. Post-hoc analyses showed that word-recognition accuracy did not vary as a function of target-set size, but that proposition-recognition accuracy was significantly lower for the four-target condition than any of the other conditions. This effect suggests that comprehension and memory for the target words compete for the same limited pool of processing resources.

The three-way interaction between text difficulty, voice and target-set size was significant [$F(2,85)=3.13$, $p=.048$]. Post-hoc analyses revealed an unexpected effect. Accuracy was greater for fourth-grade text than college-level text for listeners in the natural speech condition that were not required to monitor for targets. However, no differences in performance between fourth-grade and college-level texts were observed for the other voice and target-set size combinations.

Insert Table 3 about here

Regression Analyses

Regression analyses were carried out with segmental intelligibility as the predictor variable and word monitoring and recognition accuracy data as predicted variables. Data were collapsed across the voice, text difficulty, and target-set size variables. The results are presented in Table 3. The left column displays the correlation coefficients (Pearson's r) and the right column displays the probability values

Table 3

Correlation of MRT accuracy and comprehension measures in Experiment 1.

Predicted Variable	Correlation	Probability
Word Monitoring		
Prob. of Correct Detection	0.31	0.01
Response Latency	-0.34	<0.01
Recognition Accuracy		
Word Recognition	0.46	<0.01
Proposition Recognition	0.47	<0.01

associated with F-statistics that were used to test whether the observed coefficients were different from zero.

The magnitudes of the correlation coefficients, while significant, are moderate in their predictive power. For example, the correlation between speech intelligibility as measured by the MRT and proposition-recognition accuracy is +.45, indicating that variation in segmental intelligibility accounts for only 21% of the observed variation in the proposition-recognition accuracy scores. Although direct statistical comparisons are not possible, these correlations are, on the whole, smaller than those reported by Manous et al. (1985). Because higher-level cognitive contributions are very important for the comprehension of complex linguistic materials such as passages of connected speech, recognition memory can be predicted only moderately well from intelligibility scores alone.

Summary

The major findings from this experiment can be summarized in terms of the results obtained in each of the three tasks. First, results from the MRT were consistent with previous findings (Logan et al., 1989). Segmental intelligibility was much poorer for Votrax speech than for natural speech. Second, results from the word monitoring task demonstrated that this on-line measure of comprehension was highly sensitive to several experimental variables. Monitoring responses were more accurate and faster for detecting words in passages of natural speech compared to passages of synthetic speech. Accuracy and speed of responding in the word monitoring task were also related to target-set size. As the number of word targets in the memory set increased, accuracy of detection decreased and the latency of the monitoring responses increased. Monitoring responses were also slower when the word targets were embedded in more difficult passages. Finally, a significant interaction was observed between voice and text difficulty in the latency data. The increase observed in monitoring latency for difficult texts was larger for synthetic speech than for natural speech. The results indicate that the comprehension of passages of connected synthetic speech is poorer than natural speech. Further, the observed interaction suggests that stimulus encoding and comprehension processes share a common pool of processing resources (Kahneman, 1973; Posner & Boies, 1971; Wickens, 1987). This point is important -- comprehension deficits for synthetic speech appear to be exacerbated by task difficulty (see also Logan et al., 1989). Several differences in comprehension between natural and synthetic speech were also observed in the recognition memory data. First, recognition responses were more accurate following passages of natural speech compared to synthetic speech. Second, a significant interaction between voice and sentence type was observed. Although no difference was observed between sentence types following natural speech, accuracy was significantly poorer for proposition-recognition sentences following passages of synthetic speech.[11]

Regression analyses revealed moderate, but significant relationships between segmental intelligibility and comprehension accuracy. These findings are not surprising considering that comprehension, especially for more complex texts, involves much more than simply perceiving sound patterns. Spoken language comprehension involves accessing word meanings, parsing information within sentences, integrating information across sentences, and drawing inferences from a semantic propositional base (Kintsch & Van Dijk, 1978). These resource-intensive processes may be more variable across listeners than processes underlying the perception and recognition of words. Thus, it is not surprising

[11] The form of two three-way interactions in the verification data were unexpected and are difficult to explain within the limited capacity framework (Wickens, 1987). In both cases, performance differences between grade four- and college-level texts decreased with increasing target set size.

that comprehension performance for long passages of connected speech is not predicted very well simply by considering segmental intelligibility of isolated words in a highly constrained task like the MRT.

EXPERIMENT 2

A second experiment was conducted to obtain converging evidence about differences in the time-course of comprehension for passages of natural and synthetic speech. The experiment utilized a sentence-by-sentence listening task that was similar to procedures developed in reading research (Miller & McKean, 1964; Kintsch & Keenan, 1973). In the reading studies, passages of continuous text were presented visually, one sentence at a time. Subjects controlled the presentation of successive sentences by pressing a button. The dependent variable, "reading time," was measured as the difference between the onset of visual presentation and the following button press.

Using multiple regression techniques, several studies have demonstrated that sentence-by-sentence reading times are sensitive to a number of linguistic variables that affect reading comprehension (see Haberlandt, 1984, for an exhaustive listing). These have included sentence-level variables such as word frequency (Kieras, 1974), the number of propositions in the sentence (Kintsch & Keenan, 1973), and the imagery value of words (Kieras, 1974). Passage-level factors that influence sentence-by-sentence reading times include the number of required reinstatement searches (Lesgold, Roth, & Curtis, 1979), the presence of "new" topic nouns (Haviland & Clark, 1974), and the "height" or importance of a sentence within the overall text structure (Cirilo, 1981). Typically, the more difficult a passage is, according to any of the above metrics, the longer sentence-by-sentence reading times are using this procedure. The reading times are assumed to be a composite of several time intervals that make up the comprehension processes (Haberlandt, 1984).

In the present experiment, the passages used in Experiment 1 were segmented into individual sentences and presented to listeners one-at-a-time under the control of the subject (Mimmack, 1982). The use of the sentence-by-sentence listening time task (SBSLTT) was motivated by two factors. First, we wanted to develop another procedure that would complement existing on-line techniques that have been used to measure comprehension, such as phoneme and word monitoring tasks. Both phoneme and word monitoring encourage listeners to allocate their attention to the surface structure of a passage rather than its abstract propositional content (Bruner & Pisoni, 1982). Because the SBSLTT made use of the same passages and comprehension task, the results could be compared directly to the results obtained in the first experiment.[12] This comparison provided a baseline to assess the effects of the SBSLTT task on comprehension performance.

The second motivation for using the SBSLTT was to provide converging evidence about the effects of synthetic speech on comprehension. Because the SBSLTT measures on-going performance during comprehension, it provides a second on-line index of processing load. As such, it can be used to assess the generality of the effects observed in Experiment 1 using the word monitoring procedure.

[12] All comparisons with Experiment 1 were carried out only with data from the zero target-set size condition. This condition is comparable to the listening conditions of Experiment 2 because subjects were not required to maintain any target words in memory.

Method

Subjects

The subjects were thirty adult native speakers of English. They were paid $5.00 each for their participation. All subjects were living in the Bloomington, Indiana area at the time of testing, and none reported any speech or hearing difficulties at the time of testing. None of the subjects were in the previous experiment.

Materials

The MRT word lists, the comprehension passages, and the recognition sentences were the same as those used in Experiment 1.

Stimulus Preparation

Modified Rhyme Test. The MRT preparation and presentation methods were identical to those used in Experiment 1.

Passages. Digitized versions of the ten passages used in Experiment 1 were segmented into individual sentences using a digital waveform editing program. All sentences were equated for RMS amplitude.

Procedure

Subjects were run individually in a small sound-treated cubicle that was equipped with a three-button response box and a CRT monitor. Subjects were given a packet containing the MRT instructions and a 50-item MRT test form. Subjects were instructed to circle in their answer booklet the printed form of the word presented over the headphones.

After completing the MRT, subjects were given instructions for the sentence-by-sentence listening time procedure. Instructions were presented both visually one sentence at a time on the CRT monitor in front of the subjects and auditorily over the headphones. At the end of each sentence, a cue light on the response box was illuminated. After the subject pressed a button labeled "Continue," the next sentence from the instructions was presented. Thus, subjects controlled the rate of presentation -- that is, the temporal interval between sentences in each text by pressing the "Continue" button on the response box. Two-hundred-fifty milliseconds after the subject pressed the "Continue" button, the next sentence began. After the instructions, a practice passage and the ten test passages were presented using the same self-paced, sentence-by-sentence format. Sentences from the comprehension passages were presented only over headphones.

At the end of each passage, the same recognition memory sentences used in Experiment 1 were presented to assess comprehension of the information in the passage. However, because subjects were run individually, they had an unlimited amount of time to respond to each sentence.

Latencies between the offset of each sentence and the following button press in the SBSLTT, as well as the comprehension responses, were recorded by a PDP 11/34 minicomputer. Subjects were given a five minute break after the fifth passage. The order of the passages and the recognition sentences was identical to that used in Experiment 1. Each experimental session lasted approximately one hour.

Results and Discussion

The results will be presented separately for the MRT, the SBSLTT and the recognition memory test. Correlations were also computed between the MRT accuracy scores and several different comprehension measures. Voice (natural or synthetic) was treated as a between-subjects variable; text difficulty (fourth-grade or college) and sentence type (word-recognition or proposition- recognition) were treated as within-subjects variables. Newman-Keuls tests were used as post-hoc probes to assess statistical effects revealed by ANOVA's.

Modified Rhyme Test

Performance on the MRT was similar to that observed in the first experiment. Subjects who listened to natural speech averaged 96% correct, while subjects who listened to synthetic speech averaged 63% correct. This difference was highly significant [$t(14)=-17.14$, $p<.001$].

Insert Figure 4 about here

Sentence-by-Sentence Listening Time Task

In order to reduce variability, response latencies greater than or less than two standard deviations from the mean for each voice group were replaced with the mean value for the same group. Using this procedure, 2.1% of the responses from the natural speech condition and 1.7% of the responses from the synthetic speech condition were replaced. Data for each subject were then averaged across the different sentences within a passage as well as across the fourth-grade and college passages. Figure 4 shows the overall sentence-by-sentence listening times.

Sentence listening times were significantly longer for synthetic speech compared to natural speech [$F(1,28)=6.74$, $p=.015$]. Listening times were also longer for the college passages than fourth-grade passages [$F(1,28)=38.93$, $p<.001$]. The interaction between voice and text difficulty was not significant. These results demonstrate the validity of the sentence-by-sentence listening time procedure for measuring on-line processing activities. Not only did the methodology index differences in comprehension between natural and synthetic speech, but it also was sensitive to the text difficulty variable.

Insert Figure 5 about here

Further evidence of these differences is revealed when the sentence listening times were analyzed separately for each passage. Figure 5 displays the sentence listening times for each passage. The left panel shows data for the five fourth-grade passages and the right panel shows data for the five college-level passages. Within each panel is the passage name and the results of t-tests that were conducted for each passage. The differences between natural and synthetic speech were highly significant for four of the five fourth-grade passages. The difference just missed significance for the fifth passage [$t(28)=1.93$, $p=.06$]. The differences between natural and synthetic speech were also significant for four of the five college-level passages. Although not significant, the difference between voices for the fifth passage was in the same direction as in the other passages [$t(28)=1.70$, $p=.10$]. In short, the differences in listening times for natural and synthetic speech were robust across the individual passages used in the experiment.

Insert Figure 6 about here

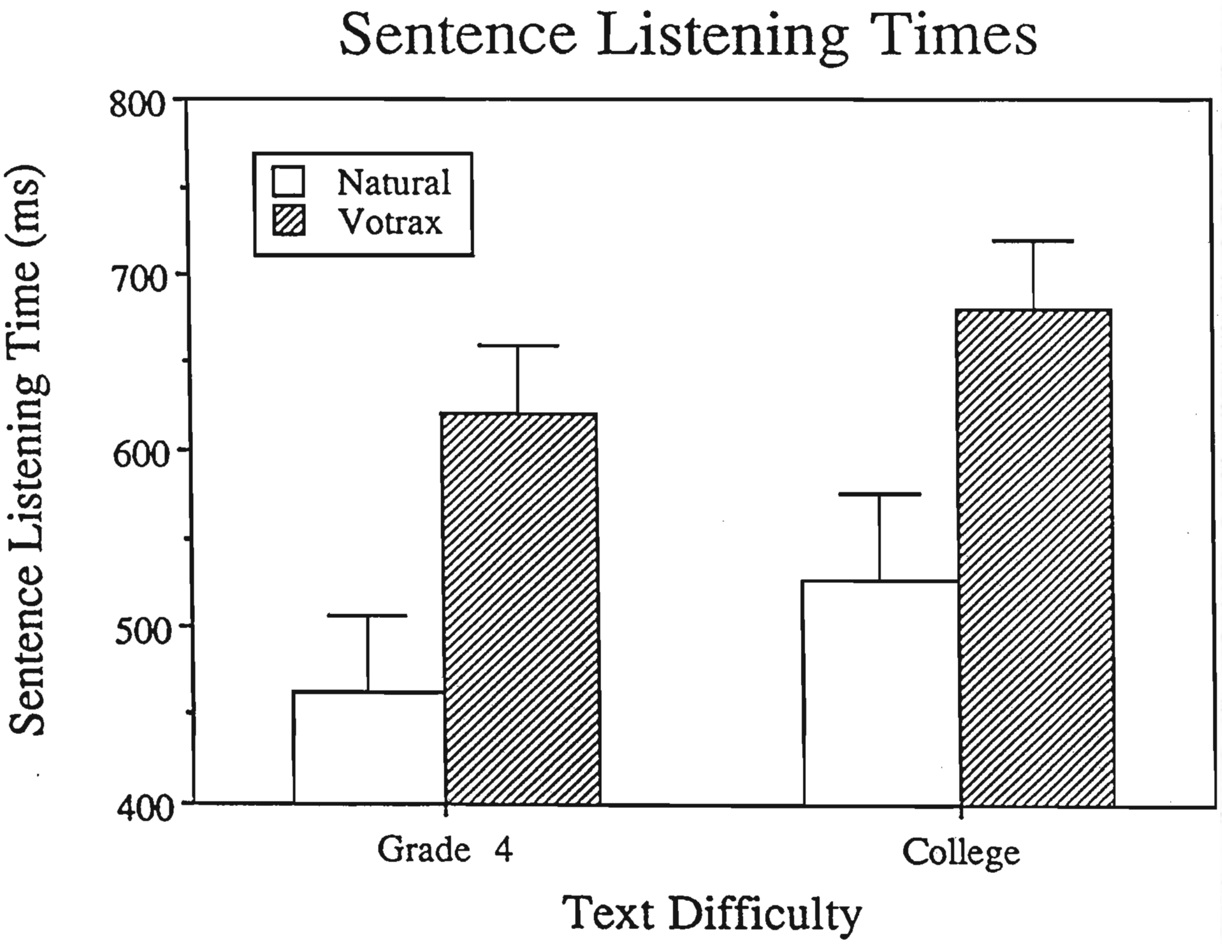


Figure 4. Sentence-by-sentence listening times as a function of voice and text difficulty in Experiment 2. Open bars represent response latencies for passages of natural speech, and striped bars represent response latencies for passages of synthetic speech. Error bars represent one standard error of the sample means.

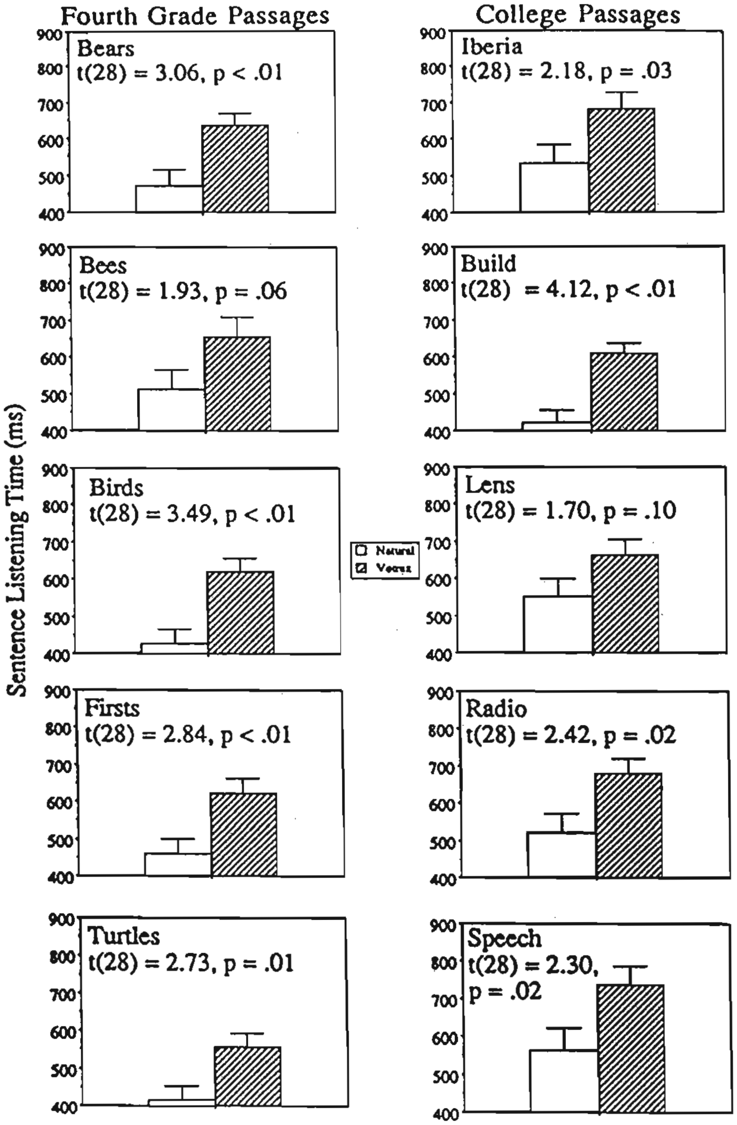


Figure 5. Sentence-by-sentence listening times as a function of voice in Experiment 2. The panel on the left displays response latencies for the five fourth-grade passages, and the panel on the right displays response latencies for five college-level passages. Open bars represent response latencies for passages of natural speech, and striped bars represent response latencies for passages of synthetic speech. Error bars represent one standard error of the sample means.

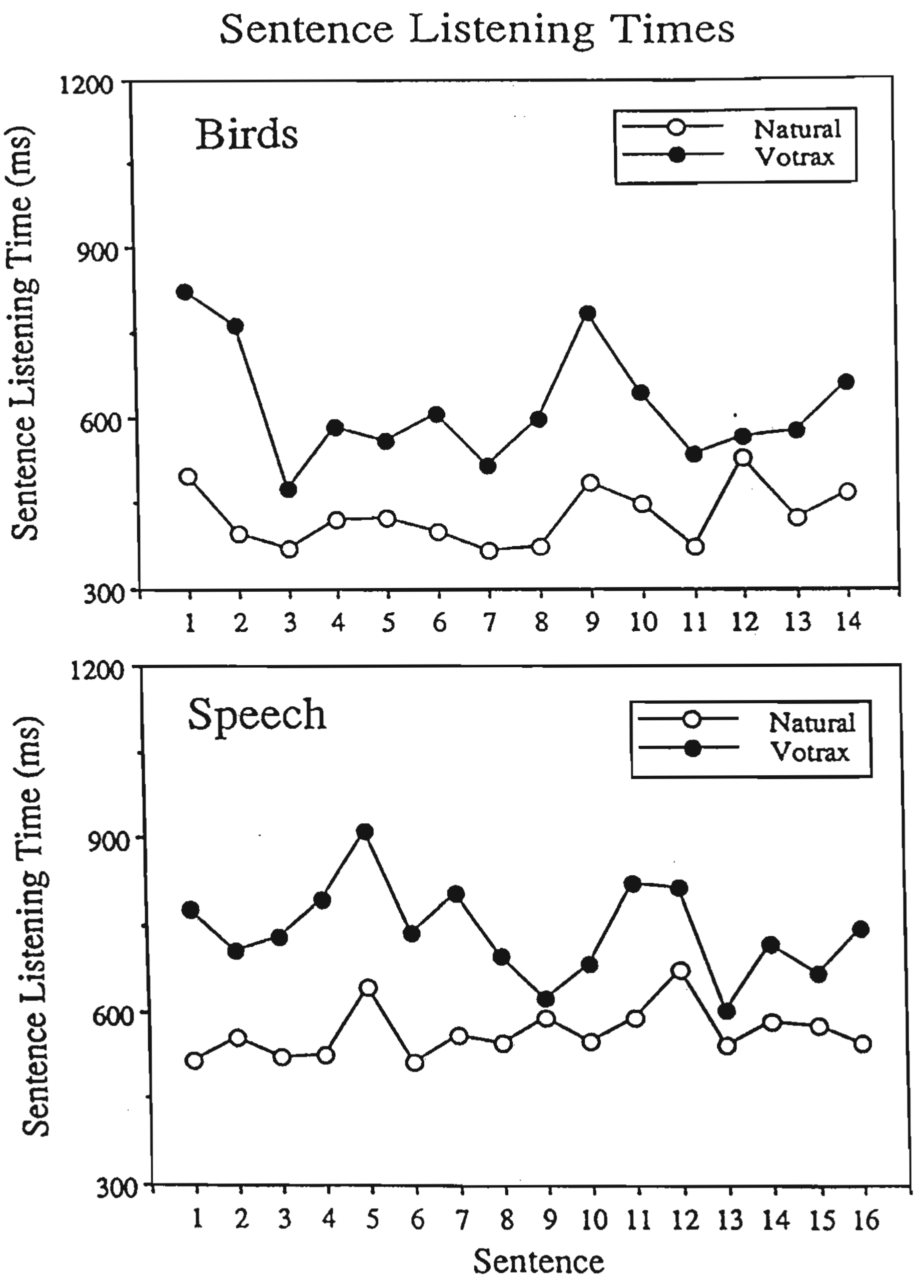


Figure 6. Sentence-by-sentence listening times as a function of the position of a sentence within a passage in Experiment 2. The upper panel displays data for a representative fourth-grade passage, and the lower panel displays data for a representative college-level passage. The open circles and connecting lines represent response latencies for passages of natural speech. The solid circles and connecting lines represent response latencies for passages of synthetic speech.

Figure 6 shows the sentence listening times for two passages as a function of voice and serial position of each sentence within a passage. The upper panel shows the listening times for a fourth-grade passage composed of 14 sentences, and the lower panel shows the listening times for a college-level passage with 16 sentences. These two representative examples reveal that differences in sentence listening times are highly consistent across individual sentences within passages as well as across passages.

In short, the sentence-by-sentence listening time data reveal on-line differences in the speed of comprehension processes. We believe that these processes include components involved in word-recognition, lexical access and parsing for individual sentences and the integration of information across different sentences (Just & Carpenter, 1980; Kintsch & Van Dijk, 1978). Thus, the SBSLTT provides a converging procedure that can be used to index on-line spoken language comprehension processes. Both the word monitoring data from the first experiment and the SBSLTT data from the present experiment suggest that comprehension proceeds more slowly for passages of synthetic speech than passages of natural speech, and that comprehension proceeds more slowly for difficult passages than easy passages. The present results replicate the major findings obtained in Experiment 1 using a word monitoring task.

Insert Figure 7 about here

Recognition Memory

Figure 7 displays recognition accuracy data. The upper panel shows data for the word-recognition sentences and the lower panel shows data for the proposition-recognition sentences. Accuracy was greater for sentences following passages of natural speech compared to passages of synthetic speech [$F(1,28)=27.54$, $p<.001$]. This finding demonstrates not only that real-time comprehension proceeds more slowly for passages of synthetic speech, but that the products of comprehension also appear to be more fragile than natural speech. This result is also consistent with the findings obtained in the first experiment. Although the main effects of sentence type and text difficulty were not significant, both of these factors entered into significant interactions.

Insert Figure 8 about here

As in the first experiment, the interaction between voice and sentence type was significant [$F(1,28)=4.91$, $p=.035$]. This interaction is shown in Figure 8. When comparing accuracy for word- and proposition-recognition sentences, a large increase in performance was observed for sentences following passages of natural speech, but only a small decrease was observed for sentences following passages of synthetic speech. Post-hoc analyses revealed that although the increase from word to proposition-recognition sentences for natural speech was significant, the decrease for synthetic speech was not significant. This effect is similar to the cross-over interaction observed earlier by Luce (1981) and replicates the interaction between voice and sentence type found in Experiment 1. No other interactions involving voice were significant.

Insert Figure 9 about here

In order to specify the nature of the interaction between voice and sentence type more precisely, difference scores were computed on the raw data. These scores were obtained for each subject by subtracting the accuracy for proposition-recognition sentences from the accuracy for the word-recognition sentences. These scores were compared to data obtained in the zero-target condition in Experiment 1. The results of this procedure are presented in Figure 9.

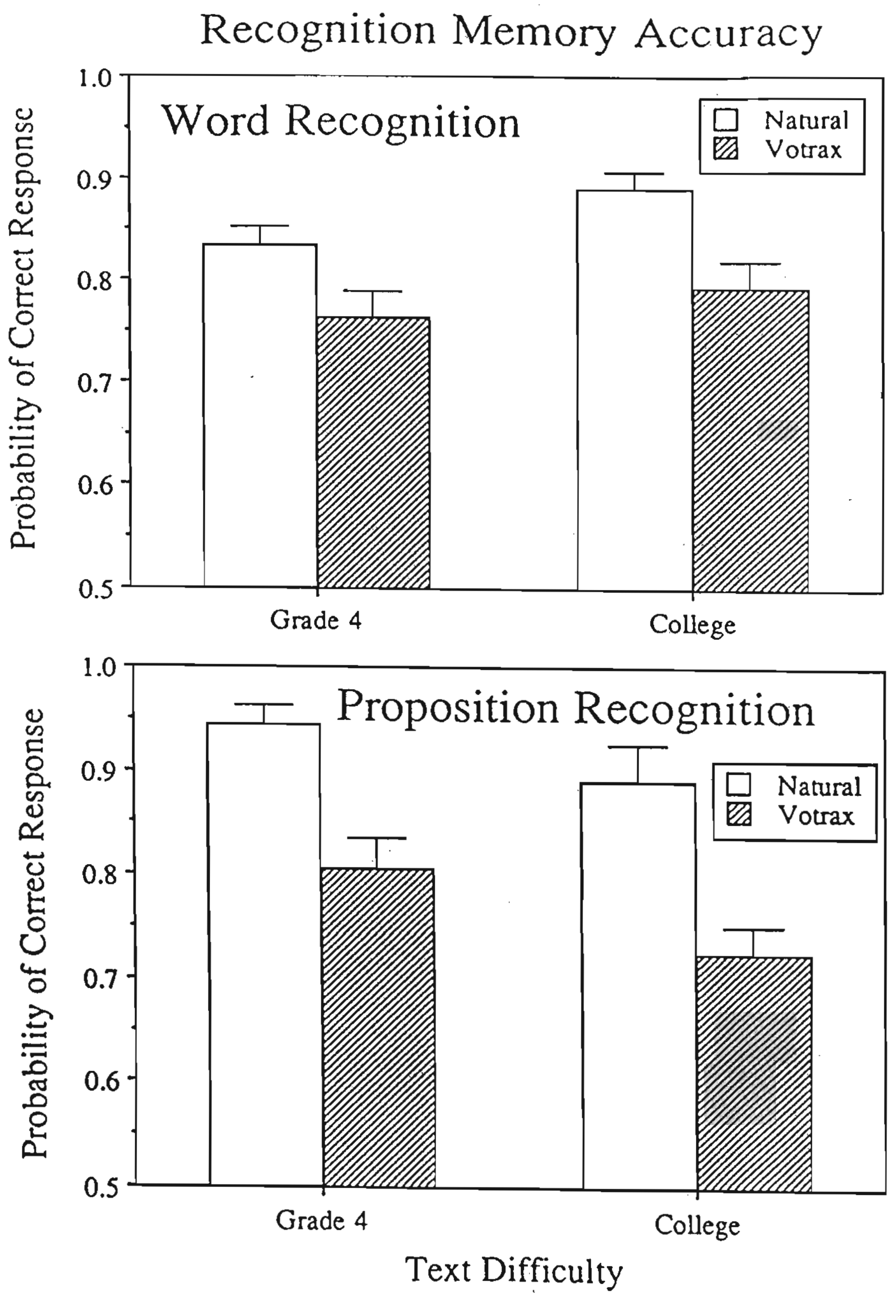


Figure 7. Recognition memory accuracy (probability correct) as a function of text difficulty in Experiment 2. The upper panel shows data for word-recognition sentences, and the lower panel shows data for proposition-recognition sentences. Open bars represent accuracy for sentences after passages of natural speech, and striped bars represent accuracy for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

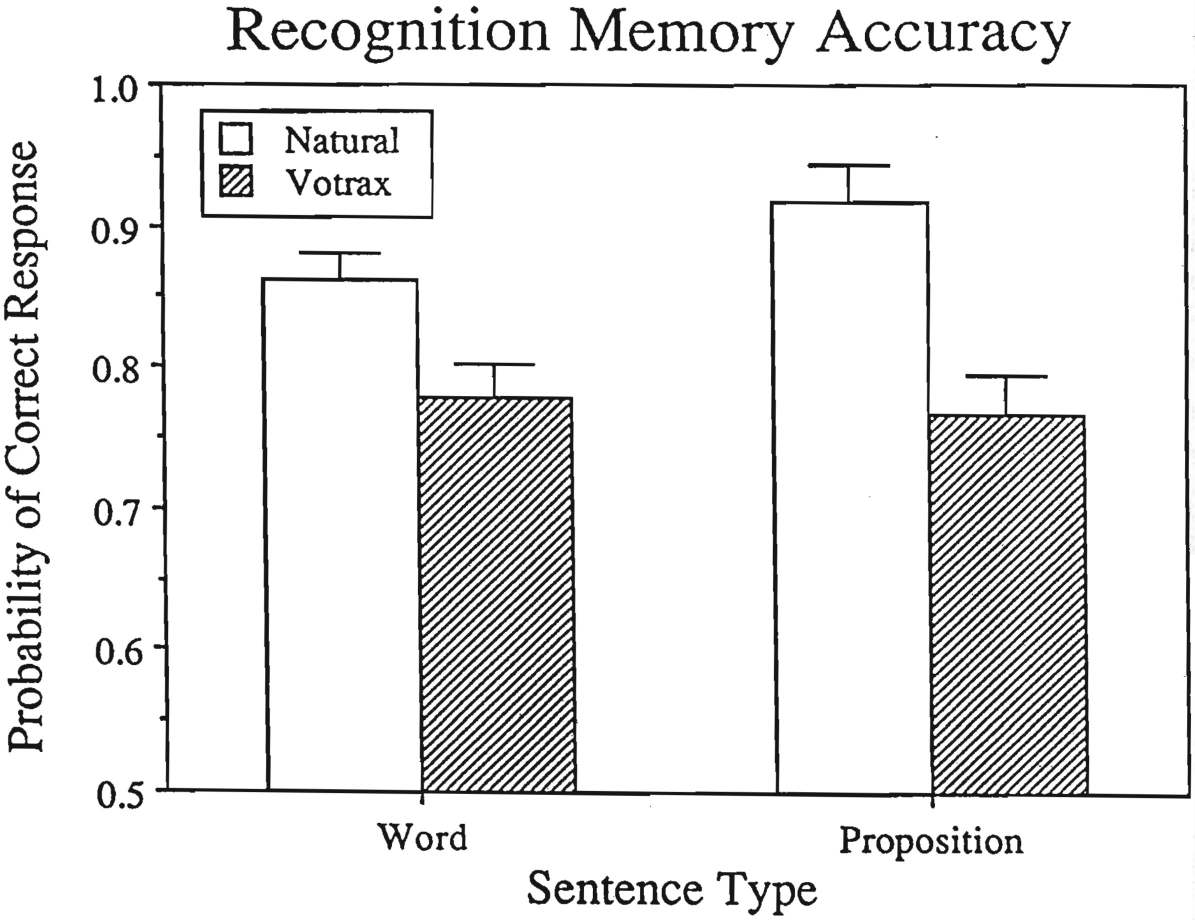


Figure 8. Recognition memory accuracy (probability correct) as a function of sentence type in Experiment 2. Open bars represent accuracy for sentences after passages of natural speech, and striped bars represent accuracy for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

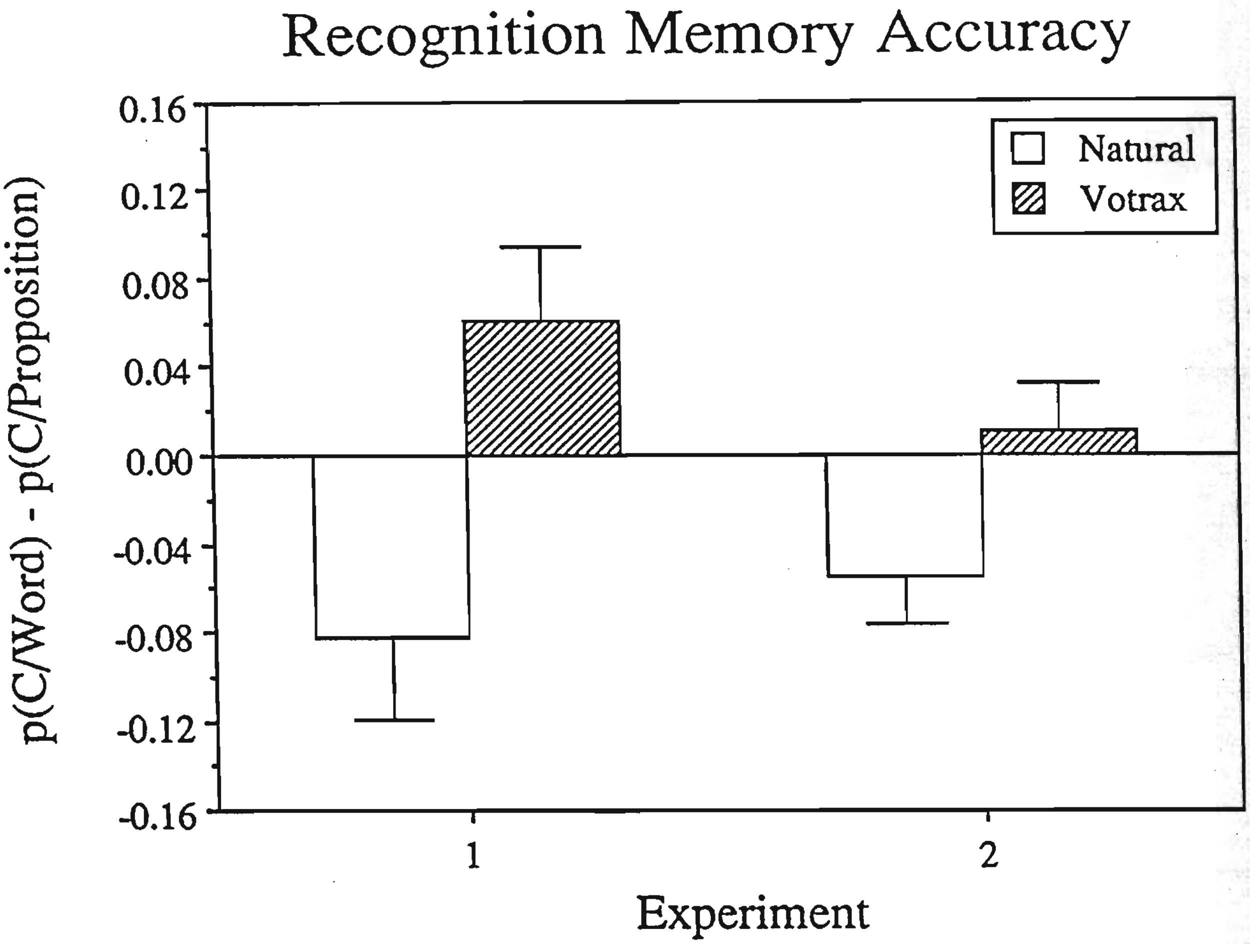


Figure 9. Accuracy difference scores (probability correct) for Experiments 1 and 2. Difference computed by subtracting proposition-recognition accuracy minus for word-recognition accuracy. Open bars represent differences for sentences after passages of natural speech, and striped bars represent differences for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

Both sets of difference scores show clearly that proposition-recognition sentences were recognized more accurately than word-recognition sentences for subjects who listened to passages of natural speech. However, the pattern was reversed for subjects who listened to passages of synthetic speech. These subjects recognized word-recognition sentences more accurately than proposition-recognition sentences. The difference scores were entered into an ANOVA with experiment and voice as between- subjects variables. This analysis revealed a significant main effect of voice [$F(1,59)=12.00, p=.001$], but neither the effect of experiment nor the interaction between voice and experiment was significant.

Assuming that the representations incorporate those stimulus attributes that are attended to during encoding, then the recognition data can be used as evidence for differential resource allocation strategies during the process of comprehension. Several investigators have argued that there are limited attentional resources that are distributed to the various comprehension subcomponents (Haberlandt, 1984; LaBerge & Samuels, 1974; Pisoni, 1982). The recognition memory data from the present experiments suggest that subjects listening to synthetic speech allocated a relatively larger proportion of their resources to perceptual encoding, leaving relatively fewer resources for processing the meaningful aspects of the text (Hersch & Tartarglia, 1983; Luce, 1981; Pisoni, 1982). Consequently, attention to acoustic-phonetic structure produced a relatively robust memory trace for words compared to meaningful information. In contrast, subjects who listened to natural speech appeared to attend to and remember much more propositional information from the same passages.

An analysis was also carried out to compare the recognition accuracy results with the findings obtained in Experiment 1. Except for two effects, the pattern of results in the two experiments were identical. The first difference emerged in the overall level of accuracy across the two experiments. Subjects were more accurate at recognizing sentences in Experiment 2 than Experiment 1 [$F(1,59)=26.22, p<.001$]. This result is not surprising because the SBSLTT encourages subjects to seek optimal levels of comprehension performance. The second difference was revealed in a three-way interaction between experiment, voice and difficulty [$F(1,59)=10.47, p=.002$]. The interaction was due entirely to the lowered performance in Experiment 1 on the college passages for natural speech. The decrease was not present in Experiment 2, again suggesting that subjects attempted to reach optimal levels of performance in the SBSLTT. Despite these differences, the recognition memory results obtained in Experiment 2 closely mirror those observed in Experiment 1. This suggests that the SBSLTT does not adversely affect the final products of comprehension. Indeed, subjects in Experiment 2 appear to have utilized the self-paced nature of the procedure to enhance their overall comprehension performance.[13]

Insert Table 4 about here

Regression Analyses

Table 4 displays the correlations between segmental intelligibility and several of the comprehension measures. Data were collapsed across voice and text difficulty. The left column shows correlation coefficients (Pearson's *r*) and the right column shows the probability values associated with *F*-statistics that were used to test whether the observed coefficients are different from zero. Overall, the correlations of the MRT with the sentence-by-sentence listening times and recognition memory accuracy were moderate and statistically significant. The correlation of the MRT accuracy with sentence listening time was negative, indicating that as segmental intelligibility increased, response latencies decreased.

[13] In fact, the increased verification accuracy obtained in Experiment 2 may be due in part to a speed-accuracy trade-off introduced by the unlimited response interval in the second experiment (Reed, 1976).

Table 4

Correlation of MRT accuracy and comprehension measures in Experiment 2.

Predicted Variable	Correlation	Probability
Sentence Listening Time		
Response Latency	-0.40	0.03
Recognition Accuracy		
Word Recognition	0.55	<0.01
Proposition Recognition	0.63	<0.01

In contrast, the correlations of MRT accuracy with recognition accuracy were positive, indicating that as segmental intelligibility increased, recognition accuracy increased. Although the results show a moderate and reliable relationship between segmental intelligibility and comprehension, the magnitude of the correlation coefficients [$r=.405$ to .631] suggests that comprehension involves more than just phoneme and word recognition. Not surprisingly, some proportion of the variance in comprehension is produced by processes that go beyond the acoustic-phonetic information present in the speech waveform.

Summary

The results of Experiment 2 corroborate and extend the major findings obtained in Experiment 1. First, the results demonstrate that the SBSLTT is a useful tool to study spoken language comprehension, particularly on-line processes employed during the initial stages of comprehension. Although the paradigm has been used profitably for several years with printed prose (Graesser, Hoffman, & Clark, 1980; Miller & McKean, 1964; Kieras, 1974; Cirilo & Foss, 1980), it has never been used to study spoken language comprehension. The finding that the observed latency measures varied with text difficulty demonstrates that the measurement technique is sensitive to several components of the comprehension process (see Haberlandt, 1984, for comparable arguments for the visual analog of the paradigm).

The SBSLTT data also demonstrate that comprehension proceeds more slowly when the sentences are produced by a speech synthesizer compared to sentences produced by a natural talker. This result is consistent with the results of the word monitoring data, in which detection latencies were reliably slower for words in passages of synthetic speech compared to natural speech. Not only was the effect of voice present in the data averaged across sentences and passages, but more detailed analyses revealed that the differences between natural and synthetic speech were present at the individual passage and individual sentence level as well.

The recognition memory data largely replicate results obtained in the word monitoring experiment. That is, accuracy was poorer for sentences presented after passages of synthetic speech compared to natural speech. The results were also compared to results obtained from the zero target condition of the first experiment in which comprehension was unconstrained by task demands. The analyses revealed very similar patterns of results. The major exception was that recognition accuracy was higher in Experiment 2 compared to Experiment 1.

Finally, the interaction between voice and sentence type was also observed in the recognition data. Subjects who listened to passages of synthetic speech performed better on the word-recognition sentences than the proposition-recognition sentences, whereas subjects listening to passages of natural speech were more accurate for proposition-recognition sentences than for word-recognition sentences. The same effect was present in the zero target set size condition in Experiment 1.

GENERAL CONCLUSIONS

The results obtained in the present set of experiments are similar to findings reported in previous comprehension studies. In experiments by Luce (1981), Hersch and Tartarglia (1983), and Moody and Joost (1986), reliable differences in several comprehension measures were observed between natural and synthetic speech using post-perceptual tests. The present findings suggest that passages of synthetic speech do indeed provide a more degraded representation of textual information.

As noted earlier, other studies have found relatively small differences in comprehension between natural and synthetic speech when post-perceptual measures were used (Jenkins & Franklin, 1981;

McHugh, 1976; Pisoni & Hunnicutt, 1980). We believe these studies may have utilized test procedures that were simply too insensitive to measure differences in comprehension. For example, in one of their experiments, Jenkins and Franklin (1981) used a free recall task that encourages subjects to adopt a wide variety of post-perceptual strategies. In the studies conducted by McHugh (1976) and Pisoni and Hunnicutt (1980), the multiple-choice test questions may have been so easy that subjects could have responded to them correctly without even listening to the passages. The results of the present investigation show clearly that measures of performance that rely entirely on memory for the products of comprehension are not able to assess differences in the speed and efficiency of the processing operations that are used to generate these representations. Once a representation has been created in long-term memory, it may not contain very detailed information about earlier stages of processing. Thus, it is not surprising that post-perceptual measures are unable to measure differences between natural speech and relatively poor quality synthetic speech such as Votrax.

The two on-line measures utilized in the present experiments are largely complementary. Word monitoring performance measures the speed of comprehension within the bounds of individual sentences, whereas sentence-by-sentence listening times provide a composite measure of the comprehension of entire sentences. In the present studies, both procedures have demonstrated that long passages of synthetic speech are harder to encode and comprehend than natural speech.

In summary, several conclusions can be drawn from the results obtained in the present investigations. First, comprehension of synthetic speech is poorer than natural speech. Second, the poor performance with synthetic speech is due, at least in part, to the increased encoding demands of synthetic speech that leave fewer resources available for comprehension. Third, the on-line word monitoring task and the sentence listening time task appear to be sensitive measures that can be successfully employed to study the real-time comprehension of spoken language. Fourth, comprehension accuracy for passages of connected speech can be predicted only moderately well from standardized tests of segmental intelligibility for isolated words. Finally, the present results suggest that low-quality synthetic speech may be inappropriate for applications in which there is a need for either a high degree of understanding or rapid responding to the semantic content of messages. This conclusion is particularly relevant when the messages may be drawn from a potentially infinite number of novel utterances selected from semantically complex knowledge domains or when the listener is under high cognitive load in a resource-demanding environment.

References

Allen, J., Hunnicutt, S., & Klatt, D.H., (1987). *From Text to Speech: The MITalk System*. Cambridge, UK: Cambridge University Press.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In Bower, G.H.(Ed.), *The Psychology of Learning and Motivation, Vol. 8*. New York: Academic Press.

Bartlett, F.C. (1932). *Remembering*. Cambridge: Cambridge University Press.

Britton, B.K., Holdridge, T., Curry, C., & Westbrook, R.D. (1979). Use of cognitive capacity in reading identical texts with different amounts of discourse level meaning. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 262-270.

Bruner, H., & Pisoni, D.B. (1982). Some effects of perceptual load on spoken text comprehension. *Journal of Verbal Learning and Verbal Behavior*, **21**, 186-195.

Cirilo R.K. (1981). Referential coherence and text structure in story comprehension. *Journal of Verbal Learning and Verbal Behavior*, **20**, 358-367.

Cirilo, R.K., & Foss, D.J. (1980). Text structure and reading time for sentences. *Journal of Verbal Learning and Verbal Behavior*, **19**, 96-109.

Cooperative English Tests: Reading Comprehension (1960). Form 1B. Princeton, NJ: Educational Testing Service.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55-60.

Farr, R., & Carey, R.F. (1986). *Reading: What Can Be Measured*. 2nd Ed. Newark, DE: International Reading Association.

Flesch, R. (1949). *The Art of Readable Writing*. New York: Harper.

Foss, D.J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, **8**, 457-462.

Graesser, A.C., Hoffman, N.L., & Clark, L.F. (1980). Structural components in reading times. *Journal of Verbal Learning and Verbal Behavior*, **19**, 135-151.

Gough, P.B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, **4**, 107-111.

Gunning, R. (1968). *The Technique of Clear Writing*. New York: McGraw-Hill.

Haberlandt, K. (1984). Components of sentence and reading times. In Kieras, D.E. & Just, M.A. (Eds.), *New Methods in Reading Comprehension Research*. Hillsdale, NJ: Erlbaum.

Haviland, S.E., & Clark, H.H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, **13**, 512-521.

Hersch, H.M., & Tartarglia, L. (1983). Understanding synthetic speech. *User Research Group, Corporate Research and Architecture*. Maynard, MA: Digital Equipment Corporation.

Hoover, J., Reichle, J., van Tasell, D., & Cole, D. (1987). The intelligibility of synthesized speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*, **30**, 425-431.

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

Iowa Silent Reading Tests (1972). Level 3. Form E. New York: Harcourt Brace Jovanovich.

Jenkins, J.J., & Franklin, L.D. (1981). Recall of passages of synthetic speech. Paper presented at the 21st Psychonomics Society Meeting. Philadelphia, PA.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, **87**, 329-354.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs: Prentice-Hall.

Kieras, D.E. (1974). Analysis of the effect of word properties and limited reading time in a sentence comprehension and verification task. *Doctoral dissertation, University of Michigan.*

Kintsch, W., & Keenan, J.M. (1973). Reading rate as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, **5**, 257-274.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363-394.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In Nilsson, L.G. (Ed.), *Perspectives on Memory Research*. Hillsdale, NJ: Erlbaum.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, **82**, 737-793.

LaBerge, D., & Samuels, S.L. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, **6**, 293-323.

Larkey, L.S., & Danly, M. (1983). Fundamental frequency and sentence comprehension. *MIT Speech Group Working Papers, Vol II.* Cambridge, MA: Research Laboratory of Electronics, Massachusetts Institute of Technology.

Lesgold, A.M., Roth, S.F., & Curtis, M.E. (1979). Foregrounding effects in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior*, **18**, 291-308.

Levelt, W.J.M. (1978). A survey of studies in sentence perception. In Levelt, W.J.M. & Flores d'Arcais, G.B. (Eds.), *Studies in the Perception of Language*, New York: Wiley.

Logan, J.S., Greene, B.G., & Pisoni, D.B. (1989). Intelligibility of eight text-to-speech systems. *Journal of the Acoustical Society of America*, **86**, 566-581.

Luce, P.A. (1981). Comprehension of fluent synthetic speech produced by rule. *Research on Speech Perception Progress Report No. 7*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25**, 17-32.

McCauley, M.E. (1984). Human factors in voice technology. In Muckler, F.A. (Ed.), *Human Factors Review: 1984*. Santa Monica, CA: The Human Factors Society, 131-166.

McHugh, A. (1976). Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. *Naval Research Laboratory Report 8015*. Washington, D.C.: Naval Research Laboratory.

Manous, L.M., Pisoni, D.B., Dedina, M.J., & Nusbaum, H.C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. *Research on Speech Perception Progress Report No. 11*. Bloomington IN: Speech Research Laboratory, Indiana University.

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

Miller, G.A., & McKean, K.O. (1964). A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology*, **16**, 297-308.

Moody, T.S., & Joost, M.G. (1986). Synthesized speech, digitized speech and recorded speech: A comparison of listener comprehension rates. *Proceedings of the Voice Input/Output Society*. Alexandria, VA.

The Nelson-Denny Reading Test (1973). Form D. Boston: Houghton-Mifflin.

Nye, P.W., & Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (Modified Rhyme Test results). *Haskins Laboratory Status Report on Speech Research, SR-33*. New Haven, CT: Haskins Laboratories.

Pauk, W. (1983). Six way paragraphs. Middle Level. Providence, RI: Jamestown.

Pisoni, D.B. (1982). Perception of speech: The human listener as a cognitive interface. *Speech Technology*, **1**, 10-23.

Pisoni, D.B., & Dedina, M.J. (1986). Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report. *Research on Speech Perception Progress Report No. 12*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D.B., & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. *IEEE Conference Record on Acoustics, Speech, and Signal Processing.* New York: IEEE Press, 572-575.

Pisoni, D.B., Manous, L.M., & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on sentence verification of sentences controlled for intelligibility. *Computer Speech and Language*, **2**, 303-320.

Pisoni, D.B., Nusbaum, H.C., & Greene, B.G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **73**(11), 1665-1676.

Posner, M.I., & Boies, S.J. (1971). Components of attention. *Psychological Review*, **78**, 391-408.

Reed, A.V. (1976). List length and the time-course of recognition in immediate memory. *Memory and Cognition*, **4**, 16-30.

Schmidt-Nielson, A., & Kallman, H.J. (1987). Evaluating the performance of the LPC 2.4 kbps processor with bit errors using a sentence verification task. *NRL Report No. 9089*. Washington, D.C.: Naval Research Laboratory.

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.

Stanford Test of Academic Skills: Reading (1972). College Level II-A. New York: Harcourt Bruce Jovanovich.

Wickens, C.D. (1987). Information processing, decision making, and cognition. In Salvendy, G. (Ed.), *Handbook of Human Factors*. New York: Wiley Interscience.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Measuring the Workload of Comprehending Spoken Discourse: A First Report[1]

James V. Ralston, Scott E. Lively,
David B. Pisoni and Susan M. Rivera

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

[1] This research was supported NSF Research Grant #IRI-86-17847 to Indiana University.

Abstract

A recent study utilized an online word monitoring task to measure cognitive workload during the comprehension of passages of natural speech and synthetic speech. However, potential confounds in that experiment as well as potential limitations of the word monitoring task itself suggest the need for a non-linguistic secondary task. Therefore, an experiment employing a click monitoring task combined with a recognition memory test was conducted to measure cognitive workload during the comprehension of connected spoken discourse. Three main variables were manipulated: voice (natural or synthetic speech), difficulty of the text (fourth grade or college level), and secondary task during comprehension (click monitoring or no monitoring). Overall, monitoring latencies were 100 milliseconds faster for the natural speech compared to synthetic speech. These data suggest that the workload associated with the comprehension of synthetic speech is greater than for natural speech. Although not statistically significant, other trends in the monitoring data were also in the direction predicted by a limited capacity account of processing resources. The finding that monitoring had no effect on recognition memory performance suggests that the introduction of the monitoring task does not significantly influence subjects' resource allocation strategies. The combined results suggest that with further refinements, the click monitoring task may be a useful experimental tool for investigating the comprehension of spoken discourse.

Measuring the Workload of Comprehending Spoken Discourse: A First Report

Recently, attempts have been made to measure the workload associated with the comprehension of passages of written or spoken discourse (Britton, Westbrook, & Holdridge, 1978; Brunner & Pisoni, 1982; Inhoff & Fleming, 1989; Ralston, Pisoni, Lively, Greene, & Mullennix, 1991). Most of these experiments have required subjects to perform some secondary monitoring task simultaneous with the primary comprehension task. Experimenters have assumed that performance on the secondary task reflects the processing demands of the primary task.[2] For example, Ralston et al. (1991) required subjects to detect target words embedded in spoken texts produced by an adult male talker or a speech synthesizer. The results showed that: a) monitoring latencies were longer for targets imbedded in passages of synthetic speech compared to natural speech, b) monitoring latencies were longer for college-level passages compared to fourth grade passages, and c) there was an interaction between the voice and difficulty of the text. The results suggest greater workload demands associated with synthetic speech and with difficult texts. The mechanisms affected by these experimental manipulations (encoding and other comprehension processes, respectively) appear to place greater demands on the same limited capacity of processing resources.

However, several characteristics of the word monitoring task as well as the specific implementation of this task suggest alternative interpretations of the word monitoring results. The stimulus materials used by Ralston et al. (1991) were constructed without controlling speaking rate. Thus, the synthetic speech was slower than the natural speech. It is possible then, that the observed difference in monitoring latencies may have simply reflected differences in target word durations. A separate problem of word monitoring is the selection of word targets. Because the nature of potential word targets (e.g., length, frequency of occurrence, etc.) typically varies with textual difficulty, controlling word target properties while varying text difficulty may be cumbersome, if not impossible. A final, more theoretical problem of the word monitoring technique is that it may also be sensitive to lexical processes it is supposed to measure. Although lexical processes may be influenced by other linguistic mechanisms, one might reasonably desire a more unbiased index of cognitive workload during comprehension.

One promising technique for measuring workload during comprehension involves the monitoring of nonlinguistic events, such as clicks or lights. This technique shares with other monitoring tasks the theoretical assumption of a single, limited capacity resource pool. Click monitoring has been used previously in several experiments examining spoken sentence and discourse processing. For example, Abrams and Bever (1969) found that for unfamiliar spoken sentences, monitoring latencies were slowest for clicks at the end of clauses and fastest for clicks between and immediately after clause boundaries. They suggested that their findings demonstrate the psychological validity of phrase structures. Britton and his colleagues (Britton, Westbrook, & Holdridge, 1978; Britton, Holdridge, Curry, & Westbrook, 1979; Britton, Meyer, Simpson, Holdridge, & Curry, 1979) have used click monitoring as an index of workload during the comprehension of discourse-length written texts. Surprisingly, Britton et al. (1978) found that monitoring latency was slower for college level texts than for primary school texts. The authors argued that more difficult texts lead to more frequent "breakdowns" in comprehension, leaving more attention for the click monitoring task. In contrast, Inhoff and Fleming (1989) have recently employed visual and auditory monitoring tasks with a self-paced reading task. They found that

[2] For a description of these tasks and their underlying assumptions, see Posner and Boies (1971).

monitoring latencies reliably increase with increasing text difficulty. Inhoff and Fleming suggested that Britton et al. (1978) may have presented probes while subjects were rereading texts, an activity that is known to be more common with difficult texts. Inhoff and Fleming also reported data suggesting that the first reading of text requires more processing resources than subsequent rereading of the same text.

The present experiment was conducted to determine whether the click monitoring technique can be used to measure workload differences during comprehension as a function of the voice that produces the text (natural vs. synthetic) and the level of difficulty of the text (fourth grade vs. college). The results would provide converging support for the conclusions of the Ralston et al. (1991). Based on our earlier results, we expected that click monitoring performance would be poorer for synthetic than natural speech and poorer for more difficult passages. Finally, control conditions were included in which subjects ignored clicks and simply listened to the passages for comprehension. These conditions allowed an assessment of the extent to which click monitoring interferes with the primary comprehension task. By comparing recognition memory performance under these two conditions, we hoped to determine whether monitoring differences are due to difficulty or whether they are due to changes in subjects allocation strategies (Inhoff & Fleming, 1989).

Methods

Subjects

Subjects were 57 volunteers enrolled in introductory psychology classes at Indiana University in Bloomington. They received course credit for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing

Stimulus Materials

The test passages and recognition memory items (one practice and ten test passages) were the same as those in Ralston et al. (1991). Five of the test passages and the practice passage were intended for fourth grade readers and the remaining five passages were intended for college readers. The mean number of words for the fourth grade passages was 230.8 (ranging from 157 to 298), and the mean number of words for the college passages was 244.6 (ranging from 198 to 326). The passages were written in narrative or expository styles and generally concerned scientific or cultural topics.

Recognition Memory Test

Eight sentences were selected for each passage to test comprehension of the preceding material. These sentences were presented visually on a CRT monitor immediately following the presentation of each passage. The memory sentences ranged from five to ten words in length and were stated in a declarative form. Four of the sentences assessed listeners' memory for specific words that occurred in the passage (referred to here as "word-recognition" sentences), and four of the sentences assessed memory for propositions of the passage (referred to here as "proposition-recognition" sentences). Based on information in the preceding passage, half of the sentences were true and half were false. Word-recognition sentences all had the same frame, "The word XXX appeared in the passage." None of the target words in the word-recognition sentences appeared as word monitoring targets in the same passage. True proposition-recognition sentences described relationships between concepts that occurred in the passage. False proposition-recognition sentences contained concepts and relationships from the passage that were recombined in semantically meaningful ways.

Clicks

The clicks used in the present study were also used in the previous Ralston et al. (1991) experiment to signal to the onset of target words. The target words were monosyllabic nouns with syllable-initial consonants. The mean number of clicks per passage was 9.0 for fourth grade passages (ranging from 8 to 10) and 9.8 for college passages (ranging from 8 to 13).

Recording and Playback Techniques

A male native speaker of American English with a midwestern accent read the eleven passages. The materials were recorded in an IAC sound-attenuated booth with an Electro-Voice D054 microphone. An identical set of passages was produced by a Votrax Type-N-Talk speech synthesizer controlled by a VAX 11/750 computer. No special stress or pronunciation corrections were applied to the synthetic speech. Both natural and synthetic speech versions of each passage were recorded onto one channel of an audio tape at 15 ips using an Ampex AG500 tape recorder.

All stimuli were first low-pass filtered at 4.8 kHz, and then sampled and digitized using a 12 bit A/D converter running at 10 kHz. With a digital waveform editing program, flags were set at the onset of each target word using the upper four bits of each sixteen-bit computer word (Luce & Carrell, 1981). All stimuli were equated for RMS amplitude. The speech stimuli were then recorded back onto one channel of an audio tape using a D/A converter. At this time, the flags in the upper bits were used to trigger a tone burst generator that output a timing tone (16 cycles of a 1000 Hz tone) onto the second channel of the audio tape. Each audio tape contained a calibration vowel and 11 comprehension passages which were produced by either a natural speaker or speech synthesizer. Speech signals were amplified to 80 dB SPL and mixed with 55 dB broadband noise to mask tape hiss and ambient room noise. The timing tones on the second audio channel were used for two purposes: they were amplified and presented in one ear of the subject as targets for the monitoring task, and they were also sensed by the experimental computer and later used as temporal references when computing response latencies. The acoustic stimuli were presented dichotically through TDH-39 matched and calibrated headphones. The speech signals were presented to the right ear of subjects, the clicks were presented to the left ear.

Procedure

Subjects were run in groups of five or fewer in a quiet room used for speech perception experiments. Each subject sat at a small desk-cubicle equipped with a set of headphones, a video display monitor (GBC Standard CRT Model MV-10A), and a two-button response box that was interfaced to the computer. Each subject was given a booklet containing instructions for the comprehension task and a post-experimental questionnaire that was designed to record subjective reactions to the experimental stimuli and tasks. Subjects then read the instructions and reviewed them with the research assistant. One group of subjects (n=28) was told that they would be required to simultaneously detect all click targets and comprehend the passage and that they would be required to respond to a series of test sentences at the end of each passage. Thirteen of these subjects listened to passages of natural speech and 15 listened to passages of synthetic speech. A second group of subjects (n=29) was told that clicks would be presented in one ear, but to ignore them. Fourteen of these subjects listened to passages of natural speech and 15 listened to passages of synthetic speech. All subjects were required to comprehend the passages and respond to the test sentences after each passage. The first passage served as practice to acquaint the subjects with the experimental procedures and the quality of the voice used in each condition.

Each trial of the comprehension test consisted of two parts. In the first part, a passage was presented over headphones and subjects in the monitoring conditions pressed a button on the response box each time they detected a click. At the end of each passage, eight sentences were presented visually on the CRT monitor for four seconds each. Subjects were instructed to determine whether the sentence was "true" or "false" and to record their response by pressing one of two buttons on the response box while the sentences were presented. No feedback was provided to the subjects about the accuracy of their responses. During the recognition memory task, subjects used the right button for a 'true' response and the left button for a 'false' response. Subjects received a five minute break after the fifth passage. The experiment was controlled in real-time by a PDP 11/34 minicomputer which presented all the test sentences and recorded subjects' responses. An experimental session lasted approximately one hour.

Results

Monitoring hits were defined as detection responses with latencies from 100 to 1200 milliseconds in duration. Responses with shorter or longer latencies were assumed to reflect anticipation errors or inattention and were treated as false alarms. Monitoring and recognition memory data were analyzed with Analysis of Variance (ANOVA). Post-hoc pair-wise comparisons were made with Neuman-Keuls tests. In the following sections, statistical significance refers to differences that are less that 5% likely to occur by chance assuming a single underlying population distribution.

Click Monitoring Accuracy

Figure 1 displays the averaged monitoring accuracy data. Monitoring accuracy was computed as the number of hits divided by the total number of clicks for each passage. Overall accuracy on the monitoring task was near ceiling, with an average of 95% correct. A difference was present in the expected direction between the means for passages of natural (97% correct) and synthetic speech (94% correct). Similarly, there was a difference in the expected direction between the means for fourth grade (96% correct) and college-level passages (95% correct). However, neither of these main effects nor their interactions with other variables was significant.

Insert Figure 1 about here

Click Monitoring Latency

Figure 2 displays the means for the latency data obtained in the monitoring task. On the average, response latency was 574 milliseconds (ms). This value is much less than the averaged word monitoring latency (765 ms) observed previously by Ralston et al. (1991), and suggests that the click monitoring task may be easier than the word monitoring task. Click monitoring latencies were nearly 100 ms faster for passages of natural speech compared to synthetic speech, a statistically significant difference [$F(1,26)=6.02$, $p=.02$]. This difference suggests that the workload associated with comprehending of natural speech is less than that for synthetic speech. Although the monitoring latencies were faster for the fourth grade passages (569 ms) than for the college-level passages (580 ms), the difference was not significant [$F(1,26)=1.86$, $p=.18$]. We address the lack of an effect of text difficulty in the discussion section below.

Insert Figure 2 about here

Recognition Memory Accuracy

Figure 3 displays averaged data from the recognition memory task. The data from all 57 subjects was submitted to an ANOVA with voice (natural and synthetic) and task (monitoring and no monitoring) as between-subjects variables and sentence type (word recognition and proposition recognition) and text difficulty as within-subjects variables. Overall, subjects responses were 68% correct. Mean accuracy was higher following passages of natural speech (75% correct) than synthetic speech (63% correct). This difference was statistically significant [$F(1,53)=33.39$, $p<.0001$]. The result is consistent with those from several recent studies using the same stimulus materials. Accuracy was slightly higher for subjects who did not monitor for clicks (68% correct) than for those subjects that did monitor (66%

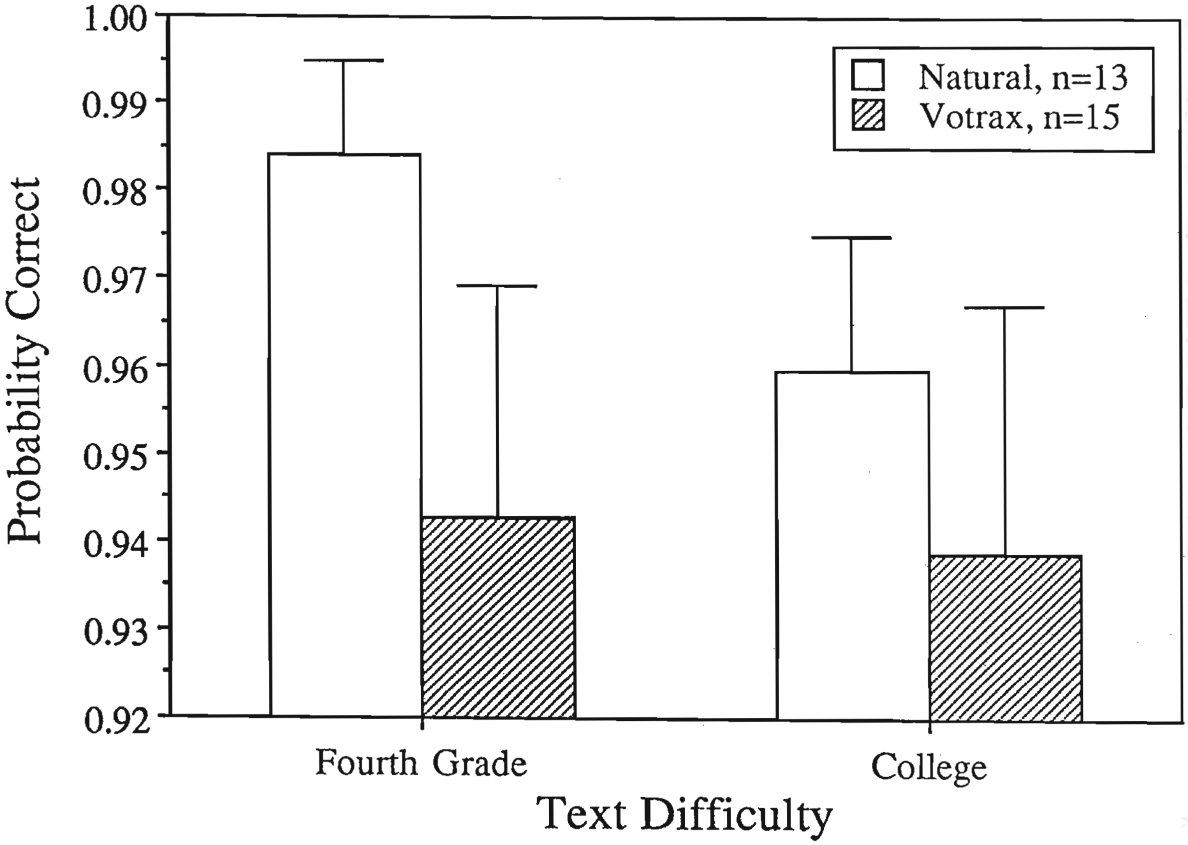


Figure 1. Click monitoring accuracy (probability of a hit) as a function of text difficulty. Open bars represent accuracy for passages of natural speech and striped bars represent accuracy for passages of Votrax synthetic speech. Error bars represent one standard error of the sample means.

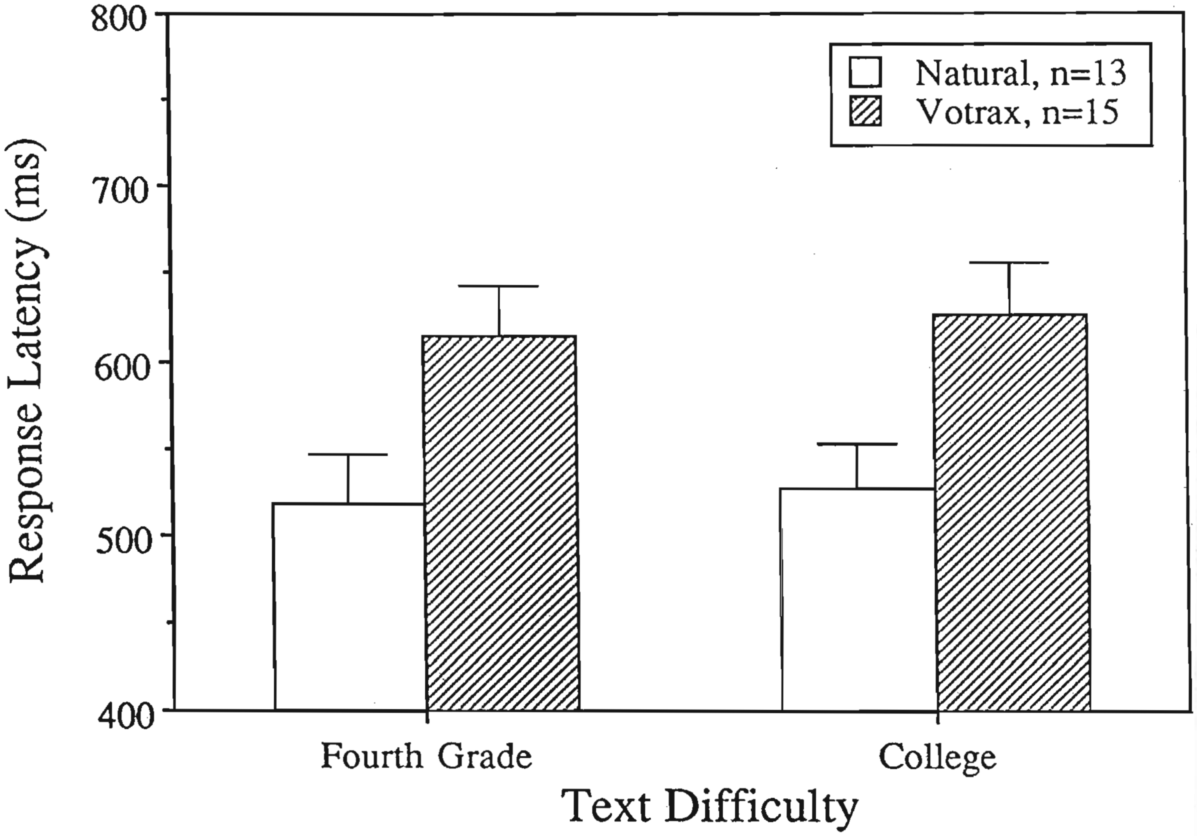


Figure 2. Click monitoring response latency (in milliseconds) as a function of text difficulty. Open bars represent latencies for passages of natural speech and striped bars represent latencies for passages of Votrax synthetic speech. Error bars represent one standard error of the sample means.

correct). However, this difference was not significant [$F(1,53)=1.47$, $p=.23$]. Mean accuracy was higher for word recognition sentences (70%) than for proposition recognition sentences (65%), a difference that was statistically significant [$F(1,53)=7.64$, $p=.0078$]. This result replicates that obtained in other recent experiments using the same stimulus materials. The interaction between sentence type and voice was significant [$F(1,53)=10.87$, $p=.0017$]. Post-hoc tests revealed that while there was no statistical difference as a function of sentence types for sentences following passages of natural speech, accuracy was poorer for proposition recognition sentences than for word recognition sentences following passages of synthetic speech. The form of this interaction also replicates findings observed in our other recent experiments using the word monitoring task.

Insert Figure 3 about here

Mean accuracy was higher for fourth grade passages (72% correct) than for college-level passages (64% correct), and this difference was statistically significant [$F(1,53)=28.01$, $p<.0001$]. Although it is tempting to conclude that this effect reflects the difference in the difficulty of the texts, it could also reflect differences in baseline difficulty of the test items. The factor of text difficulty also entered into significant interactions with sentence type [$F(1,53)=34.79$, $p<.001$] and voice [$F(1,53)=9.84$, $p=.0028$]. The interaction of text difficulty and sentence type replicates results from our earlier experiments. It is also of less interest for the same reason that the main effect of text difficulty is of little interest -- because the four cells represented in the interaction (fourth grade/word recognition, fourth grade/proposition recognition, etc.) are based on four sets of different recognition memory test items, it is possible that the interaction reflects differences in the baseline levels of performance between the different groups of test items, and may not reflect true differences during comprehension processing.

Figure 4 displays the interaction between voice and text difficulty. This interaction has never been observed in the recognition memory data in our previous studies. Post-hoc tests revealed that while accuracy was lower for the college level passages than the fourth grade passages for the natural speech conditions, there was no difference as a function of text difficulty for the synthetic speech conditions. This unexpected result, which seems to contradict predictions based on a limited capacity of processing resources, may simply reflect a near-chance performance on the memory task for the synthetic speech conditions.

Insert Figure 4 about here

Summary and Conclusions

The results from the present experiment suggest that, with further refinements, the click monitoring procedure may be a useful measure of workload during the comprehension of spoken discourse. Monitoring latencies were nearly a hundred milliseconds faster for passages of natural speech compared to synthetic speech. Following Ralston et al. (1991), we conclude that the comprehension of discourse produced by a poor-quality speech synthesizer produces a greater cognitive workload than natural speech. Although the latency differences between fourth grade and college-level text were not statistically significant as assessed by a parametric test, it was clearly in the same direction as that observed by Inhoff and Fleming (1989). Potential refinements to the click monitoring paradigm include the use of binaural stimulation, the counterbalancing of hand of response, and possibly the adjustment of click intensity.

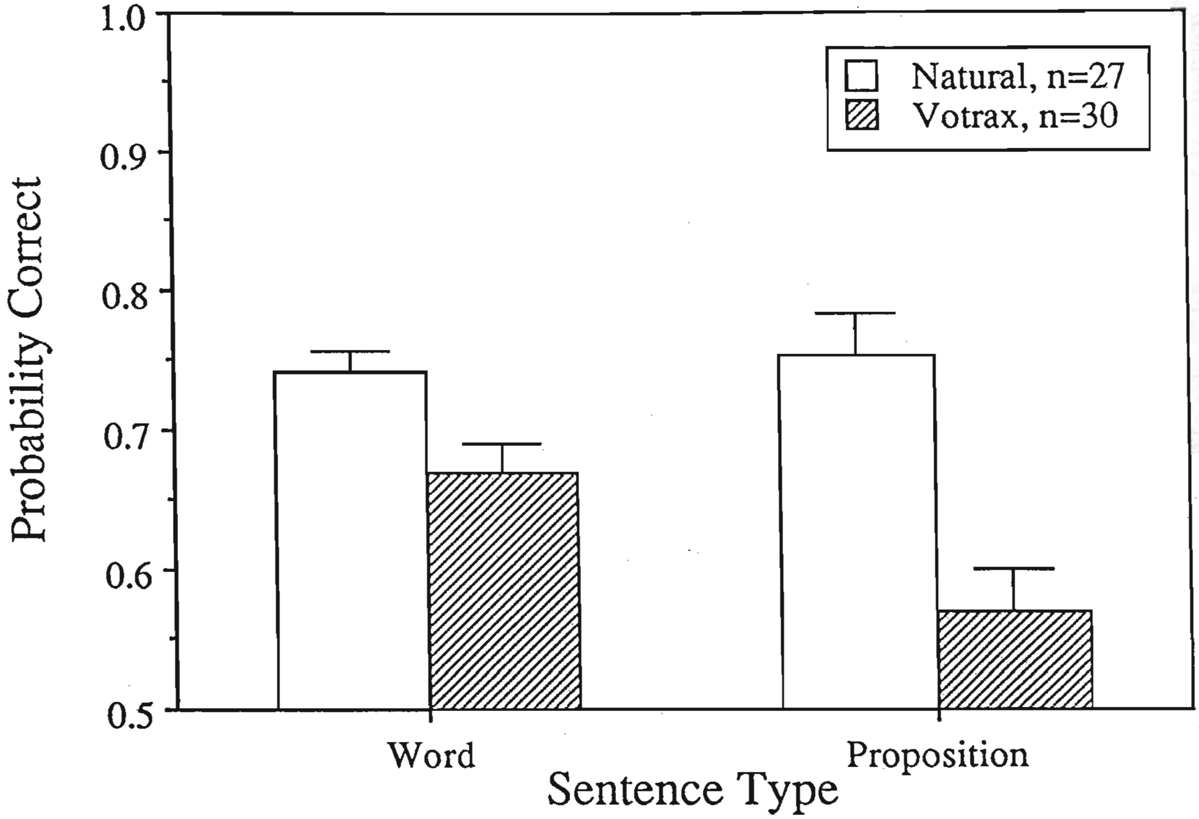


Figure 3. Recognition memory accuracy (probability correct) as a function of sentence type. Open bars represent accuracy for sentences after passages of natural speech and striped bars represent accuracy for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

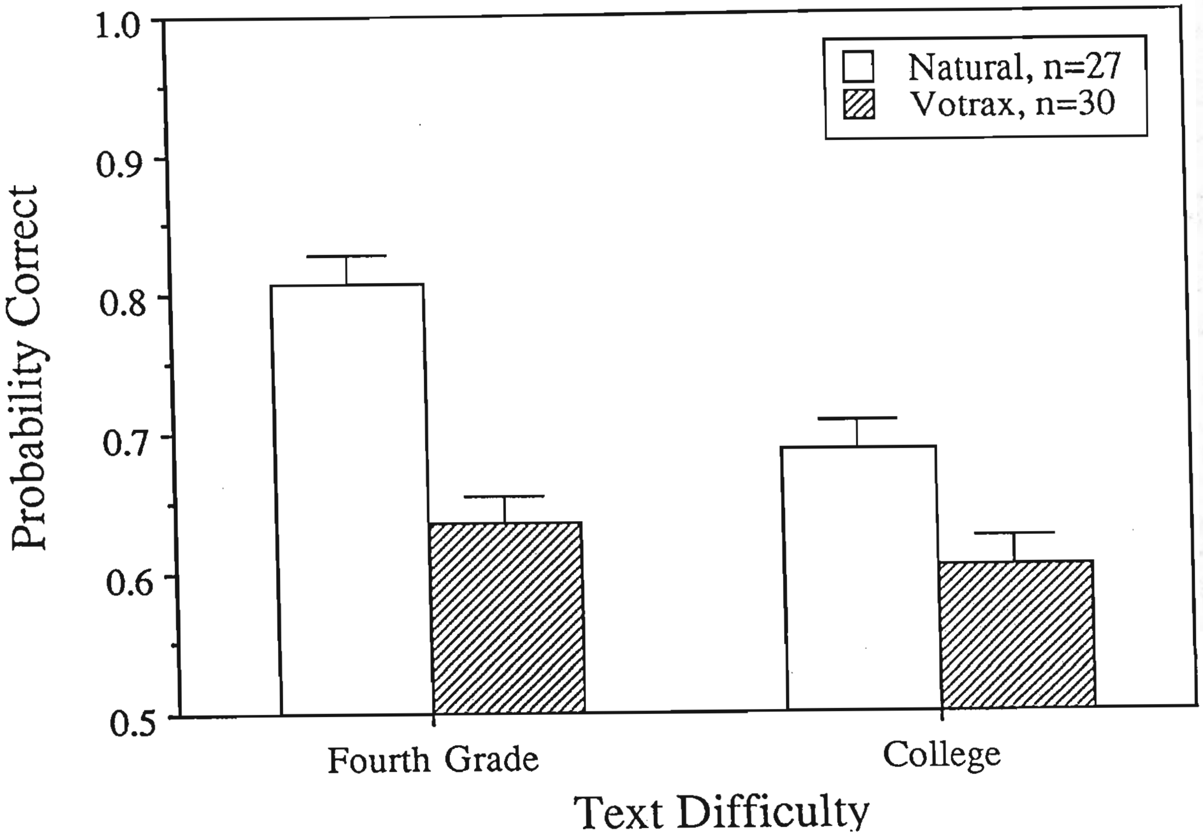


Figure 4. Recognition memory accuracy (probability correct) as a function of text difficulty. Open bars represent accuracy for sentences after passages of natural speech and striped bars represent accuracy for sentences after passages of synthetic speech. Error bars represent one standard error of the sample means.

The recognition memory accuracy data largely replicate results from our previous experiments. We observed significant main effects of voice, text difficulty, and sentence type, as well as significant interactions between voice and sentence type on the one hand and sentence type and text difficulty on the other. The monitoring task itself did not appear to significantly impact on recognition memory performance, although there was a small difference in the direction predicted by a model assuming a common pool of limited processing resources. Finally, we found an unusual interaction between voice and text difficulty. We speculated that this interaction may reflect the near-chance performance of synthetic speech listeners on the recognition memory task.

References

Abrams, K., & Bever, T.G. (1969). Syntactic structure modifies attention during speech perception and recognition. *Quarterly Journal of Experimental Psychology*, **21**, 280-290.

Britton, B.K., Westbrook, R.D., & Holdridge, T.S. (1978). Reading and cognitive capacity usage: Effects of text difficulty. *Journal of Experimental Psychology: Human Learning and Cognition*, **4**, 582-591.

Britton, B.K., Holdridge, T., Curry, C., & Westbrook, R.D. (1979). Use of cognitive capacity in reading identical texts with different amounts of discourse level meaning. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 262-270.

Britton, B.K., Meyer, B.J.F., Simpson, R., Holdridge, T.S., & Curry, C. (1979). Effects of the organization of text on memory: Tests of two implications of a selective attention hypothesis. *Journal of Experimental Psychology: Human Learning and Memory*, **5**, 496-506.

Brunner, H., & Pisoni, D.B. (1982). Some effects of perceptual load on spoken text comprehension. *Journal of Verbal Learning and Verbal Behavior*, **21**, 186-195.

Inhoff, A.W., & Flemming, K. (1989). Probe-detection times during the reading of easy and difficult text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 339-351.

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

Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G., & Mullennix, J.W. (1991, in press). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*, **33**(4).

Posner, M.I., & Boies, S.J. (1971). Components of attention. *Psychological Review*, **78**, 391- 408.

Effects of Talker Variability on Speech Perception: Implications for Current Research and Theory[1]

David B. Pisoni

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

[1] The research summarized in this paper was supported by NIH Research Grant #DC-00111-14 to Indiana University. I thank Beth Greene for her editorial assistance.

Abstract

This paper summarizes recent findings on the effects of talker variability on speech perception. The results of several experiments demonstrate that detailed information about a talker is encoded into multi-layered representations in long-term memory. These representations appear to be much more detailed than the canonical symbolic representations of speech that linguists have traditionally assumed. The representations also appear to be highly context-dependent and apparently preserve a great deal of information in the speech signal. The usefulness of these representations for optimal strategies of lexical access is discussed.

Effects of Talker Variability on Speech Perception: Implications for Current Research and Theory

Research on speech perception over the last forty-five years has been heavily influenced by formal linguistic analyses of spoken language, particularly conceptualizations by phoneticians and phonologists. According to this view, the speech signal can be represented as an idealized sequence of linguistic units distributed in time. These units, the phones or phonemes of speech, represent the "objects" of perception for human listeners and the entities of formal description for linguists. While this idealized conception of speech has been extremely useful for linguistic analysis in phonetics and phonological theory, and while many important generalizations about language have emerged from this approach, these assumptions have also created a set of troublesome problems that have occupied the attention of engineers, speech scientists and perceptual psychologists since the late 1940s. Although the general problems in the field of speech perception are, in principle, basically no different from the problems in other areas of perceptual research, stimulus variability may be unique to speech perception. The extensive variability in speech production has prevented researchers from making substantial progress in solving the "primary recognition problem," that is, mapping physical properties of the speech signal onto linguistic units derived from perception. This is not a problem for linguists who simply assume the existence of these units. But it is a serious problem for engineers and psychologists.

Until recently, there has been fairly good agreement among theorists in the field that the fundamental problem in speech perception is to determine how the continuously varying acoustic signal produced by a speaker is converted into a sequence of discrete linguistic units by the listener so the intended message can be recovered (Studdert-Kennedy, 1974; Pisoni, 1976). This general problem has been approached by examining several more specific subquestions. For example, what stages of perceptual analysis intervene between the presentation of the speech signal to a listener and comprehension of the linguistic message? What types of processing operations occur at each of these stages of analysis? What types of representations are computed at each of these stages and what perceptual and cognitive mechanisms are employed to extract the linguistic message from the speech signal? These have been the primary questions pursued by researchers working in the field of speech perception over the last 15-20 years as they more or less adopted the aims and methodologies of information processing psychology (Pisoni & Luce, 1986). But perceptual psychology has changed dramatically in a number of ways over the last few years and some of aspects of the basic problem described above have been reexamined, recast and elaborated on as knowledge from other fields is brought to bear on the same fundamental issues (Elman & McClelland, 1986).

Absent from the standard list of research problems is the issue of stimulus variability, in particular, the variability produced by different talkers. This is not too surprising because linguistic theory, with its primary emphasis on speech as an idealized representation abstracted away from the physical medium, has basically ignored the problem of talker variability. Linguists have been able to do their analyses of language using well-defined symbolic representations of speech without worrying too much about how these symbols are computed by the perceptual system. In contrast, the situation has been quite different for engineers and perceptual psychologists who do not deal with idealized symbolic representations. For them, the problem of stimulus variability in speech cannot be ignored. It has always been there and it has been difficult to deal with. One of the traditional ways of coping with stimulus variability in speech has been to simply view it as "noise" in the signal that needs to be stripped away in order to get at the symbolic representation of the linguistic message that has been encoded in the speech waveform. Stimulus variability is an inherent part of the speech production process and the resulting acoustic signal generated by the human vocal tract (Elman & McClelland, 1986). We have known for a long time that there are many sources of variability in speech. All of these affect the properties of the acoustic signal and presumably all have some consequences for perceptual analysis.

One aspect of the problem of stimulus variability derives from the physical and articulatory differences among talkers; specifically, the finding that the length and shape of the vocal tract differ quite substantially among different talkers. Moreover, the articulatory gestures used to produce individual phonemes and the strategies used to realize these in different phonetic environments also differ quite substantially among different talkers. The consequence of this is that substantial acoustic-phonetic differences exist among talkers in the physical correlates of most, if not all, of the phonetic distinctions used in spoken languages. While human listeners appear to be able to ignore or minimize differences in talker variability fairly easily through processes of perceptual compensation and talker normalization, currently available speech recognition systems have found these problems much more difficult to overcome.

Another aspect of the stimulus variability problem concerns time and rate normalization in speech perception (Miller, 1987a). Research has shown that the durations of individual speech sounds are influenced quite substantially by an individual's speaking rate. Moreover, these durations are also affected by the locations of various syntactic boundaries in connected speech, syllabic stress, and by the features of adjacent phonetic segments in words. In addition to these sources of variability, substantial differences have also been observed in the durations of segments in words when they are produced in sentence contexts compared to the same words spoken in isolation. Although human listeners show evidence of perceptual constancy in the face of enormous acoustic-phonetic variation, the precise basis for these abilities is still unknown and is a topic of intense research in the field.

In the field of speech recognition, acknowledgment of stimulus variability underlies discussions of speaker-dependent vs. speaker-independent recognition strategies. And, stimulus variability has played an important role in current accounts of human speech perception as well. Discussion of issues such as talker normalization, normalization for speaking rate and perceptual compensation for contextual effects all assume some inherent underlying stimulus variability. Unfortunately, there has been surprisingly little effort to directly pursue the study of stimulus variability in speech perception as an important theoretical problem in its own right. Variability has always been treated by linguists, engineers and psychologists as some perturbation imposed on the dynamically changing speech signal that needs to be filtered out so as to make physically different signals perceptually and functionally equivalent at a symbolic level of analysis, a level of analysis that views speech in an idealized form as a linear sequence of discrete symbols arrayed in time.

By definition, the process of perceptual normalization involves a substantial reduction in information and transformation into a common representation. In the case of talker normalization, the speech signal is stripped of its source characteristics. In the case of rate normalization, the signal is transformed into some time-invariant space to compensate for temporal variation. To what extent is this approach to the problem of stimulus variability justified?

Our research over the last few years on the role of talker variability in speech perception demonstrates that this source of stimulus variability should not be thought of as just noise in the signal. Several of our experiments summarized below demonstrate that source characteristics -- detailed information about the talker -- become an integral part of the perceptual record and are encoded into long-term memory along with the symbolic representation derived from phonetic analysis. From these results, it appears that the phonetic representation of speech in memory is actually much richer and more detailed than necessary for the linguist's description of the speech signal as an idealized sequence of segments and features. It is very likely that these more elaborate representations may provide important new clues to the pattern analyzing operations used to map speech signals onto lexical representations in long-term memory. Thus, rather than thinking of stimulus variability as noise that needs to be filtered out to get at the idealized linguistic message, stimulus variability may actually provide very useful perceptually important information to the listener about aspects of the speech signal that are used for its perceptual analysis and subsequent encoding into memory (Elman & McClelland, 1986; Nakatani & Dukes, 1977).

In the sections below, I will first review earlier studies that reported perceptual effects due to talker variability in speech intelligibility experiments. Then I will summarize several findings from my research group at Indiana. Taken together with the earlier reports, the present set of experiments demonstrates that talker variability is not filtered out or normalized at very early stages of perceptual processing. These new findings raise questions not only about the kinds of representations that have been assumed in the past but also about the close dependency of current models of speech perception on abstract linguistic units such as phonetic segments and phonemes. Despite the usefulness of these hypothetical entities in linguistic analysis, there is some reason to be skeptical of their psychological reality as perceptual units for human listeners (Klatt, 1979, 1989).

Early speech intelligibility studies

One of the earliest studies on the effects of talker variability in speech perception was reported by R.W. Peters in 1955. He studied the relative intelligibility of single-voice and multiple-voice messages in noise. The hypothesis under test was that continuity of voice during a transmission will improve listener reception. The results which are shown in Figure 1 supported his hypothesis. Single-voice transmissions were consistently more intelligible than multiple-voice transmissions. Moreover, multiple-voice transmissions also tended to be more adversely affected by increasing levels of noise than single-voice transmissions. Peters speculated that continuity of voice during a message transmission contributes to listener efficiency through changes in selective attention and continuing adaptation to the same speaker. There appears to be some fine tuning and long-term readjustment to the voice of a single talker that is retained in memory for some period of time and is subsequently used in perceptual analysis.

Insert Figure 1 about here

Another study, entitled "Case of the Unknown Talker," was reported by Creelman in 1957. Using PB words, he collected speech intelligibility data from a group of listeners who heard two types of audio tapes. In one condition, all the test words were spoken by a single talker; in the other condition, the test words were spoken by either two, four, eight or sixteen talkers. The test items were mixed with noise and presented at three speech-to-noise ratios. The results of Creelman's study are shown in Figure 2. As in the earlier study, Creelman also found a tendency for the articulation scores to decrease as the number of talkers increased. The average score for individual talkers under all conditions was about 7 percent greater than the average score when two or more talkers were used. Both Peters and Creelman touched on the problem of talker variability in speech perception but apparently they never continued this line of research any further.

Insert Figure 2 about here

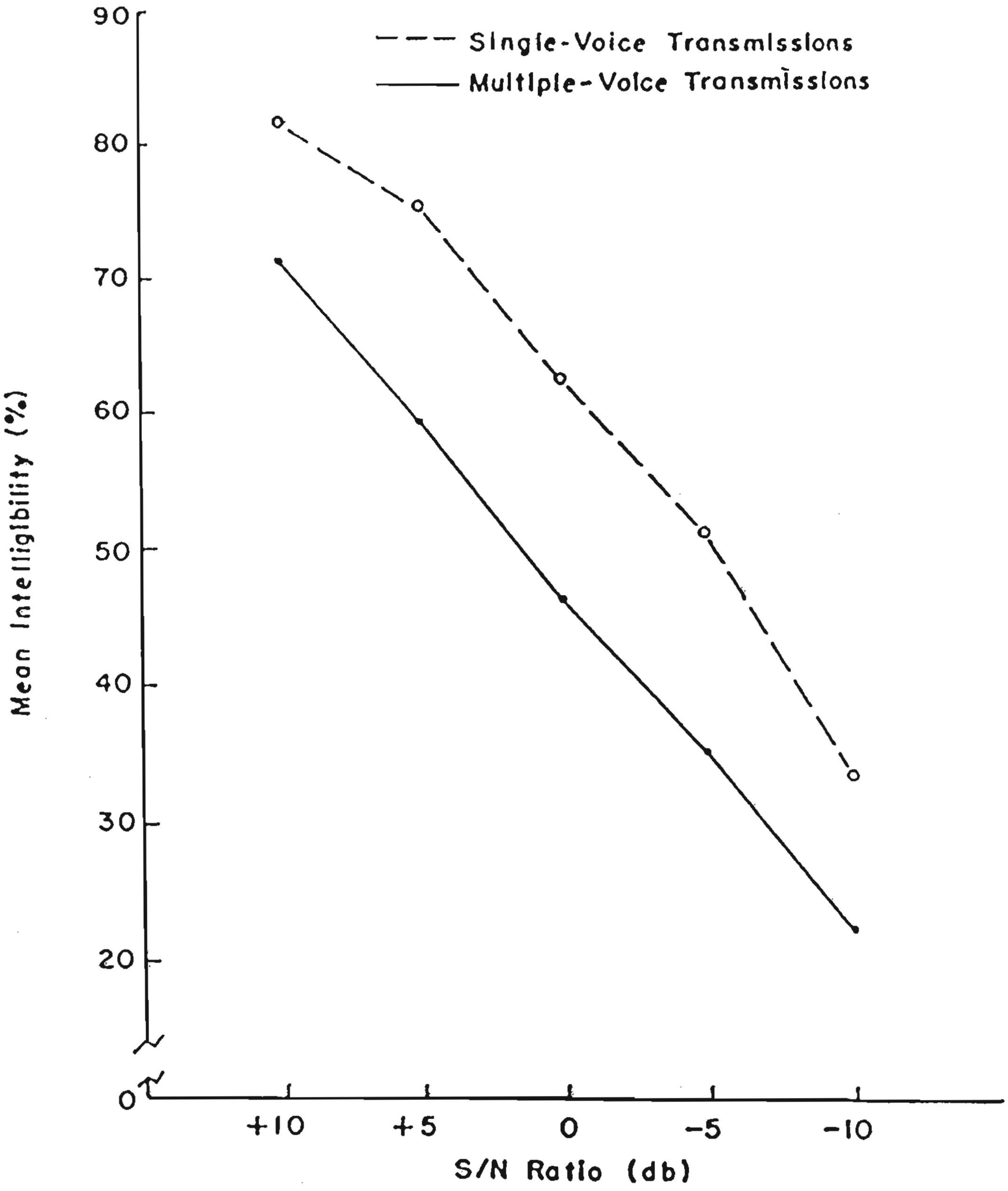


Figure 1. Mean intelligibility scores for single-voice and multiple-voice transmissions at five speech-to-noise ratios (Adapted from Peters, 1955).

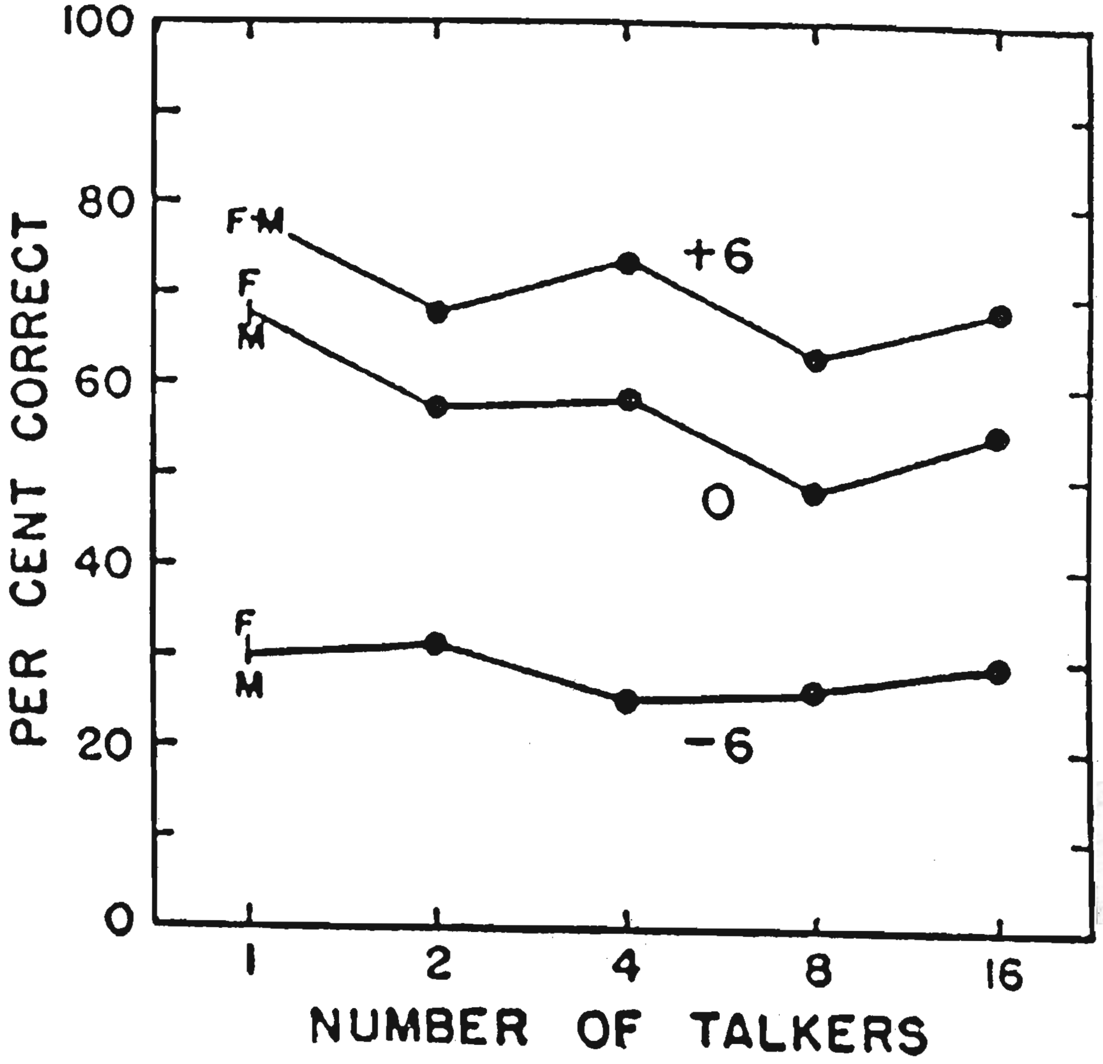


Figure 2. Relation between articulation scores and number of talkers for three different speech-to-noise ratios (Adapted from Creelman, 1957).

Recent studies on talker variability

Over the years, a small handful of perceptual studies have reported that changes in voice affect the perception of both vowels and consonants. Using an identification task, Verbrugge, Strange, Shankweiler, and Edman (1976) showed that the identification of naturally produced vowels was more accurate when the stimuli were produced by a single talker than when they were produced by multiple talkers including men, women and children. Apparently, a change in voice from trial to trial affected perceptual processing and encoding of isolated vowels. Similar changes have also been reported for consonants when the voice varies from trial to trial (Fourcin, 1968). In addition to changes in perceptual identification, processing time is also affected by talker variability. In an early study, Summerfield and Haggard (Summerfield & Haggard, 1973) demonstrated that latencies for categorizing synthetic vowels were slower when the target items were preceded by syllables designed to acoustically emulate a different voice. The authors suggested that the increase in response time due to talker variability reflected some additional processing time for vocal tract normalization to be carried out. According to Summerfield and Haggard, the perceptual system appears to "retune" itself on the basis of vocal tract characteristics each time it encounters a new item produced by a different talker.

Talker variability has also been found to affect perceptual processing time in a same-different matching task. Cole, Colheart, and Allard (1974) found that response latencies to "same" judgements were slower when the target words were produced by two different voices. Taken together, there appears to be some experimental evidence that, at least at the segmental acoustic-phonetic level, talker variability produces reliable differences across a variety of perceptual tasks. The results of these studies are all consistent with the proposal that changes in perceptual performance due to talker variability reflect the operation of some type of normalization process that operates at an early stage of speech perception, a stage typically associated with the construction of some discrete phonetic representation. However, the perceptual operations at this stage constitute only a small subset of the operations employed in the perception and comprehension of fluent speech (Pisoni & Luce, 1986).

At the present time, there is relatively little research in the literature on whether perceptual effects due to talker variability are also present at the lexical level. Current models of spoken word recognition (Forster, 1976; Marslen-Wilson, 1987; McClelland & Elman, 1986), have little, if anything, to say about the role that talker-specific information plays in the recognition of spoken words. Since the effects of talker variability on word recognition have not been addressed in these models, it is reasonable to suppose that these effects are confined to early prelexical levels of processing and probably have little impact on the recognition of spoken words or subsequent comprehension processes, which are typically assumed to occur at higher, more abstract levels of analysis. However, there have been numerous reports of modality effects in the memory literature which suggest that some information about the physical characteristics of spoken words is encoded and stored in long-term memory (Murdock & Walker, 1969; Kirsner, Milech, & Standen, 1983).

In order to determine whether talker variability affects spoken word recognition, experimental procedures must be used that are appropriate for investigating word recognition and lexical access. The perceptual studies examining talker variability that were reviewed above, with the exception of Peters (1955) and Creelman (1957), all involved perceptual tasks that assessed the perception of acoustic cues in nonsense syllables. In order to generalize these earlier results, we used perceptual identification and naming tasks with familiar words. These two tasks are well suited to measuring perceptual performance at a point after which word recognition has already occurred, thus insuring that responses will be made on the basis of the lexical status of the word and not on the acoustic cues or segments contained in the stimulus.

In our first experiment, we attempted to replicate the findings of Peters and Creelman using a similar experimental procedure with a larger set of highly familiar words (Mullennix, Pisoni, & Martin, 1989). In this particular experiment, talker variability and lexical density were manipulated. Talker variability was manipulated by having listeners identify, in one condition, words that were produced by a single talker, or, in a second condition, words produced by 15 different talkers. The stimulus items differed in lexical density, a measure used to index the perceptual similarity of words in the mental lexicon (Luce, 1986). Subjects identified these words at three speech-to-noise ratios, +10 dB, 0 dB and -10 dB. The results are shown in Figure 3 for each of the three S/N ratios. A clear effect of talker variability can be seen. Across all conditions, identification was more accurate for the single-talker lists than the multiple-talker lists. This finding replicates the earlier results reported by Peters and Creelman. A change in voice from trial to trial does, in fact, produce detrimental effects on spoken word recognition.

Insert Figure 3 about here

Unfortunately, the use of the perceptual identification task does not permit an assessment of the effects of talker variability on perceptual processing time. Moreover, the task is dependent on the use of degraded stimuli. Because of these considerations, a second experiment was conducted using a naming task. A number of researchers have used the naming procedure to examine word recognition and lexical access because it provides a method of collecting latency data to stimuli that are uncorrupted by noise (Balota & Chumbley, 1984).

As in the previous experiment, subjects heard words over headphones. However, now subjects were required to repeat the words aloud as fast and as accurately as they could. The data were analyzed in terms of both overall percent correct identification and response latencies. Table 1 shows the mean latencies collapsed over subjects for the single-talker and multiple-talker word lists for high- and low-density words. The effect of talker variability was highly significant ($p < .01$). Response latencies were faster for words in the single-talker condition than words in the multiple-talker condition. A similar pattern was observed for the identification data shown in Table 2. Identification performance was better in the single-talker condition than in the multiple-talker condition.

Insert Tables 1 and 2 about here

Overall, the effects due to talker variability found in the first experiment were replicated in this study using a naming paradigm in which the stimulus items were not degraded by noise. Performance as measured by identification and latencies was consistently worse in the multiple-talker condition compared to the single-talker condition. These results provide additional evidence that talker variability not only affects overt identification responses, but also affects the time course of perceptual processing. The context that the test items are presented in appears to reliably affect identification and response times.

In another experiment, we also examined a factor related to the ease of encoding of the input signal. This factor involved degradation of the acoustic information using a novel signal processing technique (Horri, House, & Hughes, 1971). The digital signal was degraded by flipping the sign of the amplitude value on a certain number of randomly determined samples. This method of degradation was chosen over alternative methods, such as imposing a uniform background of white noise over the stimulus, because any effects due to degradation are a direct consequence of physical disruption and/or distortion of the original information in the signal; that is, the stimulus information which is presented is not degraded by masking noise.

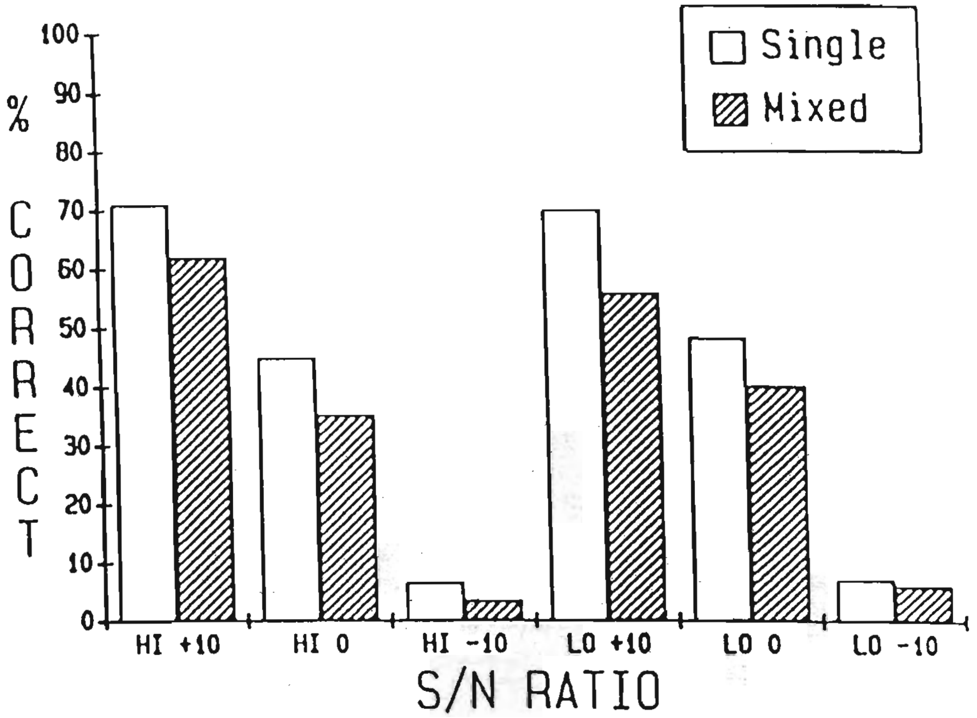


Figure 3. Mean percent correct identification scores for single-talker and multiple-talker conditions as a function of lexical density and speech-to-noise ratio (From Mullennix et al., 1989).

Table 1

Mean response latencies (in ms) for correct naming responses in single-talker and mixed-talker conditions for high and low density words.

	Density	
	High	**Low**
Single talker	611.2	605.7
Mixed talker	677.2	679.4

Table 2

Mean percent correct identification for naming responses in single-talker and mixed-talker conditions for high and low density words.

	Density	
	High	**Low**
Single talker	96.6	95.0
Mixed talker	91.8	91.1

Insert Figure 4 about here

The results of this experiment are shown in Figure 4. As expected, performance was significantly better ($p < .01$) for the single-talker condition (69.1% correct) than the multiple-talker condition (48.1% correct). The figure also shows that performance decreased more for items from the multiple-talker condition than the single-talker condition when the degradation level was increased from 10% to 20%. These differences were also significant ($p < .01$).

The results of this experiment indicate that when the processing of low-level acoustic cues in the signal becomes disrupted as a result of signal degradation, the effects of talker variability on perception become even greater. This finding is consistent with the view that talker normalization processes are intimately related to perceptual operations involved in encoding the sensory input into a phonetic representation rather than search or retrieval processes associated with word recognition and lexical access.

Processing dependencies

Another issue that we studied concerned the relationship between talker normalization and phonetic coding. Do the perceptual processes used to encode voice information function independently of processes that are used to encode phonetic information? Are talker normalization processes and phonetic coding processes interrelated? One way to determine whether perceptual processes are related to one another is to assess whether stimulus dimensions relevant to both types of processes are perceived independently or whether there is some dependency relation between then. We examined the processing relations between talker normalization and auditory-to-phonetic coding processes using a speeded classification technique (Garner, 1974).

One proposal that has been used to account for talker variability effects is a resource-limited talker normalization process at encoding (Mullennix et al., 1989). We suggested that perceptual deficits due to a change in voice occur because of competition for resources used by talker normalization processes and other perceptual operations involved in speech perception. It is conceivable that each time a different voice is encountered, resources must be allocated or reallocated to talker normalization processes until speaker-dependent perceptual operations are completed. If this is the case, perceptual deficits may arise from the additional processing load induced by changes in voice from trial to trial in an experiment. If phonetic coding processes are interfered with by processes involved in talker normalization, then the effects of talker variability may be dependent on selective attention.

To study this problem, we used the two-choice speeded classification task developed by Garner (1974). Subjects are required to attend selectively to one stimulus dimension while simultaneously ignoring another stimulus dimension. Two stimulus dimensions are combined in various ways. In the control set, the unattended dimension is constant while the attended dimension varies randomly. The control set for each dimension provides a baseline measure for classifying each dimension and permits one to assess whether both dimensions are, a priori, equally discriminable. In the orthogonal set, both the attended and the unattended dimensions vary randomly. The degree to which response latencies increase from the control set to the orthogonal set for each dimension indicates the extent to which the stimulus dimensions are processed separably or in an integral fashion. If stimulus dimensions are classified as quickly in the orthogonal conditions as they are in the control conditions, then the stimulus dimensions are said to be processed independently. That is, decisions about the relevant dimension are unaffected by the variation on the irrelevant dimension. However, if there is a significant increase in

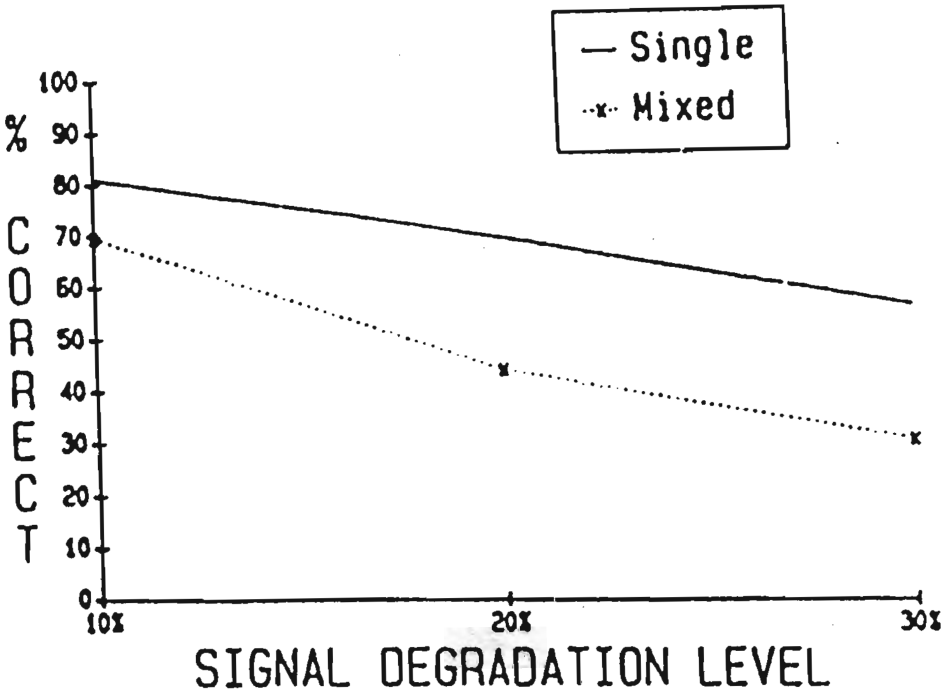


Figure 4. Mean percent correct identification scores for single-talker and multiple-talker conditions as a function of signal degradation level (From Mullennix et al., 1989).

response latencies from the control conditions to the orthogonal conditions, the stimulus dimensions are said to be processed in a dependent manner. In this case, subjects cannot ignore or "filter out" variation in the irrelevant dimension. This result, which is termed orthogonal interference, indicates that a failure of selective attention to the attended dimension has occurred.

Figure 5 shows the amount of orthogonal interference (in ms) for the word and voice dimensions for each of four stimulus conditions from an experiment carried out by Mullennix and Pisoni (1990). Across all four conditions, significant increases in orthogonal interference were obtained when subjects were required to attend to either the word or the voice dimension. The pattern of results shows clearly that the processing of each dimension affects classification of the other dimension. Moreover, this effect increased as stimulus variability increased. Thus, each dimension affects decisions on the other dimension and does so to a greater degree as stimulus variability increases.

Insert Figure 5 about here

A closer examination of the amount of orthogonal interference present for each dimension reveals that the amount of interference was greater for the word dimension than for the voice dimension. Although the two dimensions are processed in a mutually dependent manner, a reliable processing asymmetry is present in these data. Overall, the amount of orthogonal interference obtained for the word dimension was significantly larger than the amount of interference obtained for the voice dimension. Subjects apparently can attend to voice and selectively ignore irrelevant variation in the words. However, they have much more difficulty attending to words when there is irrelevant variation in the voice of the talker.

Because neither voice nor phonetic information can be selectively ignored when subjects are required to attend to specific aspects of a spoken word, we conclude that the processes involved in phonetic coding and those used to encode a talker's voice do not operate independently of one another. The presence of interference effects in the speeded classification task also demonstrates that the processing of voice information is a mandatory encoding operation in speech perception (Miller, 1987b). Given the present findings, it seems reasonable to conclude that decreases in spoken word recognition performance produced by changes in voice may be due to changes in selective attention caused by the mandatory processing of the talker's voice along with phonetic coding of the stimulus pattern into memory.

Memory studies

In addition to these perceptual experiments, we have also completed a series of memory studies that were designed to explore the effects of talker variability on recall of spoken word lists (Martin, Mullennix, Pisoni, & Summers, 1989). Subjects in these experiments were required to recall lists of isolated words in the exact order in which the items were originally presented. In one condition, all of the items on a list were spoken by a single talker; in the other condition, each item on the list was spoken by a different talker. The results of the study by Martin et al. (1989) are shown in Figure 6 which displays serial position curves for single-talker and multiple-talker word lists. Both functions in Figure 6 show the expected primacy and recency effects in the serial position curves for early and late items on the list. However, as shown in the figure, recall of early list items was better for the single-talker lists than the multiple-talker lists. The effect is small but statistically quite reliable. Recall of items from the early serial positions of a list is typically explained by memory theorists in terms of greater opportunities for rehearsal which, in turn, increases the probability that these particular items will be transferred into long-term memory. Thus, the lower performance on multiple-talker lists suggests that these items are not being rehearsed as often or as efficiently as the items from the single-talker lists. The changes in rehearsal may therefore be the result of increased processing demands when items from multiple-talker

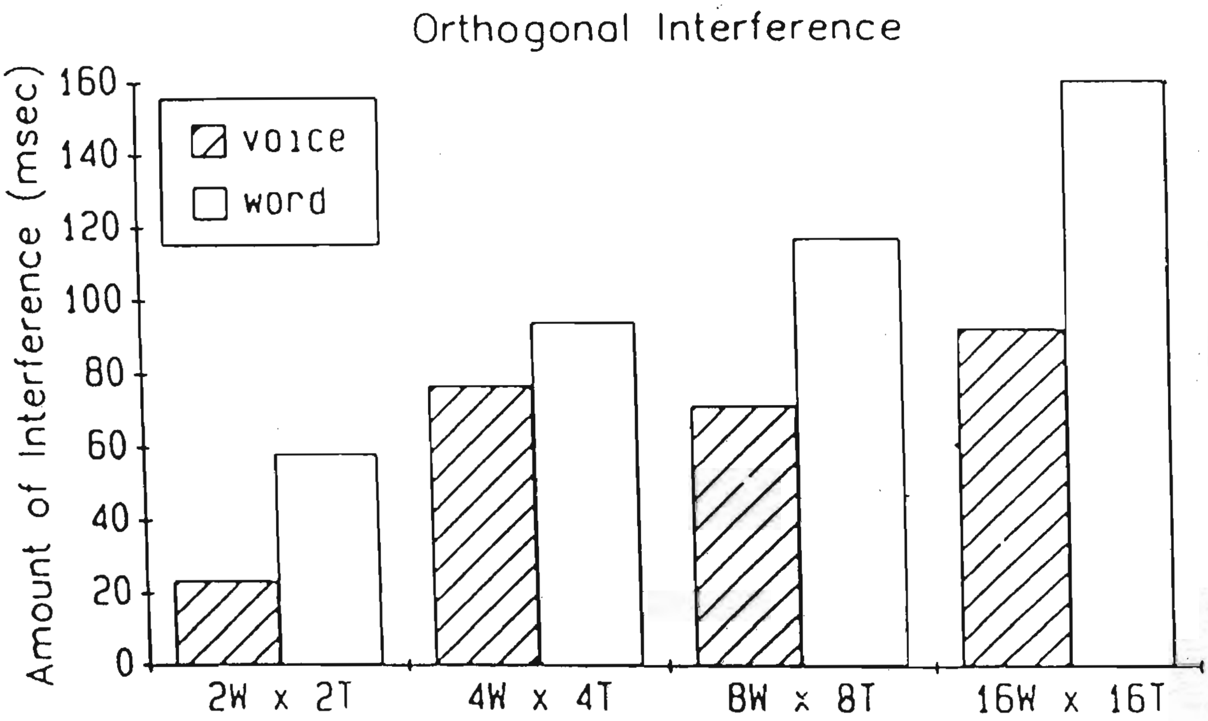


Figure 5. Amount of orthogonal interference (in ms) for word and voice dimensions at each of four levels of stimulus variability (From Mullennix & Pisoni, 1990).

lists are encoded into short-term working memory. The listener may be forced to allocate greater attentional resources to cope with the changing source characteristics of each new talker who is producing each successive word on the list. In short, the processing of words produced by different talkers appears to require more resources in working memory than the processing of words produced by only a single talker. Martin et al.'s findings are consistent with other studies reported in the memory literature showing that stimulus variability reduces recall of early list items (Tulving & Colotla, 1970; Watkins & Watkins, 1980).

Insert Figure 6 about here

The precise nature of the differences in capacity demands between single-talker and multiple-talker word lists was unclear at the time the Martin et al. study was completed although several accounts of the findings were considered. One possibility is that more processing resources are needed only for the initial perceptual encoding of words produced by different talkers. This would result in fewer available resources for subsequent rehearsal of the items on the list. Any differences in the amount, efficiency, or speed of rehearsal would thus be due to the initial differences in encoding these items at the time of input.

A second possibility is that talker variability does not affect the speed or efficiency of initial encoding operations, but rather affects only the efficiency of rehearsal processes that occur after the stimulus items have been encoded into working memory. In this case, more processing resources would be required only for the rehearsal of multiple-talker lists because of the increased uncertainty about the voice of each item.

The present findings on recall of spoken word lists raise several important questions about our current understanding of rehearsal processes in working memory, the transfer of speech from working memory into long-term memory, and the subsequent retrieval of these representations at the time of recall. In order to learn more about the transfer of speech from working memory into long-term memory, another memory study was carried out recently by Goldinger, Pisoni, and Logan (1991). In this study, we varied the rate of presentation of items over a range from 250 ms to 4.0 s. The results of this study are shown in Figure 7 for the five different presentation rates examined. A strong interaction with presentation rate was observed. At fast rates, shown at the top of the figure, we found the same effects reported by Martin et al. (1989). Recall of early list items was worse for multiple-talker lists than single-talker lists. However, at slow presentation rates, shown at the bottom of the figure, the two serial position curves reversed so that now items from the multiple-talker lists were actually recalled better than items from the single-talker lists. The effects of presentation rate influenced recall of multiple-talker lists much more than single-talker lists. The results of the Goldinger et al. (1991) study provide support for the proposal that talker variability affects recall of spoken word lists in two ways. First, talker variability apparently slows down the initial perceptual encoding of speech into a phonetic representation. Second, because of the increased processing demands at the time of encoding, talker variability also reduces the efficiency of the rehearsal process that has been proposed to transfer items from working memory into long-term memory.

Insert Figure 7 about here

Conclusions

The results of our experiments provide new information about speech perception and spoken word recognition and permit us to make several tentative conclusions about the effects of talker variability.

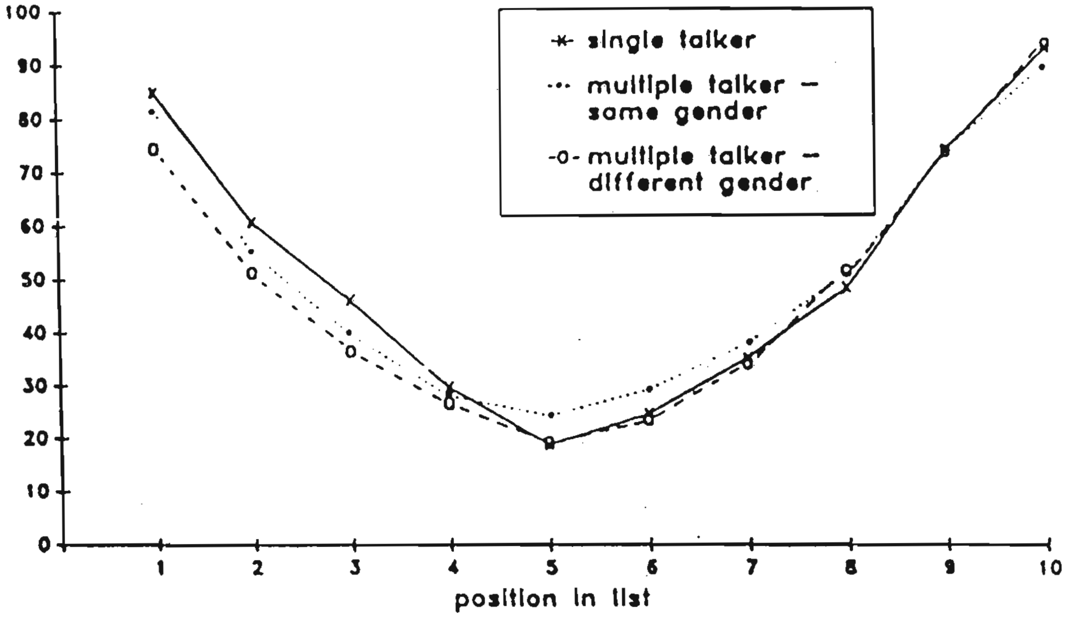


Figure 6. Mean percent correct serial recall collapsed over subjects as a function of serial position and voice.

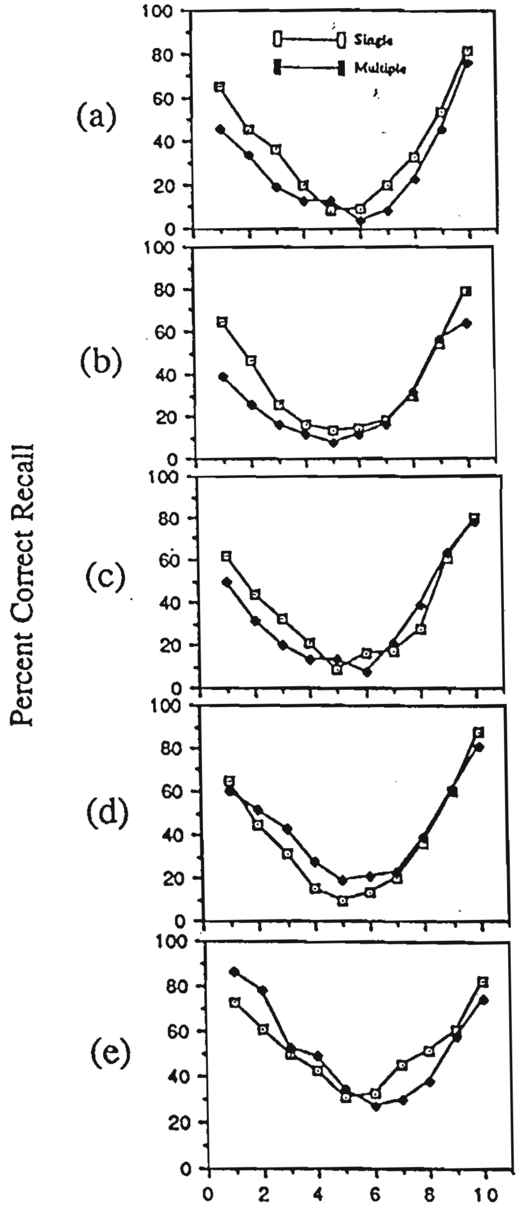


Figure 7. Mean percent correct recall of words for single-talker and multiple-talker lists as a function of rate of presentation. The five panels display the serial position curves at each rate of presentation: (a) one word every 250 ms, (b) 500 ms, (c) 1.0 s, (d) 2.0 s, and (e) 4.0 s.

First, regardless of what the precise nature of the long-term memory representation is, quite detailed information about the source characteristics of the talker is retained in memory for some period of time. Second, most current accounts of speech perception have assumed that talker normalization processes eliminate variability in the speech signal and produce an idealized symbolic representation that is isomorphic with the linguists' description of speech as a sequence of phonemes. The present findings suggest that this view may be incorrect. Third, talker variability not only affects speech perception but findings from our memory studies show that this kind of variability also influences recall performance in auditory memory tasks. Listeners apparently encode talker-specific attributes and use these distinctive properties to encode item and order information. Finally, the present findings suggest that the indexical and linguistic properties of speech signals may not be encoded or represented in memory independently of each other. Rather, these two attributes of the speech signal form an intricate multi-layered representation that may also provide information about the specific pattern-analyzing operations used to encode speech into memory. Our findings indicate that speech signals are not encoded into a canonical segmental representation consisting of a string of idealized phonemes. The representations appear to be very detailed; they appear to be highly context-sensitive and apparently preserve a great deal of information carried in the speech signal.

These general conclusions are consistent with proposals made by Dennis Klatt over ten years ago (Klatt, 1979, 1989). He argued that traditional phonetic representations discard detailed acoustic information that would be useful for lexical access. The loss of this detailed acoustic information, according to Klatt, may produce errors in lexical interpretation that would be difficult to correct if only an idealized symbolic representation were available in memory. Klatt believed that canonical phonetic representations were sub-optimal in human speech perception and machine speech recognition because they violated the principle of "delayed commitment." According to this principle, information in the speech signal should not be discarded until it is no longer of any potential use in perception or recognition. We believe that the human listener is the best example of an optimal speech recognition system and because of this, a great deal can be learned from continued research on the nature of these very detailed representations of speech in long-term memory.

References

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Cole, R.A., Colheart, 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.

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

Elman, J.L., & McClelland, J.L. (1986). Exploiting lawful variability in the speech wave. In Perkell, J.S., & Klatt, D.H. (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum Associates, 360-380.

Forster, K.I. (1976). Accessing the mental lexicon. In Wales, R.J., & Walker, E.C.T. (Eds.), *New Approaches to Language Mechanisms*. Amsterdam: North-Holland, 257-287.

Fourcin, A.J. (1968). Speech-source interference. *IEEE Transactions on Audio and Electroacoustics*, **ACC-16**, 65-67.

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

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

Horri, Y., House, A.S., & Hughes, G.W. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, **49**, 1849-1856.

Kirsner, K., Milech, D., & Standen, P. (1983). Common and modality-specific processes in the mental lexicon. *Memory and Cognition*, **11**, 621-630.

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

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

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.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. In Frauenfelder, U.H., & Tyler, L.K. (Eds.), *Spoken Word Recognition*. Cambridge, MA: MIT Press, 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, and Cognition*, **15**, 676-684.

McClelland, J.L., & Elman, J.L. (1986). Interactive processes in speech perception: The TRACE model. In McClelland, J.L., & Rumelhart, D.E. (Eds.), *Parallel Distributed Processing, Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press, 58-121.

Miller, J.L. (1987a). Mandatory processing in speech perception. In Garfield, J.L. (Ed.), *Modularity in Knowledge Representation and Natural-language Understanding*. Cambridge, MA: MIT Press, 309-322.

Miller, J.L. (1987b). Rate-dependent processing in speech perception. In Ellis, A. (Ed.), *Progress in the Psychology of Language, Vol. III*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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 and Psychophysics*, **47**, 379-390.

Murdock, B.B., & Walker, K.D. (1969). Modality effects in free recall. *Journal of Verbal Learning and Verbal Behavior*, **8**, 665-676.

Nakatani, L.H., & Dukes, D.K. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, **61**, 714-719.

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

Pisoni, D.B. (1976). Speech perception. In Estes, W.K. (Ed.), *Handbook of Learning and Cognitive Processes*. Hillsdale, NJ: Lawrence Erlbaum Associates, 167-233.

Pisoni, D.B., & Luce, P.A. (1986). Speech perception: Research, theory, and the principal issues. In Schwab, E.C., & Nusbaum, H.C. (Eds.), *Pattern Recognition by Humans and Machines*. New York: Academic Press, 1-50.

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

Summerfield, A.Q., & Haggard, M.P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of Speech Research in Progress, Volume 2*. Belfast, North Ireland: Queen's University, 2, 1-12.

Tulving, E., & Colotla, V. (1970). Free recall of trilingual lists. *Cognitive Psychology*, **1**, 86-98.

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

Watkins, O.C., & Watkins, M.J. (1980). Echoic memory and voice quality: Recency recall is not enhanced by varying presentation voice. *Memory and Cognition*, **8**, 26-30.

The Role of Cognitive Factors in the Perception of Synthetic Speech[1]

David B. Pisoni and Beth G. Greene

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

[1] This research was supported by NSF Research Grant #IRI-86-17847 to Indiana University.

Abstract

In this paper, we summarize the results of ten years of research on the perception of synthetic speech produced by rule. A variety of studies on phoneme intelligibility, word recognition, and comprehension have been carried out to learn more about how human listeners perceive and understand synthetic speech. In carrying out this research, we have also been interested in the differences in perception between natural speech and various kinds of synthetic speech produced by rule. While some of our research, particularly the intelligibility studies, has been directed towards issues of perceptual evaluation and assessment, other aspects of the research program, such as the memory and comprehension studies, have been more theoretically motivated to provide new fundamental knowledge about speech perception and spoken language understanding. Our findings suggest that the perception of synthetic speech depends on several factors including the acoustic-phonetic properties of the signal, the specific cognitive requirements of the task, and the previous experience of the listener. Suggestions for future research on comprehension and habituation are also considered.

The Role of Cognitive Factors in the Perception of Synthetic Speech

Introduction

We first became interested in studying the perception of synthetic speech back in 1979 when the MITalk text-to-speech system was nearing completion (Allen, Hunnicutt, & Klatt, 1987). At that time, a number of people including Dennis Klatt, Sheri Hunnicutt, Rolf Carlson, and Bjorn Granstrom were working in the Speech Group at MIT on various aspects of the system. Given our own interests in speech perception, it seemed quite appropriate to carry out several perceptual studies with human listeners to assess how good the synthetic speech actually was. Since that time, we have conducted a large number of experiments to learn more about differences in perception between natural speech and various kinds of synthetic speech produced by rule.

In placing this earlier work in context, it is important at the outset to draw a distinction between basic research on the perception of synthetic speech and questions dealing with assessment and the development of evaluation techniques. While we have been involved with both kinds of activities, most of our research has been oriented toward basic research issues that deal with the perceptual analysis of speech. In particular, we have been concerned with identifying and understanding some of the important perceptual differences between natural speech and several kinds of synthetic speech. By studying the perception of synthetic speech produced by rule, we hoped to learn more about the mechanisms and processes used to perceive speech more generally (Klatt, 1987; Pisoni, Nusbaum, & Greene, 1985).

Perceptual evaluation and assessment was only a side line, although an important topic in its own right. A few years after we began this research program, we generated a list of about a dozen basic research issues that seemed to be important topics for future research (Pisoni, 1981). Table 1 lists these research issues.

Insert Table 1 about here

More recently, several researchers have taken our initial set of findings and have proposed much more detailed assessment and evaluation techniques to test various types of voice output devices. The goal of this work has been to develop reliable methods of evaluation and assessment so that standards can be formulated for use in a variety of languages (Fourcin, Harland, Berry, & Hazan, 1989).

One approach to assessment was proposed recently by Pols (1989). He suggests that assessment techniques be categorized into four broad classes: (1) *global*, including acceptability, preference, naturalness, usefulness; (2) *diagnostic*, including segmentals, intelligibility, prosody; (3) *objective*, including the speech transmission index (STI), articulation index (AI); and (4) *application-specific*, including newspapers, Kurzweil Reading Machine (KRM), telephone information services, weather briefings (NOAA). Much of our own research on the perception of synthetic speech has been concerned with global and diagnostic issues.

Although some aspects of our research has been concerned with evaluation and assessment such as the intelligibility studies using the Modified Rhyme Test (MRT), many of our other studies over the years have not been concerned with practical issues surrounding assessment. Instead, we have focused our research on somewhat more theoretically motivated problems that would provide new insights into why some types of synthetic speech are hard to perceive and understand and how listeners compensate for the generally poor quality acoustic-phonetic information in the signal. In the sections below, we

Table 1

Needed research on the perception of synthetic speech.
(From Pisoni, 1981)

1. Processing Time Experiments
2. Listening to Synthetic Speech in Noise
3. Perception under Differing Attentional Demands
4. Effects of Short- and Long-Term Practice
5. Comprehension of Fluent Synthetic Speech
6. Interaction of Segmental and Prosodic Cues
7. Comparisons of Different Rule Systems and Synthesizers
8. Effects of Naturalness on Intelligibility
9. Generalization to Novel Utterances
10. Effects of Message Set Size

provide a brief summary of the major findings and conclusions from our research program over the last ten years. Both the perceptual evaluation studies and the experimental work have suggested a number of general conclusions about the cognitive factors that affect the perception of synthetic speech. These factors are discussed briefly below. Finally, we offer several suggestions for future research.

Intelligibility of Synthetic Speech

A text-to-speech system can produce three different kinds of errors that may affect the overall intelligibility of the speech: (1) the spelling-to-sound rules, (2) the computation and production of suprasegmental information, and (3) the phonetic implementation rules that convert the internal representation of allophones into a speech waveform (Allen et al., 1987; Pisoni et al., 1985).

In the studies described below, we have focused much of our attention on measures of segmental intelligibility, assuming that the letter-to-sound rules used by a particular text-to-speech system were applied correctly. For most of our work, we ignored the suprasegmentals.

Phoneme Intelligibility

The task that has been used most often in previous studies evaluating synthetic speech and the one we adopted as the "de facto" standard measure of the segmental intelligibility of synthetic speech was the Modified Rhyme Test (Nye & Gaitenby, 1974). In the Modified Rhyme Test (MRT), subjects are required to identify a single English word by choosing one of six alternative responses that differ by a single phoneme in either initial or final position (House, Williams, Hecker, & Kryter, 1965). All the stimuli in the MRT are consonant-vowel-consonant (CVC) monosyllabic words; on half the trials, the responses share the vowel-consonant portion of the stimulus and on the other half, the responses share the consonant-vowel portion. Thus, the MRT provides a measure of how well listeners can identify either the initial or final phoneme of a set of spoken words. In the last few years, several alternative, but similar, tests have been developed (Bezooijen & Pols, 1989; Carlson & Granstrom, 1989; Spiegel, Altom, Macchi, & Wallace, 1989). Some examples of data obtained in the MRT for ten text-to-speech systems are shown in Figure 1.

Insert Figure 1 about here

In addition to the standard forced-choice closed-response MRT, we have also used an open-response format. In this procedure, listeners are instructed to simply write down the word that they heard on each trial. This format provides a measure of performance which minimizes the constraints on the response set; that is, all CVC words known to the listener are possible responses compared to the six alternative responses in the closed-response MRT. This procedure also provides information about the intelligibility of vowels that is not available in the closed-response set version. The procedure requires more cognitive effort and attention because the listener must first encode the auditory stimulus, search through his/her lexicon for one or more appropriate words, and finally select one word as the match for the auditory stimulus. By comparing the results obtained in the closed- and open-response versions of the MRT, we have obtained a great deal of useful information about the sources of error in a particular system (Logan, Greene, & Pisoni, 1989).

Our results have shown large differences in intelligibility between the closed-response and the open-response format. Although the rank ordering of intelligibility remains the same across the two forms of the MRT, it is clear that as speech becomes less intelligible, listeners rely more heavily on response-set constraints to aid performance. Examples of these comparisons are shown in Figure 2.

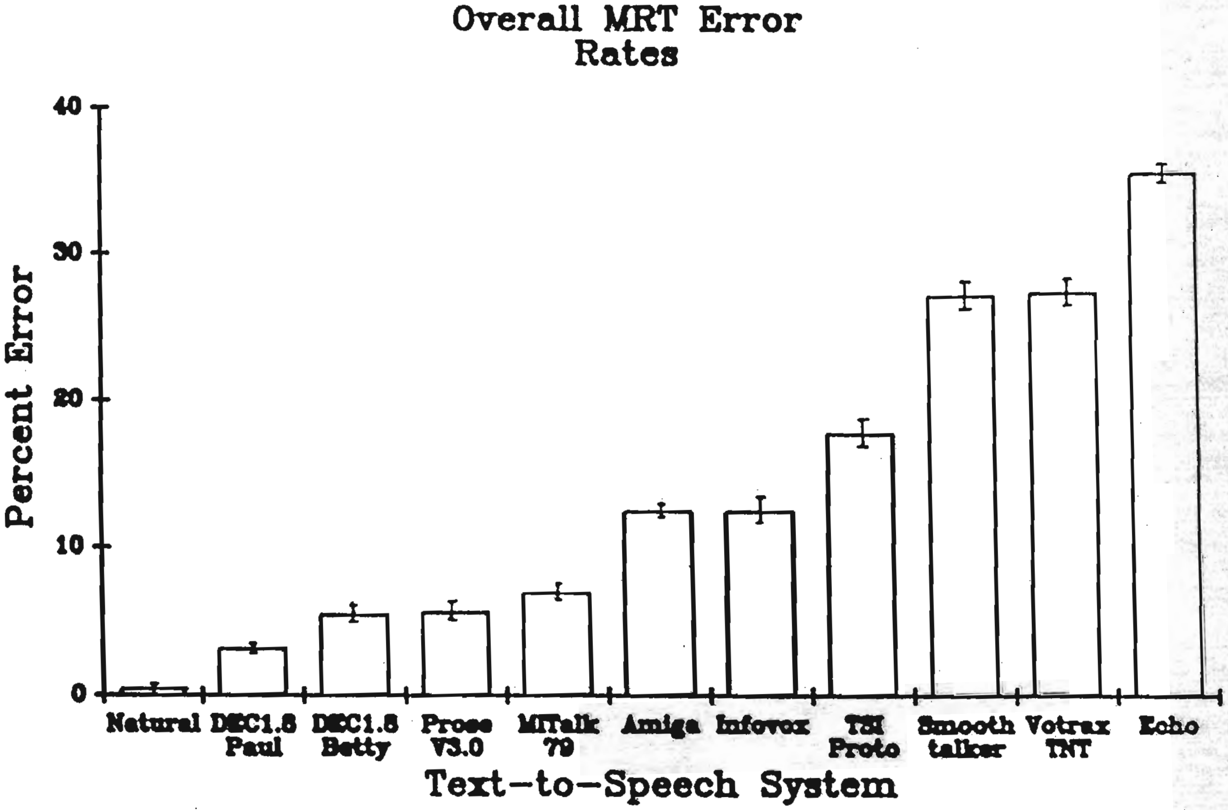


Figure 1. Overall error rates (in percent) for each of the ten text-to-speech systems tested using the MRT. Natural speech is included here as a control condition and benchmark (From Logan et al., 1989).

Insert Figure 2 about here

Nonnative Speakers of English

We have carried several out studies in which nonnative speakers of English listened to both natural and synthetic speech materials (Greene, 1986). Nonnative speakers reveal essentially the same pattern of results as we have found for native speakers: their performance is better when listening to natural speech than synthetic speech. These results were obtained for intelligibility of isolated words in the MRT task and for word recognition in a sentence transcription task. However, the absolute levels of performance were substantially lower for nonnative speakers of English than for native speakers using the same materials.

Our results revealed a wide range of individual differences. Some nonnative speakers showed performance at comparable levels to native speakers. Other listeners showed relatively poor performance. Some of the nonnative listeners had a great deal of experience with English, whereas others had a lot less exposure to the language. We concluded that nonnative speakers responded to synthetic speech in a manner reflecting their general English language ability and experience. In one sense, they responded to the synthetic speech in much the same way as they respond to natural speech -- they either do well on both natural and synthetic speech or they do poorly on both sets of materials.

Sentences: Transcription Performance

To examine the contribution of several linguistic constraints on performance, we compared word recognition in two types of sentence contexts: syntactically correct and meaningful sentences -- "Add salt before you fry the egg" and syntactically correct but semantically anomalous sentences -- "The old farm cost the blood." A recent modification of this task called the Semantically Unpredictable Sentences (SUS) task (Grice, 1989; Hazan & Grice, 1989) uses five different syntactic structures for the anomalous sentences.

By comparing word recognition performance in these two types of sentences, we were able to determine the influence of sentence meaning and linguistic constraints on word recognition. For both natural and synthetic speech, word recognition was much better in meaningful sentences than in the semantically anomalous sentences. Not surprisingly, meaningfulness helps listeners understand sentence length materials. Both top-down and bottom-up processes are required to carry out this transcription task. Furthermore, a comparison of correct word identification in these sentences revealed an interaction in performance suggesting that semantic constraints are relied on much more by listeners when the speech becomes progressively less intelligible (Pisoni, 1987; Pisoni & Hunnicutt, 1980; Pisoni et al., 1985). Subjects have great difficulty inhibiting the use of semantic constraints in word recognition even when it is not helpful to them. Some examples of this are shown in Tables 2 and 3.

Insert Tables 2 and 3 about here

Comprehension of Synthetic Speech

In addition to our studies on segmental intelligibility and word recognition in isolated sentences, we have also been interested in the verification of isolated sentences and the comprehension of long passages of continuous synthetic speech produced by rule.

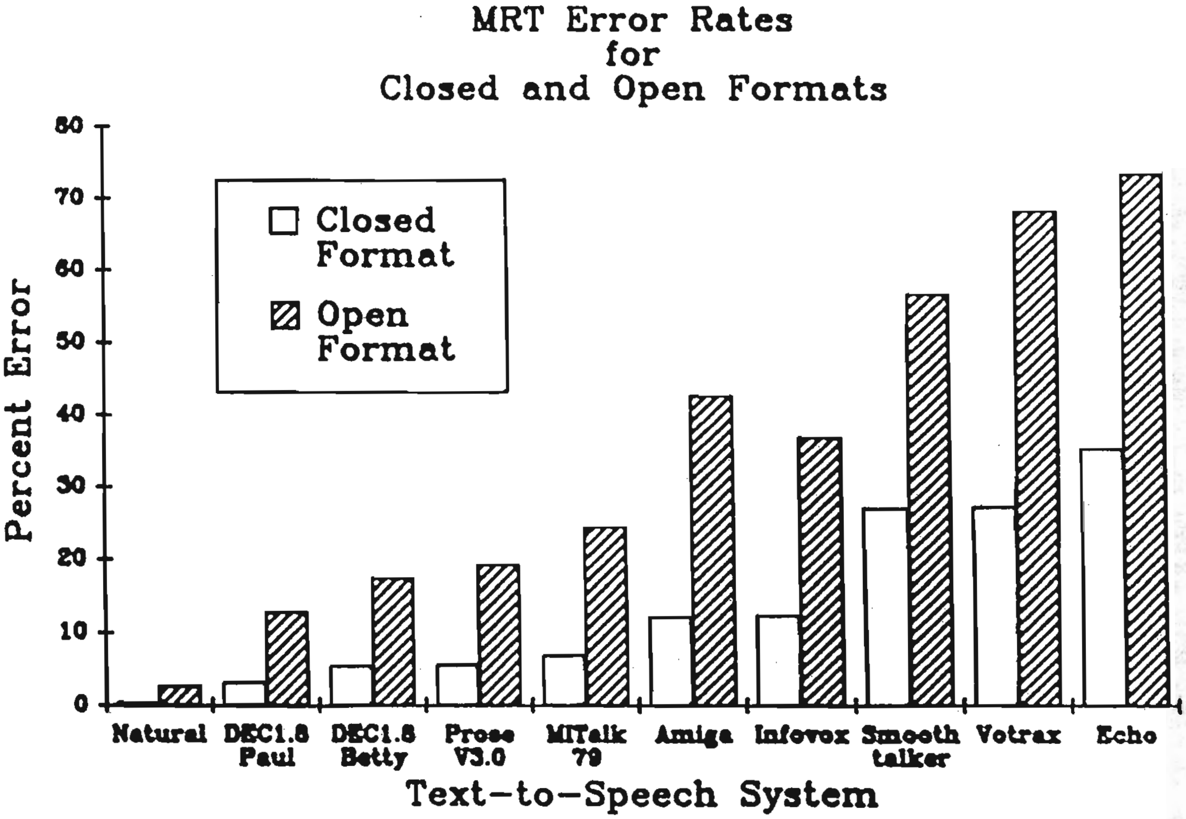


Figure 2. Error rates (in percent) for ten systems tested in both the closed- and open-response format MRT. Open bars designate error rates for the closed-response format and striped bars designate error rates for the open-response format (From Logan et al., 1989).

Table 2

Harvard sentence targets and examples of responses.

Harvard Sentences	
Target:	The juice of lemons makes fine punch.
Responses:	The juice of lemons makes Hawaiian punch.
	The goose lemon makes fine punch.
	The juice of lemons makes a high punch.
Target:	Four hours of steady work faced us.
Responses:	Four hours of study work faced us.
	Four hours of study work pleased us.
	Four hours of study work pays off.
Target:	Smoky fires lack flame and heat.
Responses:	Smoking fires lack flame and heat.
	Smoke is higher than flame or heat.
	Smoky fires lack amounts of heat.
	Smoke and fire lack flame and heat.

Table 3

Haskins sentence targets and examples of responses.

Haskins Sentences	
Target:	The far man tried the wood.
Responses:	The fireman dried the wood.
	The fireman tried the wool.
	The farm hand dried the wood.
Target:	The bright guide knew the glass.
Responses:	The bright guy threw the glass.
	The bright guide mowed the grass.
	The bright guy broke the glass.
Target:	The big bank felt the bag.
Responses:	The big bag felt bad.
	The milk man felt the bag.
	The big man filled the bag.

Sentence Verification Studies

In a series of experiments, we used three and six-word sentences that were either true or false: "Cotton is soft," "Snakes can sing," "You boil water to make rice." These sentences were pretested to determine whether the final word in each sentence was predictable. The sentences were also pretested for intelligibility to insure that they could be transcribed correctly with no errors. Subjects were required to respond to the truth value of the sentences. Results of our verification tests indicated that subjects were faster in responding to natural speech than to synthetic speech. For both natural and synthetic speech, they were faster for high-predictability sentences than low predictability sentences (Pisoni, Manous, & Dedina, 1987). The results showed that although the sentences are highly intelligible, synthetic speech, even high-quality synthetic speech, is still not perceived in the same way as natural speech. As easy as these sentences were to understand in terms of transcription scores, the additional cognitive effort required to understand synthetic speech produced longer response times.

Insert Figure 3 about here

Comprehension of Connected Text

Spoken language understanding is a very complex cognitive process that involves the encoding of sensory information, retrieval of previously stored knowledge from long-term memory and the subsequent interpretation and integration of various sources of knowledge available to a listener. Language comprehension therefore depends on a relatively large number of diverse and complex factors, many of which are only poorly understood by cognitive psychologists at the present time.

One of the factors that plays an important role in listening comprehension is the quality of the initial input signal -- that is, the intelligibility of the speech itself. But the acoustic-phonetic properties of the signal are only one source of information used by listeners in speech perception and spoken language understanding. Additional consideration must also be given to the contribution of higher-levels of linguistic knowledge to perception and comprehension.

In an early study, subjects either listened to synthetic or natural versions of narrative passages, or they read the passages silently. All three groups answered the same set of multiple-choice test questions immediately after each passage. While there was a small advantage for the natural speech group over the synthetic speech group, the differences in performance appeared to be localized primarily in the first half of the test.

The somewhat higher performance on natural speech was eliminated by the second half of the test. Performance for the groups listening to synthetic speech improved substantially whereas performance for the natural speech and the control group that read the texts remained about the same.

The finding of improved performance in the second half of the test for subjects listening to synthetic speech is consistent with our earlier results on word recognition in sentences. We found that recognition performance improved for synthetic speech after only a short period of exposure (Carlson, Granstrom, & Larssen, 1976; Nye & Gaitenby, 1974; Pisoni & Hunnicutt, 1980). These results suggest that the overall differences in performance between the three groups is probably due to familiarity with the output of the synthesizer and is not due to any inherent difference in the basic strategies used in comprehending or understanding the content of these passages.

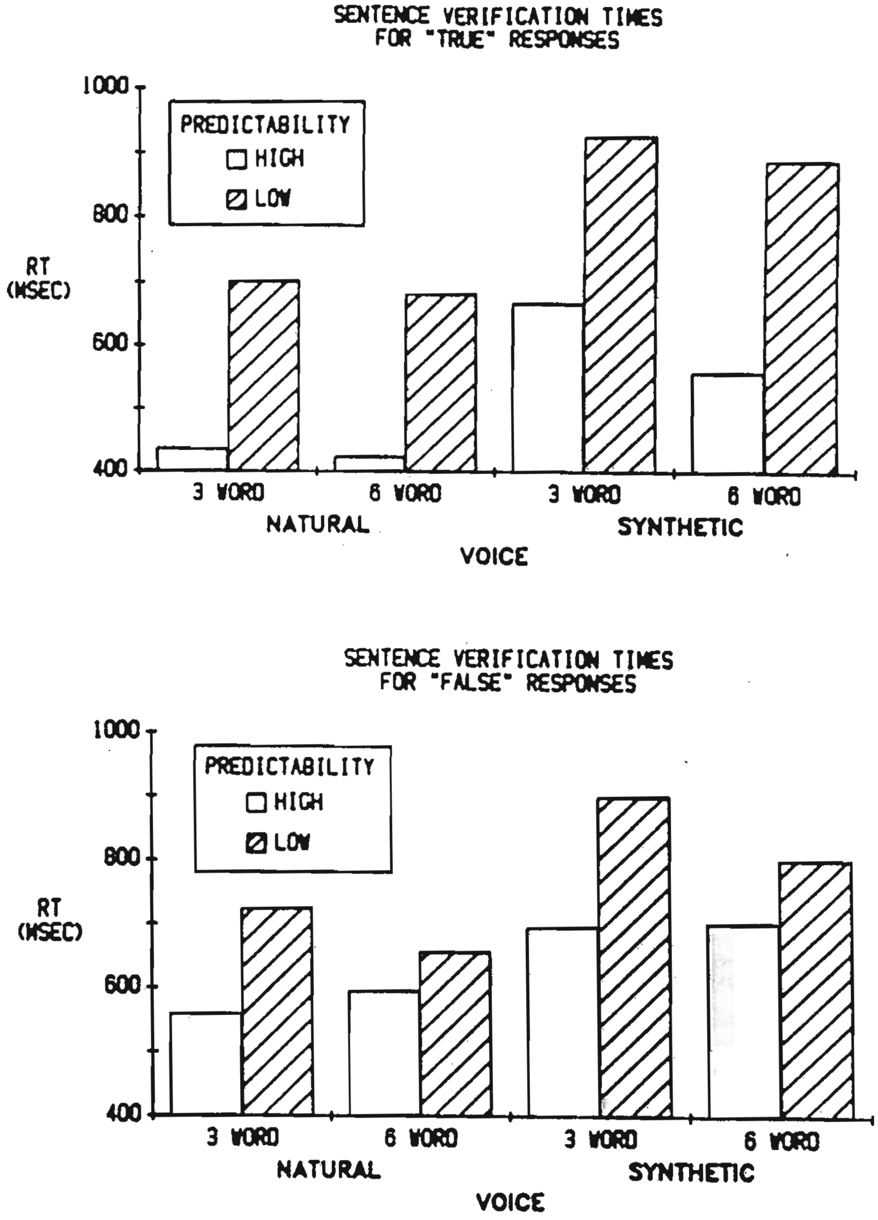


Figure 3. Mean sentence verification latencies (in ms) for the "True" responses (top panel) and "False" responses (bottom panel) for natural and synthetic speech at each of two sentence lengths. The high-predictability sentences are displayed with open bars; the low-predictability sentences are displayed with striped bars. The latencies shown in this figure are based on only those trials in which subjects responded correctly and also transcribed the sentence correctly (From Pisoni et al., 1987).

Comprehension: Online Measures

Recently, we have begun to explore the comprehension process in greater detail using several online measurement techniques. In one study, Ralston, Pisoni, Lively, Greene, & Mullennix (1990), used a word monitoring task to investigate comprehension of natural and synthetic speech. Subjects were required to monitor a spoken passage for a set of target words. Specifically, they had to memorize a set of target stimuli, rehearse the items, and then press a response button whenever they heard one of the target words as they listened to a spoken passage. To ensure that subjects understood the content of the passage, we had them answer a set of comprehension questions after each passage. Word monitoring performance was better for subjects listening to natural speech compared to synthetic speech. Word monitoring performance also decreased as the number of words in the target set increased. Listeners were more accurate in answering questions following presentation of naturally-produced passages than synthetic passages. Thus, both speech quality and memory load affected monitoring performance.

Insert Figure 4 about here

In another study, we used a novel self-paced listening task to measure how much processing time subjects need to understand individual sentences in a passage of connected speech (Lively, Ralston, Pisoni, & Rivera, 1990; Ralston et al., 1990). As expected, we found that listeners required more time to understand synthetic speech than natural speech. When the sentences in the passages were scrambled, listeners required even more processing time for *both* natural and synthetic speech. However, the differences were much larger for the passages of synthetic speech which required greater cognitive effort and processing resources.

Insert Figures 5 and 6 about here

Mechanisms of Perceptual Encoding

The results of the MRT and word recognition studies revealed that synthetic speech is less intelligible than natural speech. In addition, these studies demonstrated that as synthetic speech becomes less intelligible, listeners rely more and more on linguistic knowledge and response-set constraints to facilitate word identification. However, the results of these studies do not provide a theoretical explanation for the differences in perception between natural and synthetic speech. Several different studies were carried out to pursue this problem.

Lexical Decision and Naming Latencies

In order to investigate differences in the perceptual processing of natural and synthetic speech, we carried out a series of experiments that measured the time needed to recognize and pronounce words produced by a human talker and a text-to-speech system (Pisoni et al., 1985). To measure the time course of the recognition process, we used a lexical decision task.

Subjects responded significantly faster to *natural* words and nonwords than to *synthetic* words and nonwords. And, because the differences in latency were observed for both words *and* nonwords alike, and did not depend on the lexical status of the test item, the extra processing effort appears to be related to the initial analysis of the acoustic-phonetic information in the signal and not to the process of accessing words in the lexicon. In short, the pattern of results suggested that the perceptual processes used to encode synthetic speech require more cognitive "effort" or resources than the processes used to encode natural speech.

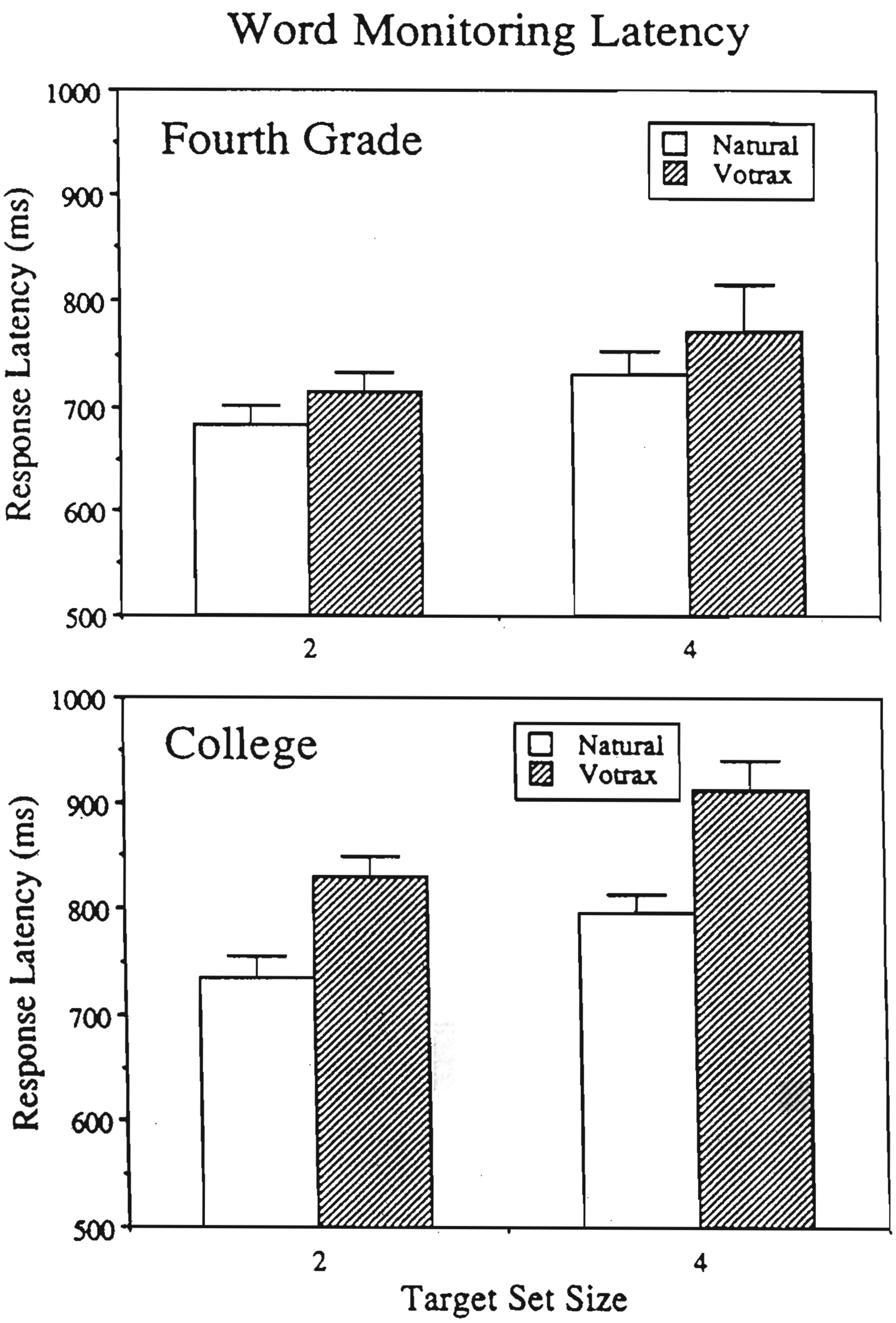


Figure 4. Word-monitoring latencies (in ms) as a function of target set size. The upper panel shows data for fourth-grade passages; the lower panel shows data for college-level passages. Open bars represent natural speech; striped bars represent latencies for passages of Votrax synthetic speech (From Ralston et al., 1990).

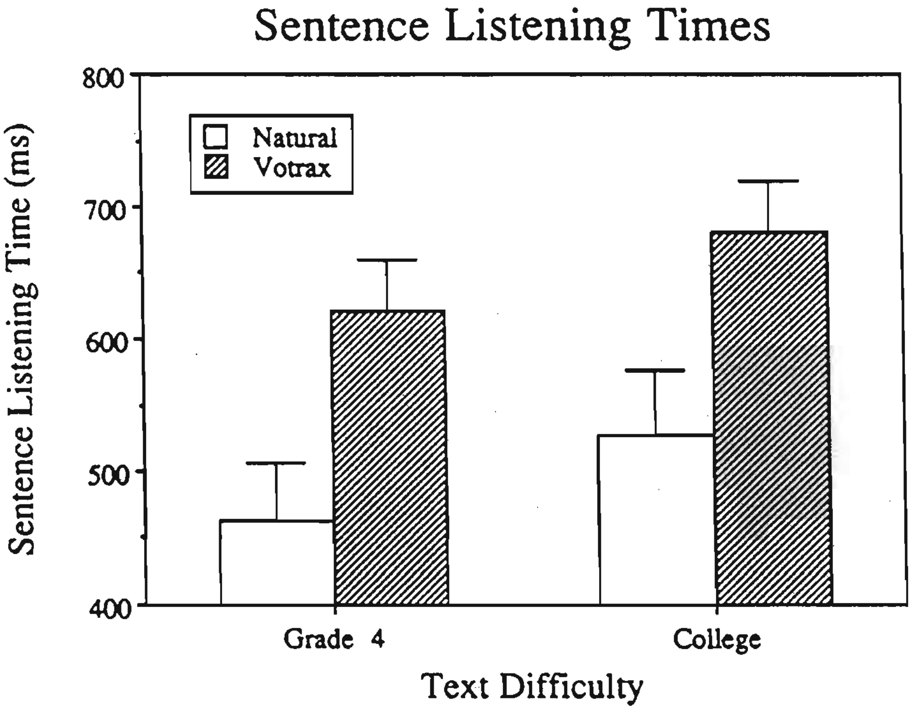


Figure 5. Sentence-by-sentence listening times for natural speech (open bars) and synthetic speech (striped bars) as a function of text difficulty (From Ralston et al., 1990).

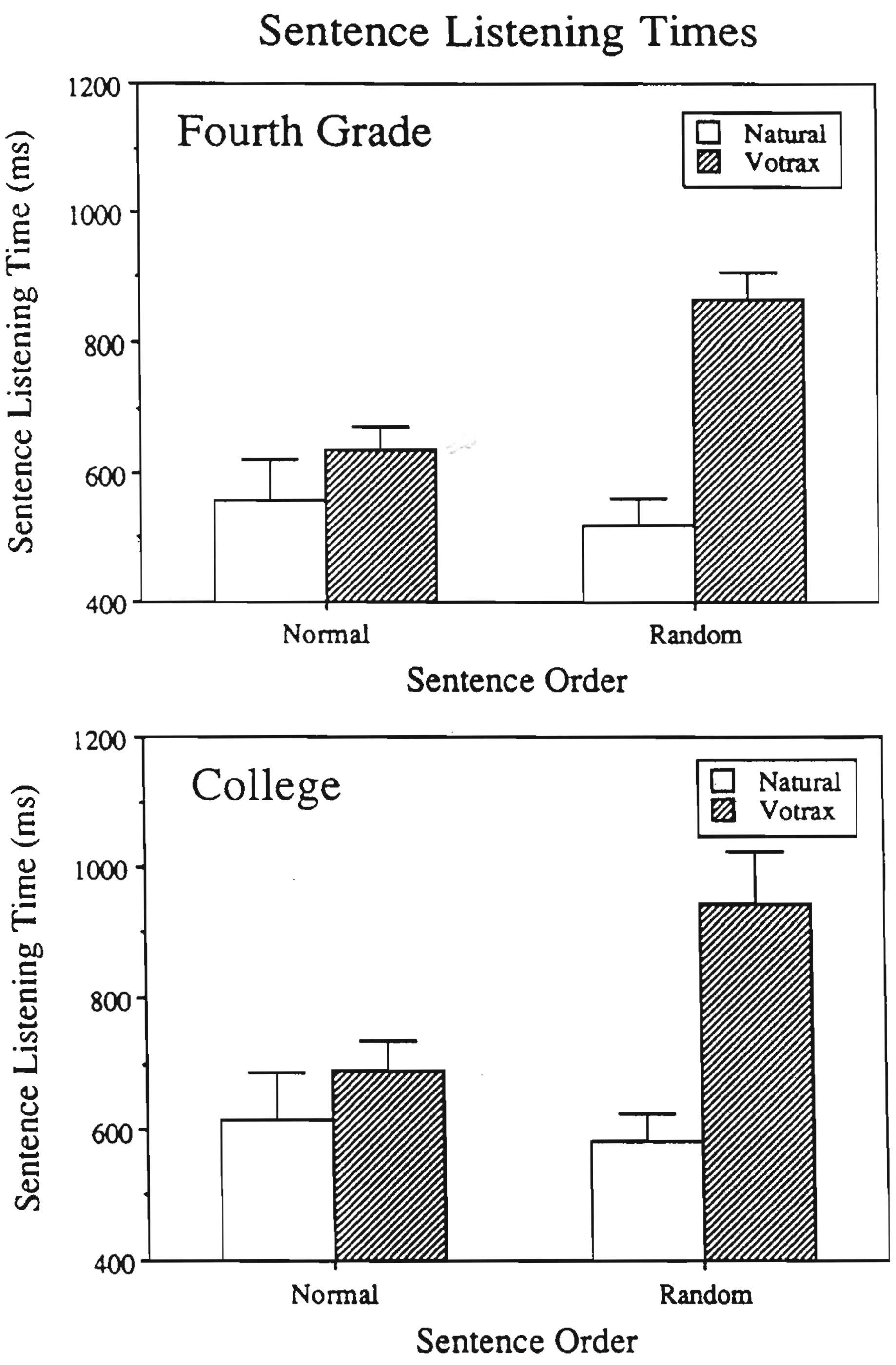


Figure 6. Sentence-by-sentence listening times for natural speech (open bars) and synthetic speech (striped bars) in normal and random sentence orders. The top panel shows data for fourth-grade passages; the bottom panel shows data for college-level passages (From Lively et al., 1990).

Insert Figure 7 about here

Similar results were obtained in a naming task using natural and synthetic words and nonwords. The naming results demonstrated that the extra processing time needed for synthetic speech does not depend on the type of response made by the listener. The results were comparable for both manual and vocal responses. Taken together, these two sets of findings demonstrate that early stages of perceptual encoding for synthetic speech require more processing time than for natural speech.

Consonant-vowel (CV) Confusions

To account for the greater difficulty of encoding synthetic speech, it has been suggested that synthetic speech should be viewed as natural speech that has been degraded by noise. In contrast, an alternative hypothesis, and the one we prefer, is that synthetic speech is not like "noisy" or degraded natural speech at all, but instead may be thought of as "perceptually impoverished" relative to natural speech. Thus, synthetic speech is fundamentally different from natural speech in both degree and kind because it contains only a minimal number of acoustic cues to each phonetic contrast.

To test this proposal, Nusbaum, Dedina, & Pisoni (1984) examined the perceptual confusions for a set of natural and synthetic consonant-vowel (CV) syllables. By comparing the confusion matrices for a particular text-to-speech system with the confusion matrices for natural speech, we found that the predictions made by the "noisy speech hypothesis" were incorrect. Some consonant identification errors were based on the acoustic-phonetic similarity of the confused segments. Other errors followed a pattern that can only be explained as phonetic miscues; these were errors in which the acoustic cues used in synthesis specified the wrong segment in a particular context.

Gating and Signal Duration

The results of the consonant-vowel confusion experiment support the conclusion that the differences in perception between natural and synthetic speech are largely the result of differences in the acoustic-phonetic properties of the signals. In another study, we obtained further support for this proposal using the gating paradigm to investigate the perception of natural and synthetic words. We found that, on the average, natural words could be identified after 67% of a word was heard; whereas for synthetic words, it was necessary for listeners to hear more than 75% of a word for correct word identification. These gating results demonstrate more directly that the acoustic-phonetic structure of synthetic words conveys less information, per unit of time, than the acoustic-phonetic structure of natural speech (Pisoni et al., 1985).

Our results provide strong evidence that encoding of the acoustic-phonetic structure of synthetic speech is more difficult and requires more cognitive effort and capacity than encoding of natural speech. Recognition of words <u>and</u> nonwords requires more processing time for synthetic speech compared to natural speech. The CV confusion study demonstrated that synthetic speech may be viewed as a phonetically impoverished signal. Finally, the gating results showed that synthetic speech requires more acoustic-phonetic information to correctly identify isolated monosyllabic words.

Capacity Demands in Speech Perception

We also carried out a series of experiments to determine the effects of encoding synthetic speech on short-term memory capacity (Luce, Feustel, & Pisoni, 1983). Subjects were given two different lists of items to remember: the first list consisted of a set of digits *visually* presented on a CRT screen; the second list consisted of a set of ten natural words or a set of ten synthetic words. After the spoken list was presented, the subjects were instructed to write down all the visually presented digits in the order

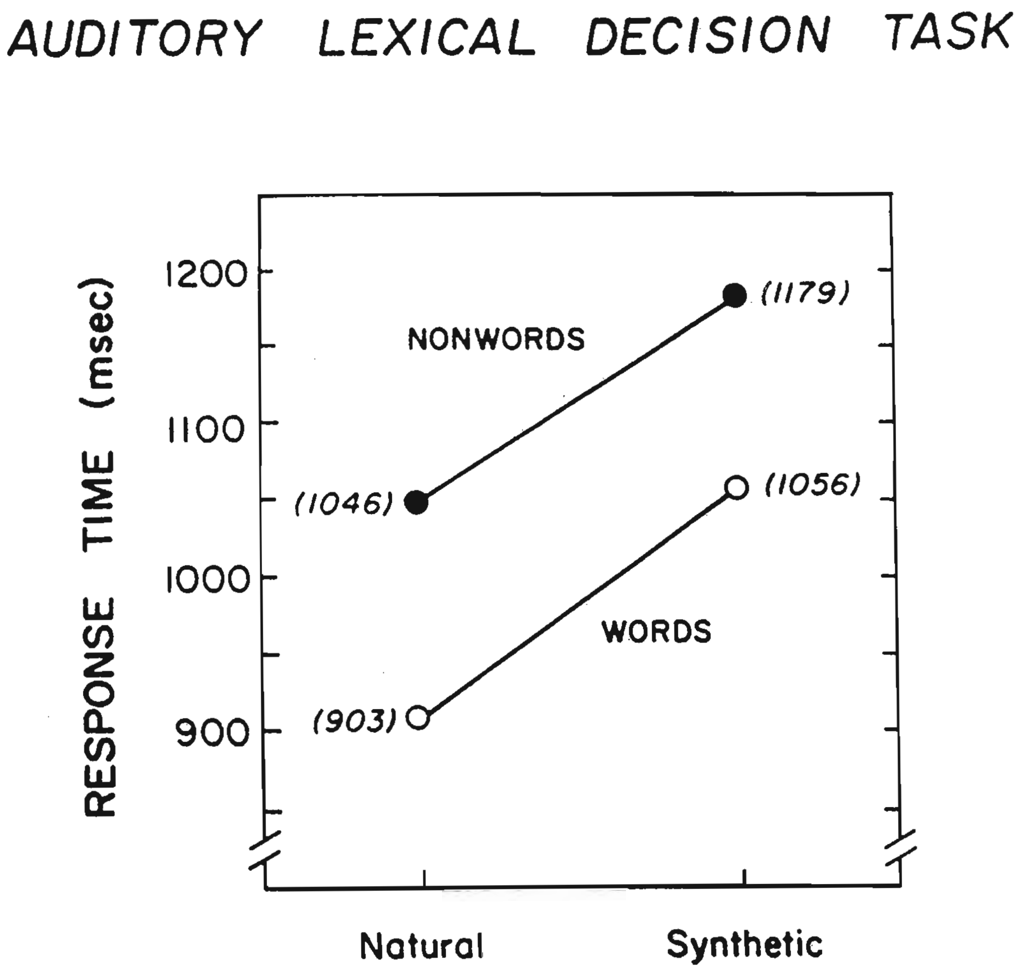


Figure 7. Response times (in ms) obtained in an auditory lexical decision task for words (open circles) and nonwords (filled circles) for natural and synthetic speech (From Pisoni et al., 1985).

of presentation and all the words they could remember from the spoken lists of words. Recall of the natural words was significantly better than recall of the synthetic words. In addition, recall of the synthetic and natural words became worse as the size of the preload digit lists increased. However, the most important finding was an interaction between the *type* of speech presented (synthetic vs. natural) and the *number* of digits presented (three vs. six). As the size of the digit lists increased, significantly fewer subjects were able to recall all the digits for the synthetic word lists compared to the natural word lists. Thus, perception of the synthetic speech impaired recall of the visually presented digits more than the perception of natural speech. These results demonstrate that synthetic speech requires more processing capacity in short-term memory than natural speech. These findings suggest that synthetic speech should interfere much more with other cognitive processes because it imposes greater capacity demands on the human information processing system than natural speech.

Training and Experience with Synthetic Speech

Schwab, Nusbaum, and Pisoni (1985) carried out an experiment to study the effects of training on the perception of synthetic speech. Three groups of subjects followed different procedures for eight days. When pre- and post-test scores were compared, the results showed that performance improved dramatically for only one group -- the subjects that were specifically trained with the synthetic speech. Neither the control group trained with natural speech nor the control group that received no training of any kind showed any significant improvement in recognition of synthetic speech. The results of this training study suggest that human listeners can easily modify their perceptual strategies and that substantial increases in performance can be realized in relatively short periods of time even with poor-quality synthetic speech.

Some Cognitive Factors in Speech Perception

The literature in cognitive psychology over the last forty years has identified several major factors that affect an observer's performance. These factors are: (1) the specific demands imposed by a particular task, (2) the inherent limitations of the human information processing system, (3) the experience and training of the human listener, (4) the linguistic structure of the message set, and (5) the structure and quality of the speech signal.

Task Complexity

In some tasks, the response demands are relatively simple, such as deciding which of two words was said. Other tasks are extremely complex, such as trying to recognize an unknown utterance from a virtually unlimited number of response alternatives, while simultaneously engaging in an activity that already requires attention. In carrying out any perceptual experiment, it is necessary to understand the requirements and demands of a particular task before drawing any strong inferences about an observer's behavior or performance.

Limitation on the Observer

Limitations exist on the human information processing system's ability to perceive, encode, store, and retrieve information. The amount of information that can be processed in and out of human short-term memory is severely limited by the listener's attentional state, past experience, and the quality of the original sensory input. These general principles apply to the study of speech perception as well as other domains of human information processing.

Experience and Training

Human observers can quickly learn effective cognitive and perceptual strategies to improve performance in almost any psychological task. When given appropriate feedback and training, subjects can learn to classify novel stimuli, remember complex patterns, and respond to rapidly changing stimuli presented in different sensory modalities.

Message Set

The structure of the message set, that is, the constraints on the number of possible messages and the linguistic properties of the message set play an important role in speech perception and language comprehension. The choice and arrangement of speech sounds into words is constrained by the phonological rules of language; the arrangement of words in sentences is constrained by syntax; and finally, the meaning of individual words and the overall meaning of sentences in a text is constrained by the semantics and pragmatics of language. The contribution of these levels varies substantially from isolated words, to sentences, to passages of continuous speech.

Signal Characteristics

The acoustic-phonetic and prosodic structure of a synthetic utterance also constrains the choice of response. Synthetic speech is an impoverished signal that represents phonetic distinctions with only a limited subset of the acoustic properties used to convey contrasts in natural speech. Under adverse conditions, synthetic speech may show serious degradation because of the lack in redundancy in the signal, redundancy that is the hallmark of natural speech.

New Avenues of Research

Most of the research on the perception of synthetic speech has been concerned with the quality of the acoustic-phonetic output. As a consequence, researchers have focused most of their attention on improving the segmental intelligibility of synthetic speech. At this point in time, the available perceptual data suggest that segmental intelligibility is quite good for many commercially available systems (DECtalk, Prose, Infovox) and, while not at the same level of intelligibility as natural speech, it may take a great deal of additional research effort to achieve relatively small gains in improvement in intelligibility. In reviewing the studies we have carried out over the last ten years, we see two important areas for future research.

Comprehension and Habituation

When subjects listen to long passages of synthetic speech, they often report difficulty in focusing attention on the linguistic content of the passage. While the results obtained in our earliest comprehension tests indicated that subjects did, indeed, comprehend these passages quite well, the use of an online word monitoring task has provided much more detailed information about how subjects allocate attention differentially across the passages. We also have anecdotal reports from some listeners suggesting that they are "tuning in" and "fading out" as they listen to long passages of synthetic speech. Is synthetic speech more fatiguing to listen to than natural speech? Can a listener fully comprehend a passage when only part of the message is processed? How is the listener's attention allocated in listening to synthetic speech? These are a few of the questions we are planning to pursue over the next few years. Research on comprehension, habituation, and the allocation of attentional resources in language processing are not only topics of practical concern with regard to the design and implementation of voice response systems, but these issues are also at the center of current theoretical work in cognitive psychology and psycholinguistics. Answers to these questions will provide us with useful insights into the reasons why synthetic speech is hard to understand and why it appears to require more attention and effort for the listener.

References

Allen, J., Hunnicutt, S., & Klatt, D.H. (1987). *From Text to Speech: The MITalk System.* Cambridge: Cambridge University Press.

Bezooijen, R.V., & Pols, L. (1989). Evaluation of text-to-speech conversion for Dutch: From segment to text. *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, the Netherlands. 20-23 September.

Carlson, R., & Granstrom, B. (1989). Evaluation and development of the KTH text-to-speech system at the segmental level. *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, the Netherlands. 20-23 September.

Carlson, R., Granstrom, B., & Larssen, K. (1976). Evaluation of a text-to-speech system as a reading machine for the blind. *Quarterly Progress and Status Report, STL-QPSR 2-3.* Stockholm: Royal Institute of Technology, Department of Speech Communication.

Fourcin, A., Harland, G., Barry, W., & Hazan, V. (1989). *Speech Input and Output Assessment.* Chichester, England: Ellis Horwood.

Greene, B.G. (1986). Perception of synthetic speech by nonnative speakers of English. *Proceedings of the Human Factors Society.* Santa Monica, CA: Human Factors Society.

Grice, M. (1989). Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages. *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, the Netherlands. 20-23 September.

Hazan, V., & Grice, M. (1989). The assessment of synthetic speech intelligibility using semantically unpredictable sentences. *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, the Netherlands. 20-23 September.

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

Klatt, D.H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, **82**, 737-793.

Lively, S.E., Ralston, J.V., Pisoni, D.B., & Rivera, S.M. (1990). Some effects of text structure on the comprehension of natural and synthetic speech. *Research on Speech Perception Progress Report No. 16.* Bloomington, IN: Speech Research Laboratory, Indiana University.

Logan, J.S., Greene, B.G., & Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566-581.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors*, **25**, 17-32.

Nusbaum, H.C., Dedina, M.J., & Pisoni, D.B. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. *Speech Research Laboratory Technical Note, 84-02*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Nye, P.W., & Gaitenby, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research, SR-38*. New Haven, CT: Haskins Laboratories, 169-190.

Pisoni, D.B. (1981). Perceptual processing of synthetic speech: Implications for voice response systems in military applications. Paper presented at the Conference on Voice-Interactive Avionics. Warminster, PA: Naval Air Development Center.

Pisoni, D.B. (1987). Some measures of intelligibility and comprehension. In J. Allen, S. Hunnicutt, & D.H. Klatt, *From Text to Speech: The MITalk System*. Cambridge: Cambridge University Press.

Pisoni, D.B., & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. *IEEE Conference Record on Acoustics, Speech and Signal Processing*. New York: IEEE Press, 572-575.

Pisoni, D.B., Manous, L.M., & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, **2**, 303-320.

Pisoni, D.B., Nusbaum, H.C., & Greene, B.G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, **73**(11), 1665-1676.

Pols, L.C.W. (1989). Improving synthetic speech quality by systematic evaluation. *Proceedings of the ESCA Tutorial Day on Speech Input/Output Assessment and Speech Databases*. Noordwijkerhout, the Netherlands. 20-23 September.

Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G., & Mullennix, J.W. (1990, in press). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*, **33**(4).

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

Spiegel, M., Altom, M.J., Macchi, M., & Wallace, K. (1989). A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases*. Noordwijkerhout, the Netherlands. 20-23 September.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Effects of Text Structure on the Comprehension of Natural and Synthetic Speech[1]

Scott E. Lively, James V. Ralston,
David B. Pisoni and Susan M. Rivera

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

[1] This research is supported by NSF Research Grant #IRI-86-17847 to Indiana University.

Abstract

Subjects were required to listen to passages of connected speech using a self paced sentence-by-sentence listening time task. Passages of natural speech and poor quality synthetic speech were presented in either normal sentence order or in a scrambled order. Following passage presentation, postperceptual word and proposition recognition statements were presented. Results indicated that the voice and order variables interacted in the listening time data. Subjects who listened to passages of synthetic speech in random order displayed the longest sentence listening times. An interaction of passage sentence order and recognition memory sentence type was observed in postperceptual comprehension data. Subjects' performance decreased for proposition recognition memory items, but remained constant for word recognition items across sentence orderings. An attentional explanation similar to the one suggested by Ralston et al. (1991) is given for the present set of results. Subjects who listen to synthetic speech attend more closely to the acoustic-phonetic characteristics of the passages. Listeners who hear natural speech, in contrast, are able to selectively attend to higher levels of the text structure. This different focus of attention may cause listeners who hear synthetic speech to have slower on-line comprehension and to have poorer performance on postperceptual comprehension measures.

Some Effects of Text Structure on the Comprehension of Natural and Synthetic Speech

The importance of the implicit given-new contract between the speaker and the listener and the writer and the reader was stressed by Haviland and Clark (1974). When each sentence of a passage reiterates information from preceding sentences, listeners quickly integrate new information into a developing schema of the discourse. Two properties of the given-new contract facilitate this integration. First, texts adhering to the contract are organized via a topicalization structure in which the first sentence of a discourse introduces its principle subject. Second, rapid integration is facilitated by presenting text with a coherent structure. Coherence is maintained by expanding on the topical sentence with each succeeding sentence and by reiterating old information for the listener.

When the given-new contract is violated, the coherence structure and comprehension suffers. Violations of the contract can occur when the speaker or writer fails to introduce the topic early in the discourse or fails to expand upon the topic. These violations have several predictable consequences. For example, reading times are expected to increase when there is no topicalization of a paragraph (Kieras, 1978). In this case, the discourse appears to be a series of unconnected sentences. Reading times are also expected to be inversely related to the causal structure of a text. Sentence-by-sentence reading times might increase as the causal relationship among sentences of a text decreases.

In addition to increases in sentence reading times, decreases in comprehension are also expected as the given-new contract is violated. Decreases in text coherence violate the listener's expectations about the text structure (Kieras, 1978). Kieras suggests that when a text conforms to the given-new contract, subjects comprehend passages by filling in slots within a frame or schema (Kintsch & Van Dijk, 1978). As expectations are violated and slots remain unfilled subjects must adopt a different comprehension strategy. One possible strategy is to rehearse unslotted information in short-term memory. Because of the limited capacity of short-term memory, however, this strategy might lead to reduced comprehension when there is a large amount of unslotted information. A second strategy expands on the first strategy. Subjects could rehearse information in a short-term store and recode it into an unstructured long-term store. Again, the information would not fill slots in the developing text frame and poor comprehension would be anticipated.

Kieras examined the roles of topicalization and coherence in comprehension by systematically scrambling simple passages of text. Comprehension was measured using a combination of a sentence-by-sentence reading time task and a postperceptual memory task. While no significant differences were observed in the reading time data across text-organization conditions, subjects recalled significantly less information from the ill-formed passages. Kieras attributed the nonsignificant difference observed in the reading time data to a trade-off between reading speed and memory accuracy. Subjects may have chosen a fixed reading rate regardless of sentence order. To account for the decreased memory for information in the ill-formed passages, he pointed out that subjects would have to hold the information from each sentence in short-term memory until it could be integrated with information occurring later in the passage. Failure to completely encode information contained in a sentence or failure to integrate the short-term memory information with information about the passage stored in long-term memory would lead to poor postperceptual performance. Either account could be caused by insufficient reading time allocated to information contained in the scrambled passages.

In a second experiment, Kieras focused on issues of workload by presenting three scrambled passages simultaneously. He found that subjects' reading times increased significantly relative to those who read passages in a normally-ordered condition. There was also a significant interaction between sentence order and memory load. Subjects who were presented with three ill-formed passages simultaneously had the slowest reading times. The results were consistent with Kieras's earlier suggestion that more information must be rehearsed in memory in the ill-formed condition. He claimed that adding new information to memory, without integrating it into a passage frame, was more difficult under high workload conditions and that this increased encoding effort inflated subjects reading times. He concluded that capacity demands in short term memory played a significant role in determining the sentence-by-sentence reading times when passages lacked a coherent text structure.

Meyer (1975), in an earlier experiment, retained topicalization and the coherence of her texts, but manipulated the importance of target paragraphs within her stimulus texts. Her data showed that subjects recalled more information from target paragraphs when they occurred high in the discourse structure of the text. Meyer attributed the different levels of recall to differences in the distribution of selective attention across levels of the passage. She suggested that subjects allocated more attentional resources to information that was critical to the understanding of the story. This information was rehearsed more often and was integrated into a higher slot in the developing frame for the text. When the same paragraph was low in informational value, however, less attention was paid to it at the time of encoding. The information from the paragraph may not have been completely integrated with the developing frame or may have assumed a low slot in the text frame. At the time of recall, the information may have been inaccessible to the subject. Thus, recall of information in the target paragraph was poor when it was low in the text hierarchy.

Britton, Meyer, Simpson, Holdredge, and Curry (1979) examined Meyer's (1975) selective attention hypothesis using a click monitoring task combined with a postperceptual free recall test. As in the Meyer study, the target paragraph was either important or relatively unimportant to the meaning of the passage. By Meyer's selective attention hypothesis, reading times for the target paragraph were expected to be longer when the paragraph played an important role in the development of the passage. Click detection latencies were also expected to be slower when the paragraph was high in the discourse structure of the text. If more attention were directed toward the paragraph in the high importance condition, then fewer resources would be available for the click monitoring task. In addition, Britton et al. predicted that more information from the high importance paragraph would be recalled than from the low context paragraph due to increased attention allocated to that paragraph.

Analyses of the on-line reading time and click detection latency data revealed no significant effects of discourse height. However, more information was recalled from the target paragraph when it was high in the discourse hierarchy. Britton et al. found no support for Meyer's selective attention hypothesis. Subjects did not appear to be dedicating any more cognitive effort to the target paragraph in the high importance condition than they did in the low importance condition. In the postperceptual data, the significant difference in the amount of information recalled from the two conditions indicated that subjects processed information from the target paragraphs in different ways. Britton et al. proposed that instead of dedicating more effort to the target paragraph at the time of encoding, more attention was directed toward the paragraph at the time of storage and retrieval. Britton et al. concluded that subjects stored information from the target paragraph in a superordinate position of the text hierarchy when it was important to the development of the ideas in the text. When the information was not relevant to the text's development, it was stored at a subordinate level. At the time of retrieval, subjects had easier access to the superordinate levels of representation. Links to the subordinate levels of representation may be weak or nonexistent. This led to inferior recall performance in the low importance condition.

Britton, Meyer, Hodge, and Glynn (1980) extended Britton et al.'s (1979) work. They showed that lowering subjects' response criteria at the time of testing did not improve subjects' memory for information that was low in the discourse hierarchy. However, a series of cued recall tasks, which varied the complexity of the probe cue, reliably reduced the difference in the amount of information recalled between importance-level conditions. Britton et al. (1980) concluded that selective attention at the time of retrieval was an important factor in the reduction of recall differences between the two conditions. Probe cues that allowed subjects to efficiently traverse discourse hierarchies to low representational levels led to reduced recall differences. They speculated that longer passages might increase the importance of response criteria. In general, however, the ease with which subjects traversed a representational frame developed for a passage determined the degree to which subjects recalled information from that passage.

Kieras (1978) and Britton et al. (1979, 1980) were interested in written text comprehension. As their dependent measures, they collected on-line sentence reading times and postperceptual recognition and cued recall memory data. It was assumed that sentence reading times reflect an on-line measure of text comprehension. Longer reading times indicated an increase in comprehension difficulty. The temporal nature of spoken language, however, does not easily permit the observation of on-line comprehension difficulty. Traditionally, comprehension of spoken passages has been measured by using postperceptual measures, such as free recall (Jenkins & Franklin, 1978), recognition memory (Bruner & Pisoni, 1981), or multiple-choice questions. While these tests may reflect the end products of comprehension, they do not reflect the on-line process of comprehension.

Ralston, Pisoni, Lively, Greene, and Mullennix (1991) assessed on-line comprehension of natural and synthetic speech using two different measures. In their first experiment, Ralston et al. measured on-line comprehension using a word monitoring task. Subjects monitored for 0, 2 or 4 word targets in passages of spoken text. In their second experiment, an auditory analog of the sentence-by-sentence reading time task was used as an on-line measure of comprehension (Mimmack, 1982). Increases in monitoring times or intersentence intervals in the listening time task were assumed to reflect difficulties in comprehension. To insure that subjects were attending to the linguistic content of the passage, a recognition memory test was used to assess memory for words and propositions in the passages. This test was given following the presentation of each passage.

The results reflected the difficulty subjects experienced when listening to the poor quality synthetic speech. Target hit rates in the word monitoring task were lower, response latencies were longer, and listening times in the sentence-by-sentence listening time task were increased for subjects who listened to synthetic speech. As the difficulty of the text changed from fourth-grade level passages to college-level passages, the subjects who listened to passages of synthetic speech showed a larger increase in response latencies in both the monitoring and listening time tasks. A similar pattern was observed in the recognition memory data; subjects who listened to synthetic speech performed more poorly than subjects who listened to passages of natural speech. Luce (1981) and Moody and Joost (1986) found similar results. In addition to a main effect of voice, Ralston et al. (1991) also observed an interaction between voice and recognition memory sentence type. Subjects who listened to synthetic speech performed significantly worse in a recognition memory test when they were required to verify propositions than when they were required to verify words. Luce (1981) reported a similar advantage for word recognition statements for subjects who listened to synthetic speech.

Ralston et al. (1991) interpreted the results of the two experiments as support for the proposal that subjects who listened to poor quality synthetic speech engaged in a different attentional strategy relative to those who listened to natural speech. When listening to natural speech, subjects quickly and easily recognized the words in the passage and therefore could attend more closely to the meaning of the passage. Little attentional effort had to be dedicated to low-level encoding. This accounted for the fast target detection and listening times and the high recognition memory accuracy scores. Subjects listening to the synthetic passages, in contrast, allocated more attention to the acoustic-phonetic structure of the input. Low-level encoding demands were assumed to tax limited capacity processing mechanisms. Few resources were available to dedicate the comprehension process because of the heavy resource allocation at the acoustic-phonetic level. This pattern of attentional allocation accounted for the relative advantage that word recognition statements had over proposition recognition statements for subjects who listened to synthetic speech.

Given that the sentence-by-sentence listening time task appears to be comparable to reading time data in that both are sensitive to certain types of text variables, we were interested in using the listening time procedure to examine text variables other than difficulty. The present study was designed to assess the role of discourse coherence in the comprehension of synthetic speech. Subjects in the our experiment listened to same randomized versions of the passages used in the Ralston et al. (1991) experiments. We expected to replicate and extend certain aspects of the data collected by Kieras (1978) and Ralston et al. (1991). First, we expected subjects to display longer intersentence intervals when listening to poor quality synthetic speech than natural speech. The longer

response intervals should reflect the difficulty that subjects have in encoding synthetic speech. Second, intersentence intervals were expected to be longer for passages presented in random order. A main effect of order would reflect the difficulty subjects have in integrating incoherent information into a representation frame for the passage. An interaction between voice type and sentence order was also anticipated. Subjects who listened to poor quality synthetic speech in the random sentence condition should show the slowest listening times. The lack of text structure should affect listeners who hear synthetic speech more because they may have little contextual support to aid them in comprehension. In the recognition memory data, main effects of voice and order were also predicted. Recognition memory for natural speech should be superior to synthetic speech. Well-ordered passages should display better recognition memory than ill-formed passages.

Method

Subjects

Subjects were 60 volunteers enrolled in introductory psychology classes at Indiana University. All subjects claimed to be audiologically normal at the time of testing and each received partial course credit for their participation. All subjects were native speakers of English.

Materials

Passages. The passages were identical to those used by Ralston et al. (1991). The difficulty of the passages was manipulated between a level appropriate for fourth grade readers and a level appropriate for college readers. Fourth grade passages were taken from Pauk's Six-Way Paragraphs (Pauk, 1983). The college level passages were drawn from a variety of written comprehension tests appropriate for college level readers. All passages were written in narrative or expository style and dealt generally with cultural or scientific topics.

Recognition Memory Test. Eight recognition memory sentences were constructed for each passage. Each sentence was between five and ten words in length and was stated in declarative form. The sentences were identical to those used by Ralston et al. (1991). Half of the sentences from each passage tested recognition memory for specific words that occurred in the passage. The remaining half of the questions tested memory for propositions in the passage. Half of the recognition memory sentences were true, while the remaining sentences served as distractors. Word recognition distractors were either phonologically or semantically similar to words that occurred in the passage. Proposition distractors were based on information stated in the passage that was recombined in a semantically meaningful way.

All of the recognition sentences had been pretested previously by Ralston et al. (1991) to insure that subjects had to listen to the passages in order to correctly answer the questions.

Stimulus Preparation. The stimuli were sentence-length digital files used by Ralston et al. (1991). A male native speaker of English produced each of the passages. Original analog recording was done in an IAC sound attenuated booth using a DO45 Electro-Voice microphone connected to an AMPEX AG500 reel-to-reel tape recorder running at 15 ips. An identical set of passages was also recorded to analog tape using a Votrax Type-N-Talk speech synthesizer controlled by a VAX 11/750 computer. No special stress or pronunciation rules were applied to the synthetic speech. Analog tapes were low-pass filtered at 4.8 kHz and were digitally sampled at a rate of 10 kHz using a 12 bit D/A converter interfaced to a PDP 11/34 laboratory computer. Digital versions of each passage were segmented into individual sentences using a waveform editor (Luce & Carrell, 1981). The resulting sentence length files were equated for RMS amplitude.

Procedure

Due to the requirements of the sentence-by-sentence listening time task, subjects were run individually in a small, sound-treated booth equipped with a set of TDH-39 calibrated headphones, a CRT monitor (GBC Standard CRT Model MV-10A) and a three-button response box, interfaced to a PDP 11/34 laboratory computer. Instructions on how to perform the sentence listening time task were presented to subjects both visually and auditorally. Instructions appeared on the monitor or over the headphones in a sentence-by-sentence manner. At the end of each spoken sentence, a cue light was illuminated, indicating that the next sentence was ready for presentation. Subjects received the next sentence in the set of instructions 250 ms after they pressed a button labelled "Continue" on the response box in front of them. Subjects were given a practice passage in the sentence-by-sentence format, followed by a practice set of recognition questions. After the practice passage, subjects listened to ten test passages. In the scrambled condition, each subject heard a different randomization of the sentences within each passage. All subjects heard the passages in the same random order. All of the stimulus materials were presented at a comfortable listening level between 75 and 80 dB SPL. Latencies between the offset of the sentence and the onset of the button press signaling a subject's readiness for the next sentence were measured for each sentence of each passage. Eight recognition memory sentences followed each of the passages. Accuracy and latency measures were recorded for each recognition memory question. Stimulus presentation and response collection were controlled by a PDP 11/34 laboratory computer.

Results

The results will be presented separately for the sentence-by-sentence listening time task (SBSLTT) and the recognition memory test. Voice (natural vs. synthetic) and Sentence Order (normal vs. random) were between-subjects variables. Text difficulty (fourth grade vs. college) and recognition memory sentence-type (word recognition vs. proposition recognition) were within-subjects variables.

Sentence-by-Sentence Listening Times

An analysis of variance (ANOVA) was conducted on the listening time latency data. Two main effects were observed in the sentence listening time data. First, subjects who listened to synthetic speech had significantly longer intersentence intervals than subjects who listened to natural speech [$F(1,56)=13.69$, $p<.01$]. The mean intersentence interval for synthetic speech was 783.56 ms, while the mean intersentence interval for natural speech was 568.61 ms. A main effect of passage difficulty was also observed [$F(1,5)=27.66$, $p<.01$]. Intersentence intervals for fourth grade passages were shorter than for college passages (644.21 ms vs. 707.96 ms). The main effect of order failed to reach significance [$F(1,56)=3.13$, $p<.09$], although there was a trend suggesting that listening times were longer for sentences in the random order condition.

Voice and sentence order participated in the only significant interaction observed in the listening time data [$F(1,56)=5.61$, $p<.05$]. Figure 1 displays this interaction. Listening times decreased slightly for passages of natural speech as the order changed from normal to random. Listeners who heard synthetic speech, in contrast, showed an increase in listening times as sentence order changed from normal to random.

Insert Figure 1 about here

Recognition Memory Results

A separate analysis of variance was performed on the recognition memory data. Voice and sentence order were treated as between-subjects factors while sentence type and passage difficulty were treated as within-subjects variables. Overall, a main effect of voice was observed [$F(1,56)=32.77$, $p<.01$]. Subjects who listened to naturally produced passages answered a higher percentage of the recognition memory questions correctly (85.4% correct natural vs. 75.3% correct synthetic). A main effect of passage difficulty was also observed. Recognition memory performance was better for fourth grade passages than for college level passages (82% correct fourth grade vs. 77.7% correct college). The main effects of order and sentence type failed to reach significance.

In addition to the two main effects observed in the recognition memory data, several interactions were also significant. First, sentence type interacted with voice [$F(1,56)=13.06$, $p<.01$]. The form of this interaction replicated the findings reported by Ralston et al. (1991) and Luce (1981). Subjects who listened to synthetic speech performed better on the word recognition questions than on the proposition recognition questions. In contrast, subjects who listened to natural speech had better performance for proposition recognition questions. This cross-over interaction is displayed in Figure 2.

Insert Figure 2 about here

Sentence type also interacted significantly with sentence order [$F(1,56)=6.55$, $p<.05$]. Word recognition performance remained constant across both orderings,while proposition recognition performance decreased when the sentences were randomly ordered. Figure 3 shows the sentence type by order interaction.

Insert Figure 3 about here

The final significant interaction observed in the recognition memory data occurred between sentence type and text difficulty [$F(1,56)=22.65$, $p<.01$]. Word recognition accuracy remained constant across the two levels of text difficulty. Proposition recognition accuracy, in contrast, was lower in the college level passages relative to the fourth grade level passages. Figure 4 displays the sentence type by difficulty interaction.

Insert Figure 4 about here

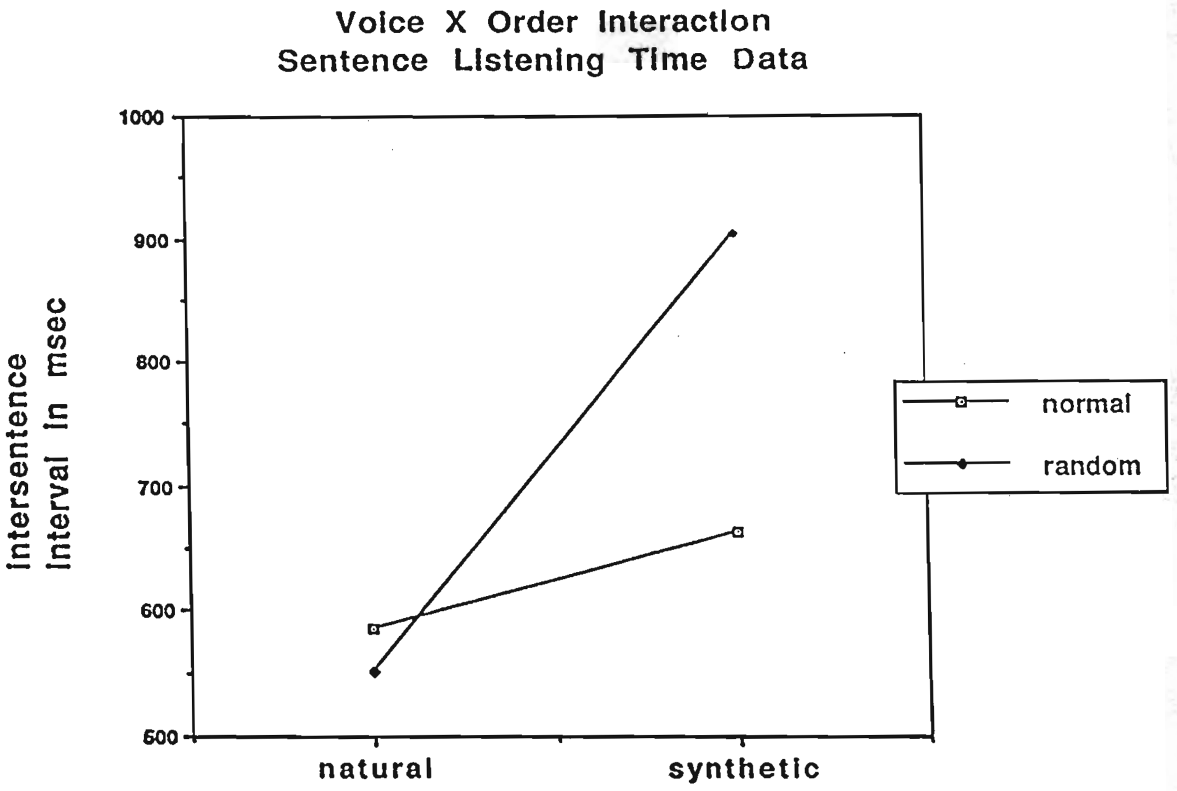


Figure 1. The interaction between voice type and sentence ordering observed in the listening time data.

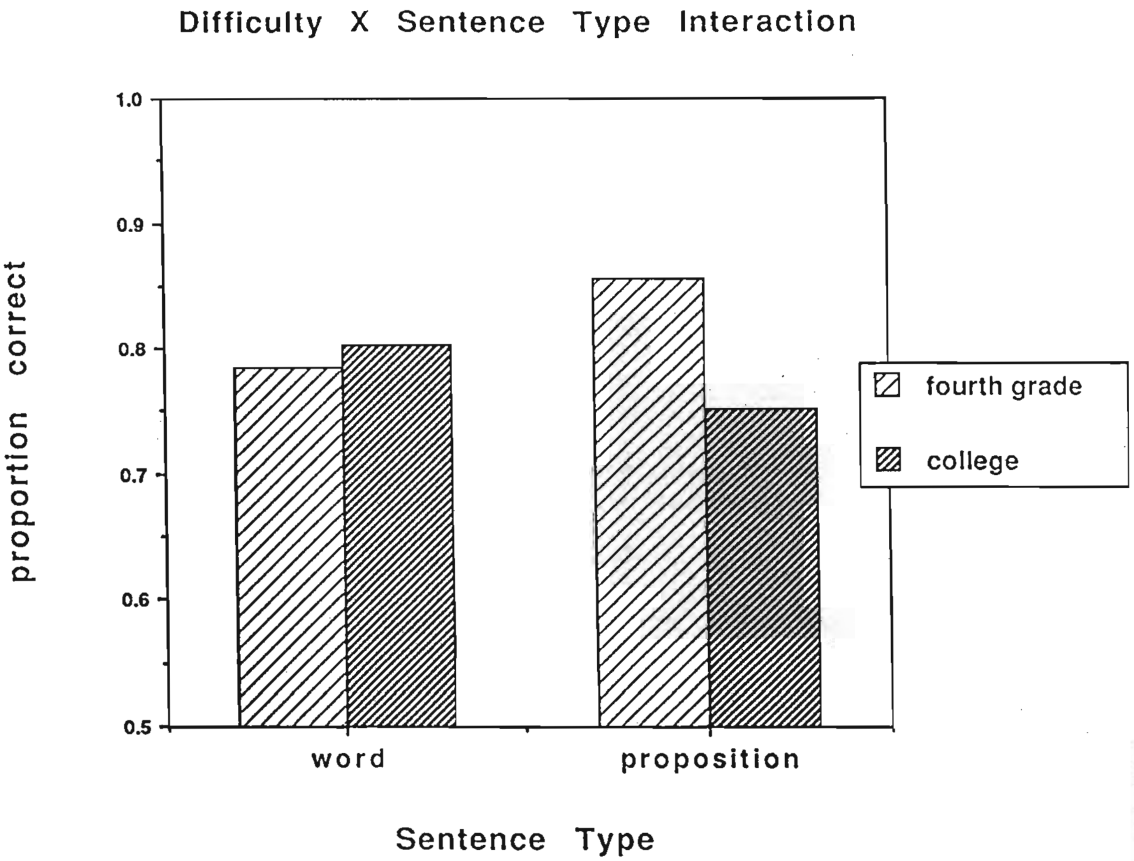


Figure 2. The interaction between voice type and sentence type observed in the recognition data.

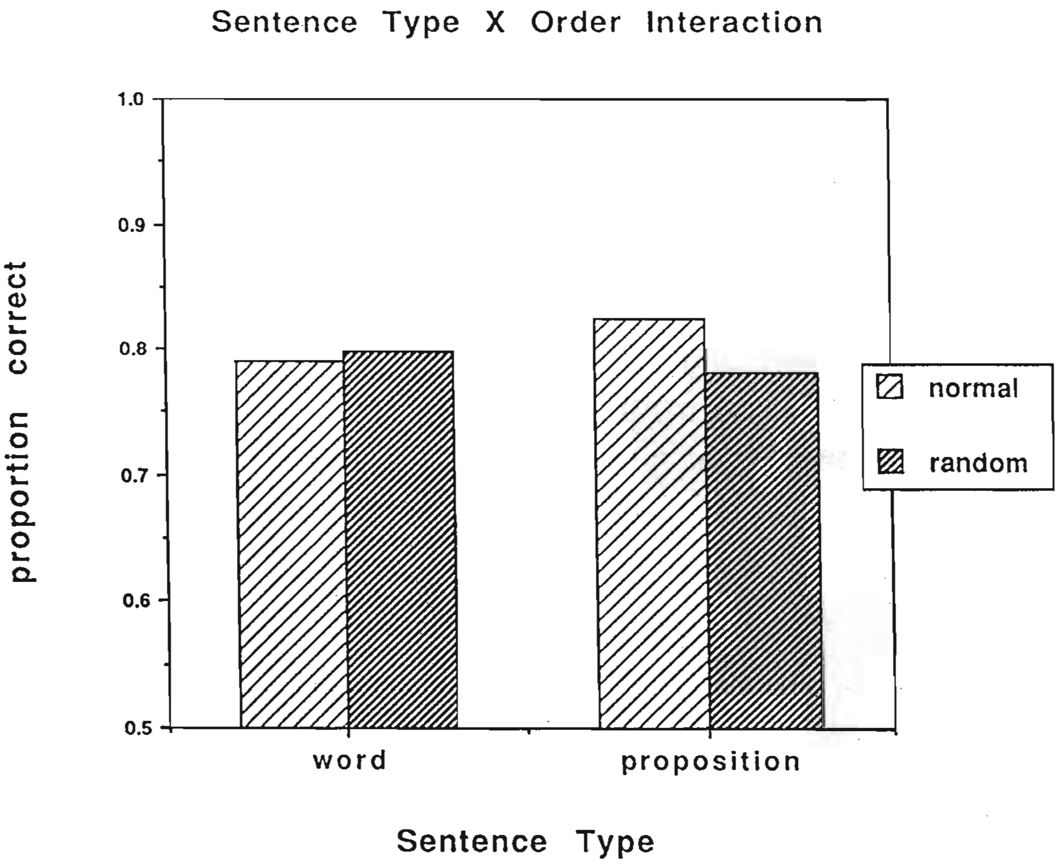


Figure 3. The interaction between sentence order and sentence type observed in the recognition data

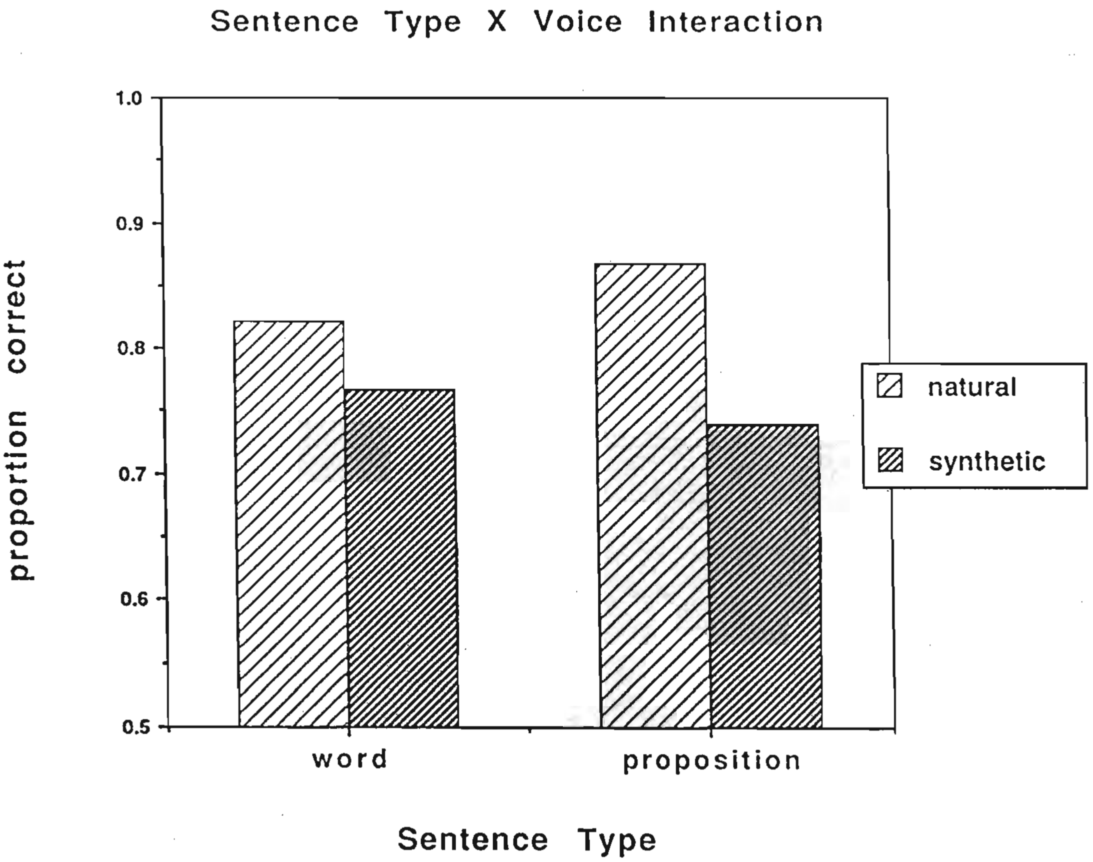


Figure 4. The interaction between passage difficulty and sentence type observed in the recognition data

Discussion

The results of the present experiment replicate and extend the earlier findings of Ralston et al. (1991). They also validate the sentence-by-sentence listening time task as a means of studying on-line language comprehension. Listening times were sensitive to voice and passage difficulty and were marginally sensitive to sentence order. Subjects who listened to synthetic speech had longer intersentence intervals in the sentence-by-

sentence listening time task and performed more poorly in the postperceptual recognition memory test. Listening times were also slower when listening to difficult passages. The effects of randomizing the sentences within each passage, however, were mixed. Listening times were not significantly longer for subjects who listened to the scrambled passages. Overall recognition memory was also not significantly worse. Order did, however, interact with voice in the listening time data and with sentence type in the recognition memory data. This may be taken as tentative evidence that scrambling the passages interfered with comprehension in different ways for listeners who heard natural speech and listeners who heard synthetic speech. In addition to the sentence order by recognition sentence type interaction, the interaction between sentence type and voice was also significant, replicating Luce's (1981) and Ralston et al.'s (1991) earlier results. Subjects who listened to synthetic speech had higher word recognition accuracy scores than proposition recognition accuracy scores. Subjects who listened to natural speech, in contrast, had better proposition recognition accuracy.

The results of the present experiment are interesting in three respects. First, they replicate and extend the earlier results of Ralston et al. (1991). Second, the results are methodologically interesting in that they further validate the use of the sentence-by-sentence listening time task as an on-line means of evaluating comprehension processes. The task was still sensitive to differences in the difficulty of the passages, despite the fact that none of the passages had a coherent structure in the scrambled condition. The main effect of difficulty in the listening time data may be attributed to the richer vocabulary and longer, more complex sentence structure used in the difficult passages. Thus, the listening time task appears to be sensitive to effects that are present at least at the sentence level.

The most interesting results of the present experiment, however, were the interaction of sentence order with voice in the listening time data and the interaction of order with sentence type in the recognition memory data. The longer listening times in the synthetic, randomly ordered passage condition can be accounted for by comparing this condition to the remaining three conditions and examining possible differences in attentional focus across the groups of listeners. First, consider the attentional focus of listeners who heard normally-ordered, natural passages and normally-ordered, synthetic passages. This condition was a replication of Ralston et al.'s (1991) second experiment. In both experiments, intersentence intervals were significantly longer for subjects who listened to synthetic speech. Ralston et al. accounted for this difference by suggesting that the listeners in the different groups attended to different levels of the signal. Listeners who heard natural speech may have focused their attention at the message level. Low-level word recognition and integration processes were assumed to require little attentional effort. Hence, listeners could focus most of their limited attentional capacity on the comprehension task. Listeners who heard synthetic speech, in contrast, may have focused their attention more closely on the acoustic-phonetic input. This shift in attention draws resources away from the message level of the text and causes the comprehension process to proceed more slowly. Reduced attention to the meaning may have also produced a more fragile end product of the comprehension process. This difference in on-line performance helps predict the observed postperceptual results. Subjects who heard natural speech demonstrated better recognition memory performance in both studies. Additionally, an interaction was found in both studies which indicated that subjects in the synthetic speech condition had better word recognition performance while subjects who heard natural speech demonstrated better proposition recognition performance. This result supports the claim that subjects distribute their limited attentional resources in different ways for natural and synthetic speech.

An analogous explanation can be offered for the differences observed in the random-order passage condition. Instead of considering global, passage comprehension, however, comprehension on a sentence by sentence basis may be more appropriate. As in the normally-ordered condition, subjects who listened to natural speech had shorter listening times than subjects who listened to synthetic speech. Again, the allocation of limited attentional resources can be assumed to be different. Subjects who listened to natural speech may have dedicated more processing resources to extracting the meaning from each sentence than to low-level encoding. One way of accounting for the small response time advantage of the random-ordered passages over the normally-ordered passages in the natural speech condition is to suggest that the difference is due to an integration of new information into an existing text frame. In the normally-ordered passages, subjects may have allotted extra processing time to integrate new information into the developing frame of the passage. In the random passages, however, subjects may have bypassed this integration stage because they lacked a coherent frame for the new information.

A similar argument can be made for subjects who listened to the synthetic random passages. As in the normally-ordered passage condition, subjects may have allocated more attentional resources to the acoustic-phonetic input. In the normally-ordered condition, subjects might have been able to use previously derived contextual information to facilitate the recognition of words and propositions in the current sentences. This implies that subjects were using global information to enhance processing of local information. No such global information was available to subjects in the randomly ordered passages. Thus, subjects may have had to rely strictly on local, acoustic-phonetic information. This would lead to a large increase in intersentence intervals for subjects who listened to randomly-ordered, poor-quality synthetic speech passages.

In addition to accounting for the voice by sentence order interaction in the listening time data, two interactions in the recognition memory data are also worth considering here. The interactions between sentence order and sentence type and sentence type and difficulty can be explained by the same basic attentional mechanism. In both interactions, word recognition questions were responded to more accurately than proposition questions. Differences in performance on word recognition sentences and proposition recognition sentences were small in the normally ordered passages and in the fourth grade passages. Differences between the two recognition measures were larger in the randomly ordered passages and the college level passages. One explanation for this finding is that word recognition performance may be independent of the difficulty of the text. The recognition of individual words falls at a low-level of the comprehension process. Proposition recognition, in contrast, relies on higher levels of representation involving word recognition, the integration of context, etc. When this high-level representation becomes more fragile or more difficult to derive, performance deficits may be anticipated.

The lack of a significant difference in the recognition memory data between the normally-ordered and the randomly-ordered passages also deserves some comment. At least two possibilities exist for why the order effect failed to reach significance. First, the passages may not have been difficult enough for the randomization to be truly effective. Subjects may have been able to derive a basic structure for the text in the random order condition. In this case, subjects would not be relying strictly on their memory for isolated sentences in the passages. Second, each of the recognition memory questions could be answered on the basis of one sentence information. No inferences across sentences of the passages were required to answer the questions. Thus, even without a basic frame for the text, subjects may have used the richness of the recognition sentence to locate isolated information in memory. These two explanations of the null result suggest that more difficult passages and more inference-oriented questions, with less associated context, may be needed to sensitively address issues of text coherence.

In summary, the present results extended the findings of Ralston et al.'s (1991) second experiment. Sentence-by-sentence listening times were longer for subjects who listened to passages of poor quality synthetic speech. Recognition memory performance for synthetic speech was also reduced relative to natural speech. While the main effect of sentence order was not significant in the on-line or postperceptual measures, sentence order did interact with voice in the listening time data and with sentence type in the recognition memory data. Differences in the allocation of attentional resources were given as an account for why subjects who listened to synthetic speech displayed longer listening times and reduced recognition memory performance relative to those who listened to natural speech.

References

Britton, B.K., Meyer, B.J.F., Hodge, M.H., Glynn, S.M. (1980). Effects of organization of text on memory: Text retrieval and response criterion hypotheses. *Journal of Experimental Psychology: Human Learning and Memory*, **6**(5), 620-629.

Britton, B.K., Meyer, B.J.F., Simpson, R., Holdredge, T.S., & Curry, C. (1979). Effects of the organization of text on memory: Tests of two implications of a selective attention hypothesis. *Journal of Experimental Psychology: Human Learning and Memory*, **5**(3), 496-506.

Bruner, H. & Pisoni, D.B. (1982). Some effects of perceptual load on spoken text comprehension. *Journal of Verbal Learning and Verbal Behavior*, **21**, 186-195.

Haviland, S.E. & Clark, H.H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, **13**, 512-521.

Jenkins, J.J. & Franklin, L.D. (1981). Recall of passages of synthetic speech. Paper presented at the 21st Psychonomic Society Meeting. Philadelphia, PA.

Kieras, D.E. (1978). Good and bad structure in simple paragraphs: Effects on apparent theme, reading time, and recall. *Journal of Verbal Learning and Verbal Behavior*, **17**, 13-28.

Kintsch, W. & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363-394.

Logan, J.S., Greene, B.G., & Pisoni, D.B. (1989). Intelligibility of eight text-to-speech systems. *Journal of the Acoustical Society of America*, **86**, 566-581.

Luce, P.A. (1981). Comprehension of fluent synthetic speech produced by rule. *Research on Speech Perception Progress Report No. 7*. Bloomington, IN: Speech Research Laboratory, Indiana University.

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

Meyer, B.J.F. (1975). *The Structure of Prose and Its Effects on Memory*. Amsterdam: Holland.

Mimmack, P.C. (1982). Sentence-by-sentence listening times for spoken passages: Text structure and listeners' goals. *M.A. Thesis*, Indiana University.

Moody, T.S. & Joost, M.G. (1986). Synthesized speech, digitized speech and recorded speech: A comparison of listener comprehension rates. *Proceedings of the Voice Input/Output Society*. Alexandria, VA.

Pauk, W. (1983). Six-way paragraphs. Middle Level. Providence, RI: Jamestown.

Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G., & Mullennix, J.W. (1991, in press). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*, **33**(4).

Novel Scientific Evidence of Intoxication: Acoustic Analysis of Voice Recordings from the Exxon Valdez[1]

J. Alexander Tanford[2], David B. Pisoni and Keith A. Johnson[3]

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

[1] Part of this article reports original research conducted under the direction of the second and third authors. The initial research was supported by a contract to Indiana University from General Motors Research Laboratories. The specific analysis of voice recordings of Captain Joseph Hazelwood was conducted by them at the request of the National Transportation Safety Board, and is based on tapes and data supplied by the NTSB. The second author may be called as a witness in some of the lawsuits pending against the Exxon Corporation. The opinions expressed in this article concerning whether this evidence meets the legal standards of reliability and admissibility are those of the first author, who is not affiliated with the Speech Research Laboratory and has not participated in either the initial research nor the analysis of the Exxon Valdez tapes.

[2] School of Law, Indiana University, Bloomington, IN 47405.

[3] Now at Department of Linguistics, UCLA, Los Angeles, CA 90024.

Novel Scientific Evidence of Intoxication: Acoustic Analysis of Voice Recordings from the Exxon Valdez

On March 24, 1989, the oil tanker Exxon Valdez ran aground in Prince William Sound, causing the worst accidental oil spill in history. Captain Joseph J. Hazelwood was in command, although he was not on the bridge at the time of the accident. Almost immediately, there were rumors and allegations that Hazelwood was intoxicated, but no proof. An oil tanker at sea is a long way from the nearest police officer with a breathalyzer. Hazelwood denied being intoxicated, and an Alaska jury found him not guilty of the charge. There seems to be no way to prove whether or not Hazelwood really had been drinking. Or is there?

Recent work in the Speech Research Laboratory at Indiana University's Psychology Department suggests that intoxication can be detected by acoustic-phonetic analyses of a suspect's voice (Pisoni & Martin, 1989; Johnson, Pisoni, & Bernacki, 1990). It turns out that tape recordings of conversations between Captain Hazelwood and the Coast Guard from before, during, and after the accident were available. Copies of these voice recordings were sent to us by the staff of the National Transportation Safety Board (NTSB) charged with investigating the Exxon Valdez accident.[4] Analyses of these tapes suggests that Hazelwood was indeed intoxicated when his ship ran aground.[5] As far as we know, this is the first real case use of this novel application of speech science techniques for measuring the effects of alcohol on speech.

Our analyses raise several obvious questions: Are the results reliable? How certain can we be that Captain Hazelwood was intoxicated? And, if this new testing procedure can indeed determine that someone was under the influence of alcohol based on voice recordings, are the results admissible in court?

The answers to these questions have implications beyond their obvious application to determining fault in the more than three hundred pending lawsuits concerning the Alaska oil spill.[6] In many cases, defendants may be far away from the nearest breathalyzer or may refuse to submit to blood-alcohol tests. If courts lack reliable, objective evidence of whether defendants had been drinking, they must rely on presumptions of guilt based on refusal to take a test, admit the speculative opinions of witnesses, or let the jury guess based on listening to tape recordings.[7] If a new measurement procedure exists that can, at least in some cases, provide objective indications of a person's physical state or condition, then courts could more reliably convict the guilty and acquit the innocent in cases where no chemical blood test results exist.

[4] Cf. New York Times v. National Aeronautics and Space Administration (1990). Voice recordings of the deceased Challenger crew were held within a Freedom of Information Act exemption; their release could unfairly invade a right of privacy.

[5] Captain Hazelwood denies that he was intoxicated. His attorney, Michael G. Chalos, has called the analytical method discussed here "voodoo stuff" (Bishop, K. Leaps of Science Create Quandaries on Evidence, *New York Times*, April 6, 1990).

[6] As of March 13, 1991, over three hundred lawsuits against Exxon were pending that are not affected by Exxon's settlement with the Environmental Protection Agency (*New York Times*, March 13, 1991, A1, col. 6).

[7] See Pennsylvania v. Muniz (1990) in which the jury heard video and audio tapes and testimony about field sobriety tests, and South Dakota v. Neville, (1983) in which the court laments carnage on highways and approves admissibility of refusal to submit to blood-alcohol test.

In this article, we describe this new testing procedure, using the analyses performed on Captain Hazelwood's voice as an example. We then discuss whether the results should be admissible under the rules governing novel scientific evidence. We conclude that the kinds of acoustic-phonetic analyses described in this article produce reliable and relevant evidence that should be admitted when supported by proper expert testimony. We do not claim that speech analyses conclusively prove that Captain Hazelwood was intoxicated when the Exxon Valdez ran aground. Rather, we believe that the analyses of Hazelwood's voice produce data consistent with intoxication. It therefore should be admitted into evidence and considered by the jury along with other relevant information.

The Scientific Evidence: Determining Intoxication from Voice Recordings

Alcohol is generally considered to be a central nervous system depressant. Significant blood concentrations of alcohol have been found to impair coordination, reflexes and nerve transmissions (see Berry & Pentreath, 1980; Pisoni & Martin, 1989). This kind of loss of motor control would seem naturally to affect speech. Indeed, controlled laboratory studies have demonstrated that alcohol produces three kinds of changes in speech production: gross effects, segmental effects and suprasegmental effects. Examples of each are listed in Table 1.

Insert Table 1 about here

Gross effects in speech production involve the alteration of entire words. A talker's ability to retrieve from memory and utter the proper sequence of words is impaired by alcohol. A talker who has consumed alcohol may revise, omit or interject words, sounds, or phrases (see Sobell & Sobell, 1972; Sobell, Sobell & Coleman, 1982).[8] One common example of a gross effect is the reversal of two words in a sentence (a spoonerism), such as "*work* is the curse of the *drink*ing class" (Borden & Harris, 1984).

Segmental effects, on the other hand, involve the misarticulation of specific speech sounds, notably the phonemes /r/, /l/, /s/ and /ts/. These mispronunciations are easily detected in spontaneous speech if one can compare examples of a person's intoxicated speech to that person's speech while sober (see Lester & Skousen, 1974; Trojan & Kryspin-Exner, 1968; Pisoni, Yuchtman, & Hathaway, 1986). Although some segmental effects may accompany any kind of loss of motor control,[9] the substitution of /š/ for /s/ seems to be unique to loss of control caused by alcohol (see Johnson et al., 1990).

Finally, suprasegmental effects involve changes in the duration, pitch and amplitude of speech. Intoxicated talkers speak more slowly, often use a lower mean pitch, and display greater pitch variation than when they are sober (Pisoni & Martin, 1989).[10] When these changes in speaking rate and pitch can be quantified, they are usually the most salient indication of alcohol impairment. Differences in rate

[8] Gross effects may be hard to recognize in spontaneous speech because the speaker's intended utterance is unknown.

[9] For example, common speech errors include consonant reversals, such as "a two-sen pet" instead of "a two-pen set", and vowel reversals such as "fool the pill" instead of "fill the pool" (Borden & Harris, 1984).

[10] Pisoni and Martin (1989) found no significant effects on fundamental frequency at moderately high levels of intoxication (0.10% BAL). (Cf. Sobell, Sobell & Coleman, 1982). Trojan and Kryspin-Exner (1968) report that effects on pitch varied.

Table 1

Effects of alcohol on speech.

Gross Effects	Syllable/word/phrase interjection Word omission Word revision Broken suffixes
Segmental Effects	Misarticulation of /r/ and /l/ /s/ becomes /š/ Final devoicing (e.g., /iz/ -> /is/) Deaffrication (e.g., "church" -> "shurch")
Suprasegmental Effects	Reduced speaking rate Decreased amplitude Mean change in pitch range Increase in pitch variability

and pitch can be measured objectively using conventional digital signal processing techniques,[11] and then easily compared to similar measurements made from samples of a talker's sober speech (see Pisoni & Martin, 1989; Klingholz, Pening, & Liebhart, 1988; Sobell, et al., 1982). Two common suprasegmental effects are vowel lengthening (Pisoni & Martin, 1989) and lengthening of consonants in unstressed syllables (Lester & Skousen, 1974).

In trying to determine if Captain Hazelwood had consumed alcohol, we applied speech analysis techniques to five samples of his speech from audio tapes provided by the NTSB. According to the NTSB, these recordings were made thirty-three hours before the accident, one hour before the accident, immediately after the accident, one hour later, and nine hours after the accident. All taped communications contained the phrase "Exxon Valdez." This utterance was the focus of our quantitative analyses because it was the same phrase used several times, and it contained tokens of the fricative consonant /s/.

We first looked for gross effects. Gross effects usually are difficult to measure in spontaneous speech, because the listener does not know what word was intended by the speaker. Nevertheless, we found several examples of gross errors where the intended word was obvious, or the speaker himself corrected the mistake. These are summarized in Table 2.

Insert Table 2 about here

We also found several segmental errors in Captain Hazelwood's speech. Several misarticulations of /r/ and /l/ in the words "northerly," "little," "drizzle," and "visibility" occurred in the recordings made near the time of the accident. We paid particular attention to the /s/ sounds in the phrase "Exxon Valdez." Although Hazelwood pronounced Exxon correctly the day before the accident, he misarticulated it as "ek*sh*on" around the time of the accident. This subjective perception was confirmed by spectral analysis.[12] We also found several examples of final devoicing, as /z/ in Valdez became an /s/ ("valde*s*"). These findings suggest that at the time his ship ran aground, Captain Hazelwood was having difficulty with the fine motor control used to produce these sounds, but it still does not prove that alcohol was the cause.

If the loss of motor control is due to alcohol consumption, the most salient indication of intoxication will usually comprise suprasegmental effects. We focused on speaking rate and voice fundamental frequency.[13] We measured the duration of the speech segments in the phrase "Exxon Valdez" in each sample, taking the average of at least two occurrences of the phrase in each time period. The results, summarized in Figure 1, show that it took Captain Hazelwood approximately 50% longer to say the phrase "Exxon Valdez" at the time of the accident than the day before.

[11] Standard operating procedures were used. The voice recordings were low-pass filtered at 9.6 kHz and digitized at a 20-kHz sampling rate through a 12-bit analog-to-digital (A/D) converter. A digital waveform editor was used with a PDP 11/34 minicomputer to edit all speech samples into separate digital files (see Pisoni & Martin, 1989).

[12] Spectral analysis techniques are commonly used to measure the distribution of energy at different frequencies as a function of time. The results are often displayed as speech spectrograms, or "voiceprints," in order to reveal the dynamic time-varying nature of speech (see Flanagan, 1972).

[13] Because the communication equipment had automatic gain controls and the distance between the speaker and the microphone probably varied, we could not measure changes in the amplitude of the speech.

Table 2

Summary of gross effects found in the NTSB tape.

First Word Used	Revision
1. Exxon Ba...	Exxon Valdez
2. Departed	Disembarked
3. I	We'll
4. Columbia Gla	Columbia Bay

Insert Figure 1 about here

Speech theory predicts that changes in duration will be most noticeable in vowel segments (see Pisoni & Martin, 1989) and unstressed consonant segments (see Lester & Skousen, 1974). Looking at the "e" and "on" in Exxon and the initial "v" in Valdez, the slowed speech is particularly apparent. These results are summarized in Figure 2.

Insert Figure 2 about here

Finally, we calculated voice fundamental frequency across the phrase "Exxon Valdez." For each of the five relevant time periods, we measured the pitch of all four vowel sounds in two productions of "Exxon Valdez", and averaged the eight measurements. Voice pitch was dramatically lower in the samples recorded around the time of the accident. Variability of fundamental frequency was correspondingly greater in the voice samples taken at the time of the accident. These data are summarized in Figures 3 and 4. Lowered pitch and increased variability would be expected in alcohol-impaired speech, so this evidence also suggests that Captain Hazelwood had been consuming alcohol at the time his ship ran aground.

Insert Figures 3 and 4 about here

The acoustic-phonetic differences found in the speech samples supplied by the NTSB are consistent with the findings of controlled laboratory studies on the effects of alcohol on speech (see Klingholz et al., 1988; Lester & Skousen, 1974; Pisoni et al., 1986; Pisoni & Martin, 1989; Sobell & Sobell, 1972; Sobell et al., 1982; Trojan & Kryspin-Exner, 1968). However, they do not *prove* that Hazelwood was drunk. This is not even strong evidence that he was intoxicated unless other obvious explanations for the pattern of changes can be discounted. Speech is affected by the physical state of the speaker, especially by stress and fatigue. Measurements made from tape recordings can be affected by malfunctions or variations in the recording equipment. In the Hazelwood case, however, these alternative explanations can be discounted.

The Physical State of the Speaker

Joseph Hazelwood would undoubtedly have been under some psychological stress following the Exxon Valdez accident, and would logically have become increasingly fatigued as the hours wore on. However, neither of these two factors appears to explain the specific changes in his speech found on the tapes.

Stress might be expected to affect Hazelwood's speech immediately after, one hour after and nine hours after the accident, but it could hardly have affected his speech one hour *before* the accident occurred. Yet, both segmental and suprasegmental effects are just as robust one hour before the accident as one hour afterwards. In addition, previous studies on speech production in stressful environments show that stress affects speech in the *opposite* direction from alcohol. Stress causes pitch and speaking rate to increase rather than decrease (see Brenner & Shipp, 1988). Psychological states similar to stress, such as fear and anger, likewise produce increases in pitch and rate (see Hansen, 1988; Williams & Stevens, 1972). Alcohol, on the other hand, causes fundamental frequency (pitch) and speaking rate to decrease. In Hazelwood's case, they decreased.

Fatigue may cause changes in speech production which are similar to the changes caused by intoxication. Surprisingly, speech scientists have conducted little controlled research on fatigue effects. In the absence of scientific data concerning the effects of fatigue on speech production, it is reasonable

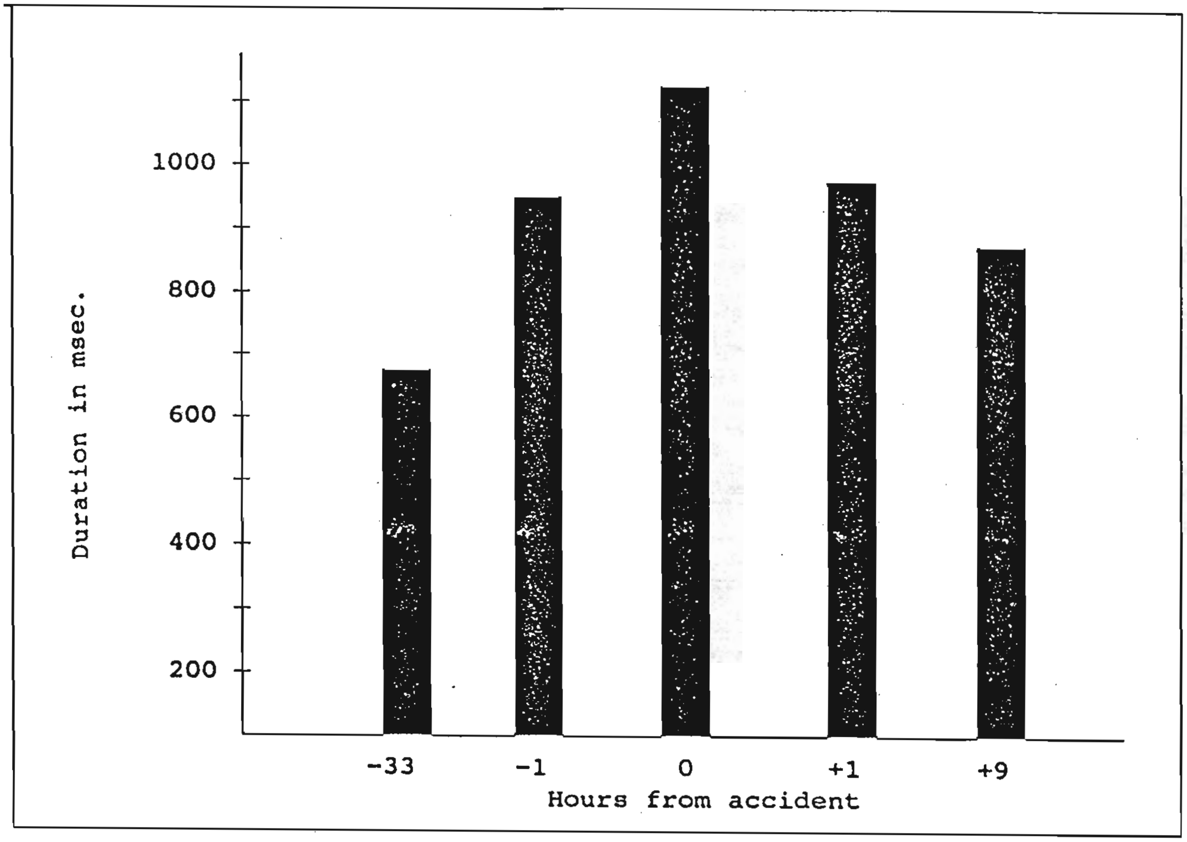


Figure 1. Cummulative duration of speech segments in "Exxon Valdez".

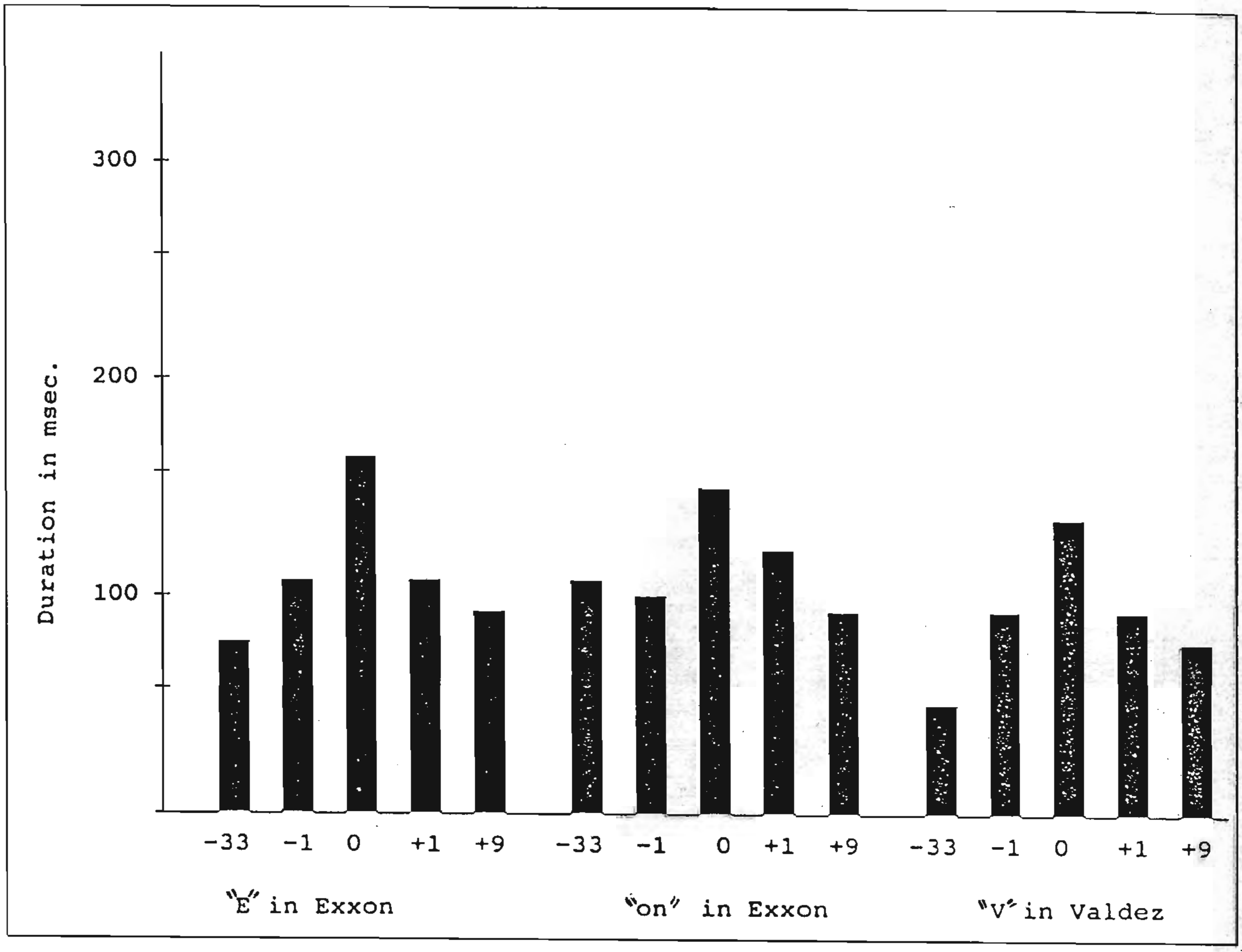


Figure 2. Duration of vowel sounds and consonants in unstressed syllables in "Exxon Valdez".

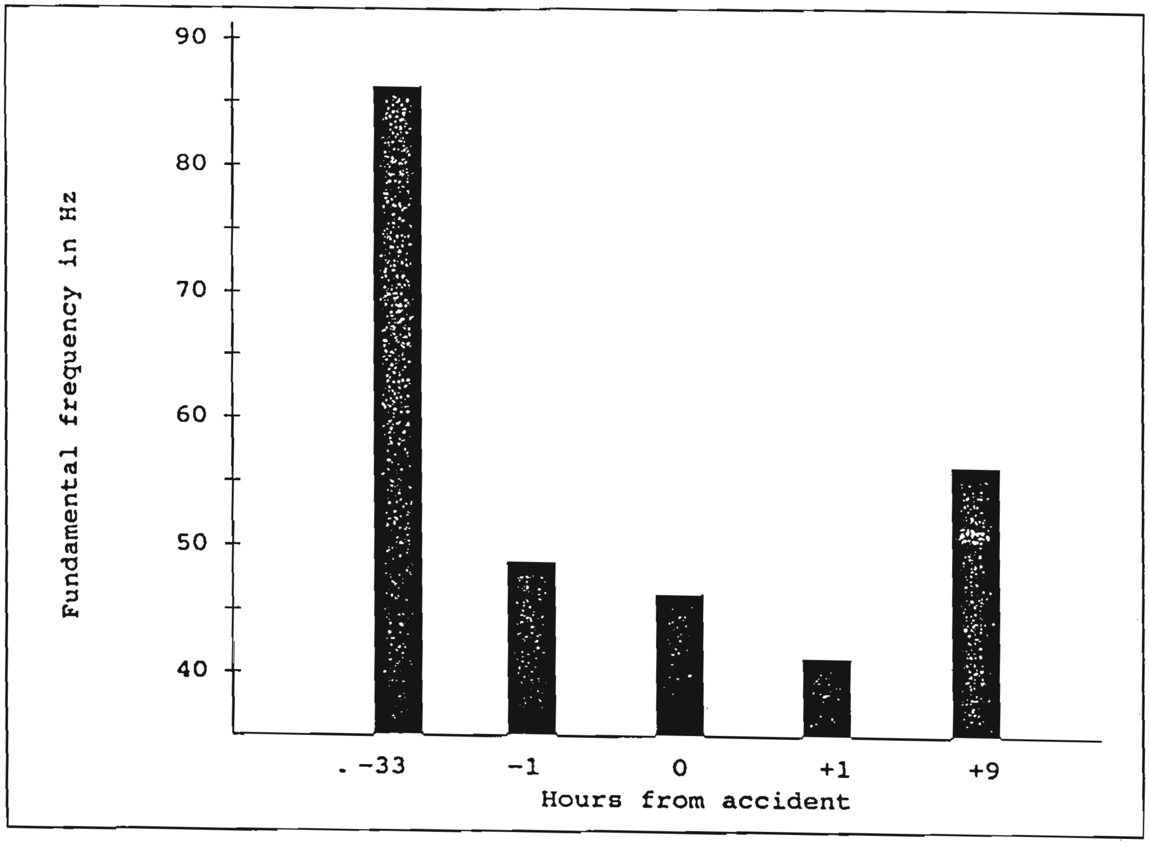


Figure 3. Average pitch of voice in "Exxon Valdez".

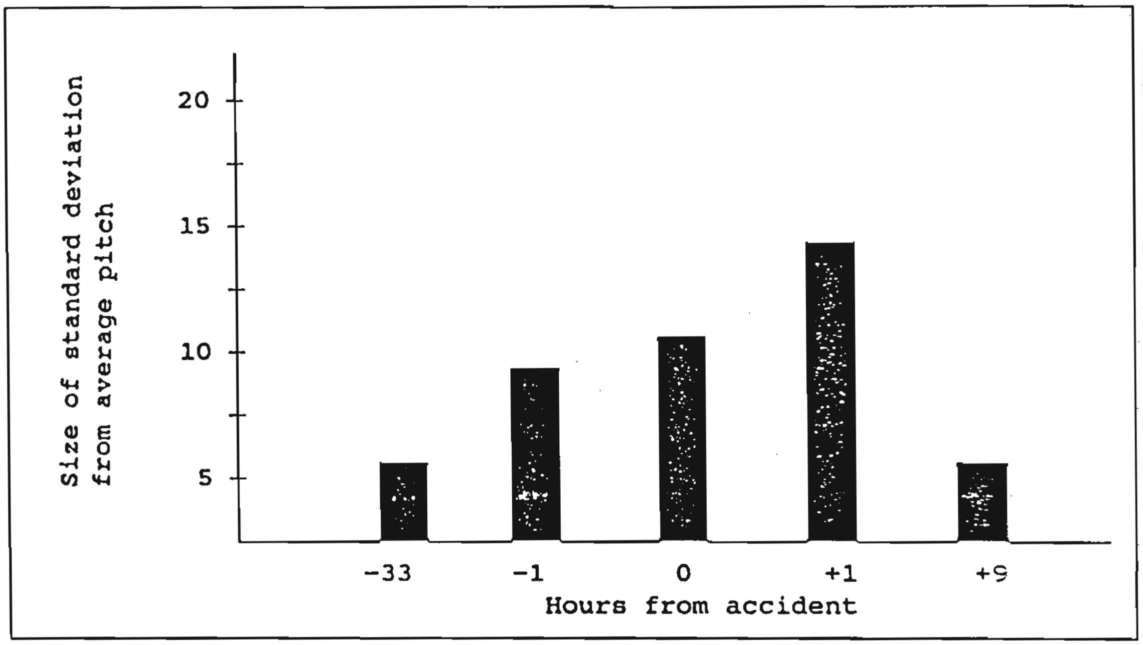


Figure 4. Variation in pitch of voice in "Exxon Valdez".

(and conservative) to assume that fatigue produces effects which are similar to the effects of intoxication: a general lowering of arousal, slower speaking rate, and lower voice pitch. It also seems reasonable to assume that a person's level of fatigue increases over time without sleep.

Given these assumptions, the pattern of phonetic changes seen in Captain Hazelwood's speech cannot be attributed to simple fatigue. He spoke more slowly and with lower fundamental frequency at the time of the accident than nine hours later. Indeed, his speech nine hours after the accident was similar to his speech the day before. If the changes could be attributed to fatigue, one would expect that fatigue-induced effects would be greatest nine hours after the accident, by which time he had been active without sleep for twenty-two hours. This expectation was not borne out by the data.

The Mechanical State of the Recordings

Measurements of pitch and duration of sounds obviously are sensitive to fluctuations in the speed of tape recordings. If an audio tape is played at a slower speed, the voice will sound lower in pitch and speech sounds will be longer. These effects are similar to those caused by alcohol impairment. To some extent, the risk of erratic tape speeds is reduced as long as both the control recording (in which the speaker is presumed sober) and the test recording (in which the state of the speaker is in question) were recorded and played on the same equipment. The test is a comparative one that does not depend on the *true* pitch and duration being known.

A more reliable control is to measure some sound that appears on both tapes other than the suspect's voice. In the Exxon Valdez case, for example, Captain Hazelwood was talking to the same Coast Guard radio operator on the various tapes. Hazelwood's voice pitch and duration changed while the radio operator's did not. It is also possible to measure the average pitch of background noise on two tapes. This was done for the Valdez tapes. Again, the background noise did not significantly change, while Hazelwood's voice did. These observations reduce the likelihood that the effects measured were the result of mechanical problems in the audio tape recording.

Other Factors

Research on environmental and emotional effects on speech production has demonstrated that other factors also can cause suprasegmental effects, similar to alcohol. In a noisy environment speech tends to be produced with increased fundamental frequency variability and at a slower rate, but tends to have a *higher*, rather than lower, fundamental frequency (see Hansen, 1988; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988). In another experiment, acceleration and vibration affected fundamental frequency and duration, but in the opposite direction from alcohol (Moore & Bond, 1987).[14] Several studies on mental workload with high cognitive demands (e.g., airline pilots, air traffic controllers) indicate that speech affected by performing a cognitively demanding task will display suprasegmental effects opposite to those observed with alcohol -- higher average frequency and shorter duration.[15] Sorrow produces lower average frequency and slower speech, but seems to cause less, rather than more, pitch variability (see Hansen, 1988; Williams & Stevens, 1972).

In short, there are no known environmental situations or emotional states that produce quite the same pattern of suprasegmental effects as alcohol impairment. These observations mean that *in theory*, it is possible to classify changes observed across two samples of speech as more like the pattern found for alcohol-affected speech than for any other probable cause of impairment. Three problems exist, however.

[14] Moore and Bond (1987) is preliminary at best; only two subjects were tested.

[15] The effects of alcohol on frequency variability are mixed (see Griffin & Williams, 1987; Hansen, 1988).

First, it is not possible to give any kind of confidence rating to such a classification. There is not enough published laboratory data on individual differences which would allow the calculation of hit rates and false alarm rates for classifications based on these acoustic measures.[16] Second, there are some possible physiological effects on speech production which have not yet been adequately studied, such as fatigue, illness and being suddenly awakened. Any of these might produce effects similar to those measured after alcohol consumption, or they might not. Third, no data have been gathered in more complex environments involving combinations of these stimuli.

Scientific Conclusion

The ultimate conclusion that can be drawn from our data is this: Analyses of audio tapes of Captain Hazelwood's speech cannot prove that he was alcohol-impaired at the time of the Exxon Valdez accident. However, our analyses provide objective tests for which of several explanations of his physical state is the most likely. For example, if phonetic analyses revealed faster speech and higher fundamental frequency, we could attribute those changes to increased arousal caused by stress or anger, and eliminate intoxication as an explanation for the changes in Hazelwood's speech. As it turns out, no such alternative explanation fits the observed pattern of changes in a simple way.

Analyses show a pattern of changes consistent with the conclusion that Captain Hazelwood had been drinking. His speech around the time of the accident is characterized by misarticulations of /r/ and /l/, changes from /s/ to /š/, final devoicing of /iz/ to /is/, reduced speaking rate, lower fundamental frequency, increased pitch variability, and a number of word and syllable revisions, compared to his speech thirty-three hours earlier or nine hours later. This pattern of segmental and suprasegmental effects is consistent with alcohol-impaired speech measured in a controlled laboratory environment. It is not consistent with patterns of speech affected by fear, anger, noise, acceleration, vibration, or increased mental workload. The pattern is partially consistent with speech affected by stress or sorrow, but these changes were observed on the audio tape made one hour *before* the accident, when Hazelwood would probably not yet have been experiencing either stress or sorrow. The effects probably were not caused by mechanical problems affecting tape speed. Our analyses cannot rule out the possibility that Hazelwood's speech was affected in whole or part by fatigue, although logic and general theories of speech motor control suggest this is a less likely explanation than alcohol consumption. Is this scientific evidence admissible in court?

The Legal Standard for Admitting Novel Scientific Evidence

The problem of separating reliable scientific evidence from quackery is not new. For centuries, courts have wrestled with the question of whether to admit the testimony of expert witnesses, not always reaching the right decision.[17] But at least courts agreed on the procedural rule by which science was to be measured: the "general acceptance" test announced in Frye v. United States (1923). Within the

[16] For example, it is not possible to offer reliable probabilistic statements such as: "Hazelwood had this pattern, and 95% of people who exhibit this pattern are intoxicated and only 10% of fatigued speakers show this pattern."

[17] In A Trial of Witches (1665), two women of Leystoff, County of Suffolk, were accused of bewitching two children. The evidence was conflicting, so the court sought the opinion of "Dr. Brown of Norwich, a person of great knowledge" concerning witchcraft. Dr. Brown was clearly of the opinion that the victims had been bewitched, and explained the procedure whereby the devil excited victims' humours through pins inserted by witches. The fact that the form of bewitchment appeared to be natural, swooning fits he attributed to the villainous "subtilty [sic] of the devil." His testimony was believed, and the witches were convicted and executed.

last ten years, however, the effectiveness of the Frye test as a screening device for scientific evidence has come under increasing attack, led by Professor Paul Gianelli's (1980) watershed article.[18]

The Frye Test

The Frye test prohibits the courtroom use of scientific evidence until it has gained "general acceptance" in relevant scientific fields. The rule was announced in a case from the District of Columbia in which the appeals court decided that test results from a systolic-pressure measurement device designed to detect lies was inadmissible. The court stated, without precedent or citation:

> Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while the courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. (Frye v. United States, 1923, p.1014)

This vague language probably was intended only to justify the exclusion of "lie detector" test results. Its negative tone expresses the presumption that scientific evidence will be *excluded* rather than admitted. This sets the Frye test at odds with the most fundamental principle of evidence: all facts having rational probative value are presumed admissible (Wigmore, 1940). Nevertheless, this language has been widely adopted by courts as a general principle of exclusion. Courts have justified the adoption of a restrictive rule in part because of the fear that jurors are "easily overawed by conclusions voiced in court by articulate experts with impressive credentials" (Moenssens, Inbau, & Starr, 1986, p.6).

The Frye test is still the controlling law in some jurisdictions (e.g, United States v. Shorter, 1987) (see McCormick, 1984; The Frye Doctrine and Relevancy Approach Controversy, 1986; Petrosinelli, 1990). Its major weakness is that it fails to clearly distinguish between novel scientific evidence and reliable scientific evidence. The general acceptance test arises from a fear of science (Tanford & Tanford, 1988), and therefore tends to exclude all new science, whether or not it is reliable. Increasingly, courts and commentators have criticized the Frye test on three grounds (see McCormick, 1984; Moenssens et al., 1986).

1. The Frye test is bad science. The Frye test was supposed to help judges distinguish scientifically reliable evidence from nonsense. However, the assumption that general acceptance of a procedure is synonymous with scientific accuracy is dubious. This is especially true if courts rely on other court opinions to determine general acceptance. Testing procedures can be in widespread use in forensic laboratories, achieving the appearance of general acceptance, yet never be scientifically reliable. Two good examples are the once-popular paraffin test to determine if a suspect had recently fired a gun (Gianelli, 1980; Neufeld & Colman, 1990) and chirography (handwriting identification) (Risinger, Denbeaux, & Saks, 1989). Both have been routinely admitted by courts as generally accepted procedures; neither has a valid scientific basis.

The opposite is also true: some scientifically reliable testing procedures are not yet generally accepted. The Frye test produces an inevitable "cultural lag" problem (Malskos & Spielman, 1967).

[18] That his article has led the trend away from the Frye test is somewhat ironic, as Gianelli intended to criticize most of the deviations from the Frye test and urge that we preserve the spirit of the Frye test by developing a more workable version of that test (Gianelli, 1980).

Time must pass in order for a new procedure to gain general acceptance, even after it has become scientifically reliable. All new techniques have to go through a probationary period when the tests are producing scientifically reliable results but trials must be conducted without them (Gianelli, 1980). This delay until a procedure is generally accepted can be considerable, especially if a new scientific advance is perceived as threatening or too radical (and therefore, unacceptable) by established, older scientists in the field (Kuhn, 1970; Moenssens et al., 1986).

2. The Frye test is bad law. The Frye test is also bad law. Its terms are vague and ambiguous, and judges have difficulty applying it to novel scientific evidence.[19]

In order to apply the Frye doctrine, a judge must first identify the appropriate field. The term "field" is ambiguous. New scientific evidence may not be easily classifiable into a traditional field, or may overlap several fields (Petrosinelli, 1990). How the judge interprets the term may be dispositive. If the judge defines the field broadly, it virtually assures that many scientists in that broad field will not have heard of a new procedure. For example, People v. King (1968) involved the admissibility of spectrogram (voiceprint) evidence. The judge defined the relevant field as anatomy, physiology, physics, psychology and linguistics. The test results were excluded when the proponent could not prove general acceptance in all these disciplines.

There also is danger in defining the field too narrowly. In case of a dispute within a discipline, those who dissent from the conventional wisdom may form their own subfield, within which untested or unreliable evidence is generally accepted. For example, People v. Williams (1958) involved the admissibility of a controversial Nalline test for detecting narcotic use. The court defined the relevant field as "those who would be expected to be familiar with its use" (People v. Williams, 1958, p.251). The test results were therefore easily admitted, although the evidence showed that the medical profession generally had never heard of this supposed test (Gianelli, 1980; Moenssens et al., 1986).

Once the field is defined, the judge must decide if a testing procedure has been generally accepted. This language is obviously ambiguous. How many scientists must agree? How unanimous must the agreement be (Gianelli, 1980)? What is the effect of the opponent producing "experts" who dispute acceptance? Is any paid consultant's opposing testimony enough to disprove general acceptance? These issues have never been answered by courts interpreting the Frye test (Gianelli, 1980).

Even if general acceptance were more clearly defined, the Frye test does not set forth any specific foundation requirements. In the first place, it does not place any restrictions on who may be called as a witness. The general acceptance language could be interpreted as requiring a disinterested foundation witness *other* than a scientist who helped create the new technique.[20] Such a disinterested witness, however, is unlikely to know enough about the new technique to answer questions (especially on cross-examination) about the details of its operation. Witnesses who know most about the details of a particular test, however, are often not scientists, but technicians or "examiners," trained to use a piece of testing

[19] If a testing procedure has been around a long time and still not gained general acceptance (e.g., the polygraph), the Frye test is indeed one measure of the scientific unreliability of the test. If the testing procedure is new, its lack of general acceptance may be either because it is unreliable or just because it is new. In the case ofacoustic anaysis of speech, the methodology has been in general use in speech science for over forty years; only the particular application is new.

[20] See e.g., People v. Tobey (1977) about rejection of voiceprint evidence because an expert witness built a career on voiceprint and was not disinterested.

equipment, but unable to provide necessary information to determine its underlying scientific reliability (Gianelli, 1980; Moenssens et al., 1986).

It also is unclear whether a *single* witness can establish *general* acceptance. A few courts have held that more than one expert is required,[21] although this conflicts with the general evidence principle that a foundation may be laid by a single person unless a rule clearly requires corroboration.

Is general acceptance a matter merely of expert *opinion*, or is some minimal level of conditional fact required? For example, may a procedure ever be said to be generally accepted as scientifically valid based on a *single* validating study? The Frye test seems to assume that additional validating studies will be done over time -- otherwise there is no reason to require a probationary period for new scientific techniques. However, some courts have admitted voiceprint evidence on the basis of a single study from Michigan State University (Gianelli, 1980). If additional validating studies are required, how many are required? Must they be done by a second research team? Is publication in a peer-reviewed professional journal sufficient? Is any particular degree of significance required in the validating studies? Must the studies show that a test is 98% reliable? 90%? 75%?

The Frye test refers to a requirement that some "thing" have gained general acceptance (Frye, 1923, p.1014). It does not explain what that thing is. Although it seems obvious that the court was referring to the specific testing procedure at issue -- use of a device that measured systolic blood pressure changes to detect deception --in the same paragraph, the Frye court refers to scientific principles, discoveries, and deductions. It leaves open the issue of whether general acceptance must be proved for the underlying scientific theory,[22] the technology,[23] the procedure and technique,[24] the particular instrument used,[25] or some combination. This can make a substantial difference, because novel scientific evidence may involve a new theory, a new application of established theory, an improved procedure, or the use of a new instrument (Gianelli, 1980).

Finally, the Frye test is bad law because it requires judges to make decisions for which they are ill-equipped. Judges generally are illiterate in science, untrained in statistics, and operate in a legal culture that is non-scientific (if not actively hostile to science) (Tanford, 1990). Judges are therefore poorly equipped to distinguish generally reliable from unreliable methods. Nor are they likely to get any help from lawyers, who are similarly untrained in science and have difficulty accessing scientific debates in science journals (Neufeld & Colman, 1990).

3. The Frye test is premised on bad psychology. One underlying premise of the Frye test is that jurors will be easily overawed by expert testimony. Gianelli asserts that "an aura of scientific infallibility may shroud the evidence and thus lead the jury to accept it without critical scrutiny" (Gianelli, 1980, p.1237). Professor Lawrence Tribe (1971) makes the same argument about statistical evidence. In U.S. v. Addison (1974), the D.C. Court of Appeals asserts that scientific evidence may "assume a posture of mystic infallibility in the eyes of a jury of laymen" (see State v. Carlson, 1978).

[21] See e.g., People v. Kelly (1976) concerning rejection of voiceprint evidence and doubt whether a single witness can ever satisfy foundation.

[22] See e.g., United States v. Addison (1974) which suggests that theory must be generally accepted.

[23] See e.g., United States v. Stifel (1970) concerning the state of technology of neutron activation analysis.

[24] See e.g., People v. Law (1974) concerning reliability of the voiceprint procedure, and Reed v. State (1978) concerning reliability of the voiceprint technique and or process.

[25] See Commonwealth v. Fatalo (1963) concerning scientific acceptance of instrument -- polygraph.

This assertion about human behavior is dubious. Social psychologists have demonstrated that people generally *undervalue* scientific data, misunderstand and under-utilize statistics, rely on anecdotes and emotion rather than empirical scientific evidence when making important decisions, and persistently hold beliefs contrary to scientific logic and mathematics (Nisbett & Ross, 1980; Saks & Kidd, 1981; Thompson & Schumann, 1987; Thompson, 1990,).[26] Some studies focusing specifically on juries have found that expert witnesses have no significant impact on verdicts.

Alternative Legal Standards

In response to the weaknesses of the Frye test, commentators have suggested that it be replaced with either of two[27] alternatives: A modified Frye test or a relevancy test.

1. Modified Frye tests. One proposal would modify the "general acceptance" part of the Frye test. Commentators have suggested the substitution of "substantial" or "reasonable" acceptance.[28] This would probably have the effect of admitting more scientific evidence, but would still be ambiguous and difficult to apply. It also would be unlikely to more effectively distinguish reliable from unreliable *new* techniques.

A second proposal would modify the definition of "field." In People v. Williams (1958), the court stated that it would accept scientific evidence unknown in a general scientific field, if it were accepted as reliable within a narrow specialty--those who would be expected to be familiar with its use. This variation would naturally tend to admit more novel scientific evidence, but it fails to address the problem that even within a specialty field, acceptance is not the same thing as reliability.

The third proposal is to modify the Frye test's requirement that the technique be generally accepted. It would require only that the scientific principles underlying a new testing procedure be generally accepted. New testing procedures designed to explore a particular problem using generally accepted principles and existing equipment would be admissible, even if the particular application were new (Ibn-Tamas v. United States, 1979; Coppolino v. State, 1969).[29]

2. Relevancy tests. The trend, however, is to reject the Frye test rather than try to modify it. As early as 1954, leading scholars were calling for the substitution of a basic relevancy test. Any relevant scientific evidence supported by a qualified expert should be presumptively admissible (McCormick, 1954). Neither the evidence's newness nor its lack of general acceptance are dispositive, for neither criterion makes scientific evidence *relevant*.[30] The primary check against unreliable or pseudo-scientific evidence is the rule permitting the opponent to contradict the expert's testimony by introducing passages from leading textbooks in the field written by reliable authorities (see Federal Rules of Evidence 803(18)).

[26] Most subjects give *less* weight to statistical evidence than it deserves.

[27] A third alternative has also been suggested: delegating the decision on reliability to independent bodies of science advisors. Some have suggested the creation of a science court, others the use of independent bodies of experts hired to advise courts on validity of scientific techniques (see Gianelli, 1980; McCormick, 1984). The idea was incompatible with the adversary system (see Bazelon, 1977), and went nowhere.

[28] See Richardson (1974) on substantial acceptance. See Latin, Tannehill, & White (1976) on reasonable acceptance. See also, *Expert Testimony Based on Novel Scientific Techniques* (1980).

[29] Gianelli (1980) argues that the Coppolino test is essentially the same thing as McCormick's relevancy test.

[30] United States v. Stifel (1970) addresses the newness and lack of absolute certainty affect of weight, as opposed to admissibility: "Every useful new development must have its first day in court".

The first requirement for admissibility under a relevancy approach is that *scientific evidence must be introduced through a properly qualified expert.* All novel scientific evidence must be sponsored by an expert witness who can explain both the theoretical and practical reliability of the new testing procedure. Although the usual rule is that either education or experience qualifies a person as an expert (e.g., Federal Rules of Evidence 702), in this case a fully-educated scientist is probably required. Mere experience with a new forensic technique is not adequate to explain the theoretical validity of a procedure, so a technician probably cannot lay the necessary foundation (Gianelli, 1980; Moenssens et al., 1986).

The second requirement is that *the subject-matter must be one on which expert testimony will assist the jury.* Expert testimony generally is permitted only in appropriate situations in which the jury could use some help. For example, experts with breathalyzers and blood test results will be of real assistance to a jury in determining a person's blood-alcohol content, but traffic safety experts are probably not needed to help a jury determine if high-speed drunk driving is dangerous. The difference is whether the jury is presumed to be competent to draw their own conclusions based on observations.

There are three possible situations: (1) The data upon which conclusions are drawn may be beyond the jurors' perceptions, so jurors are incapable of drawing any conclusions about the subject, such as quantum mechanics or brain surgery. In this case, experts will always be permitted. (2) The data may consist of common subjective evaluations based on perception completely within common knowledge, such as whether a concrete block falls to the ground when dropped off a scaffold. No physicist is necessary to explain gravity. In this situation, experts are superfluous. (3) The subject may be somewhere in between. The data may be partially hidden, or it may be an area in which lay and expert witnesses use quite different methods for reaching similar kinds of conclusions, such as whether failure to water a lawn during a drought caused grass to die.

The old version of evidence law (still followed in a few states) once again presumed that scientific evidence should not be admitted. It permitted expert testimony only in the first situation -- when the matter at issue was completely beyond the understanding and common experience of the average juror (see McCormick, 1984; see also, State v. Maudlin, 1981). Thus, scientific evidence was excluded if the jurors and the expert were similarly qualified to draw conclusions. The rule did not recognize the possibility that jurors might draw *better* conclusions with the aid of experts.

The modern version of the rule reflects the contemporary view of presumptive admissibility. It states that scientific evidence is admissible if it will "assist" the jury to understand the evidence or draw conclusions (e.g., Federal Rules of Evidence 702). Under this version, evidence should be *admitted* if jurors and experts can both draw reliable conclusions. Indeed, it is under this modern view that testimony on field sobriety tests is admitted.[31]

The third requirement is that *all equipment involved in generating scientific evidence be shown to be in good working order.* Inherent in the creation of scientific evidence is reliance on machines, technology, and laboratory equipment. The reliability of an ultimate scientific conclusion may depend on factors the scientist takes for granted -- that ordinary pieces of apparatus were in good working order and were operated by qualified persons. The law requires proof that instruments and equipment, such as microphones, tape recorders, and X-ray machines, were in good working order and were properly operated during the creation of any particular piece of scientific evidence. It also requires that human

[31] See People v. Randolph (1989) and People v. Krueger (1968). Typical field sobriety tests are described in Pennsylvania v. Muniz (1990).

beings be accounted for. If an expert relies on lab technicians or graduate students, they must be shown to be reliable and properly trained. Some jurisdictions also require proof that reliance on these supporting procedures be considered reasonable by experts in the field (e.g., Federal Rules of Evidence 703).

The fourth requirement is that *scientific evidence must be relevant* under basic relevancy rules, such as Federal Rules of Evidence 401-402. Under FRE 401, evidence is relevant whenever it helps the jury determine the facts at issue to any extent. Gianelli argues that scientific evidence will help the jury when it can be shown to be reasonably *reliable* under a three part test: (1) The underlying principles must be considered valid by the scientific community. Is the test based on conventional scientific theory? (2) The technique applying the principle must be scientifically reliable. Has basic research been conducted that demonstrates the procedure works as predicted, and generally produces statistically significant results? (3) The technique must have been applied properly on the occasion of the particular test (Gianelli, 1980; Neufeld & Colman, 1990).[32] Novelty and lack of general acceptance do not negate reliability, and thus go to the weight rather than the admissibility of the evidence.[33]

This relevancy test creates a standard of presumptive admissibility. A judge "should exclude an expert opinion *only* if it is fundamentally unsupported and would not actually assist the jury in arriving at an intelligent and sound verdict" (Christophersen v. Allied Signal Corporation, 1990, p. 362).[34] As a rule, "it is better to admit relevant scientific evidence in the same manner as other expert testimony and allow its weight to be attacked by cross-examination and refutation" (United Stats v. Baller, 1975, p.463). Thus, any relevant conclusions supported by even a *single* qualified expert witness should be received unless there are distinct reasons for excluding it.[35]

The fifth requirement is that *the probative value of the scientific evidence must be "weighed" against its potential prejudicial effect* (Tanford, 1989). Under the prevailing federal rule, the evidence will be admitted unless its probative value is *substantially* outweighed by the danger it will mislead the jury (see Gianelli, 1980).[36] If the evidence helps prove one of the central disputed issues in a case, it will almost always be admissible.[37] The mere fact the evidence is scientific in nature does not make it prejudicial.

[32] The three-part test is sensible from a scientific point of view.

[33] Degree of acceptance, including testimony by opposing experts, goes to the weight, not the admissibility (Jenkins v. State, 1980; Reed v. State, 1978).

[34] But see Gianelli (1980) which proposes a presumption of *in*admissibility and a requirement that the state prove reliability beyond a reasonable doubt.

[35] Cf. Gianelli (1980) which criticizes the test for this reason. McCormick (1984) thinks one expert is not enough. He would require some corroboration, such as publication in peer-reviewed journal, and that scientists rather than technicians serve as expert witnesses.

[36] Federal Rules of Evidence 403 also provides for the exclusion of relevant evidence that will cause undue prejudice, confusion of the issues, undue waste of time. Scientific evidence does not substantially implicate any of these. Unfair prejudice refers to the arousing of emotions in jurors, when those emotions are not inherent in the nature of the case (Tanford, 1989). Confusion of the issues generally refers to confusion about the legal issues. If evidence tends to cause jurors to apply the facts to an erroneous understanding of the law, it is prejudicial. Undue waste of time usually means that undue time would be spent *on tangential issues*, and does not apply to evidence that casts light on the central issues.

[37] If evidence has any real probative value, courts admit it (see Tanford, 1989).

Courts have generally articulated only one major danger: scientific evidence may mislead a jury into making a *factual* error because of a supposed aura of scientific infallibility (Gianelli, 1980).[38] Social psychologists suggest that the risk of undue influence is minimal -- certainly far short of the "substantially outweighs" standard of Rule 403. The better test would find substantial danger of misleading the jury only when "an exaggerated popular opinion of the accuracy of a particular technique makes its use ... likely to mislead the jury" (United States v. Baller, 1975).[39]

A supposedly different version of the relevancy test has been suggested by United States v. Downing (1985). It would base the admissibility of scientific evidence not in Rules 401-403, but in Rule 702.[40] It assumes that the use of the word "assist" in Rule 702 means more than mere relevancy. The Downing court suggested a balancing analysis that asks: (1) How reliable is the process or technique used in generating the evidence? (2) Will it overwhelm, confuse or mislead the jury? and (3) How strong is the connection between test results and factual dispute? This test seems virtually indistinguishable from the basic relevancy test.

A few appellate courts have endorsed a more informal, ad hoc determination of reliability (see State v. Hall, 1980; United States v. Stifel, 1970). These courts apparently focus primarily on the third criteria: was the test properly conducted in the particular case? The answer is clearly a question of fact for the trial judge. But it seems unwise to extend this area of discretion and grant thousands of trial judges the power to make individual rulings on the validity of scientific theories and techniques. Trial judges will not all be equally scientifically literate, and their rulings are likely to vary. The very essence of science, however, is its universality (Merton, 1973). It does not vary from county to county and case to case, so a rule of law permitting such results is unsound (see Gianelli, 1980).

Is Acoustic-Phonetic Evidence Admissible in the Exxon Valdez Case?

Admissibility Under the Frye Test

We concede at the outset that the results of our analyses of Captain Hazelwood's speech are probably not admissible under the Frye test. Although our evidence derives from generally accepted scientific theory, and uses standard techniques of speech analysis, the particular application to alcohol has not yet achieved general acceptance. Despite all the problems of interpreting the meaning of the old Frye test, it is doubtful that the evidence would be admitted in any jurisdiction that strictly follows the old rule favoring exclusion.

It is also unlikely that our evidence would be admissible under two of the modified Frye tests. The application probably has not yet gained "substantial" acceptance. It is too new for many speech scientists to have heard of and thought about it. Articles describing it are just now appearing in refereed scientific journals. Nor has a subspecialty emerged that uses this novel application to measure the effects of alcohol on speech that would justify using the "narrow field" test.

[38] If techniques are demonstrable in the courtroom and involve principles and procedures understandable to the lay jury, there is little concern with experts having undue influence. However, when the nature of the analysis is esoteric or invisible, for example DNA-typing that depends on knowledge of molecular biology, chemistry, genetics and statistics, a stronger showing of probative value could be required (see Thompson & Ford, 1989).

[39] See Taslitz (1990) concerning the popular belief in infallibility of bloodhounds' powers of scent far exceeding scientific fact.

[40] Federal Rules of Evidence 702 provides: "If scientific ... knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, [an] expert ... may testify thereto in the form of an opinion or otherwise."

This evidence might, however, be admissible under the third Frye test modification that requires a new forensic application to be derived from accepted techniques, based on generally accepted underlying scientific principles, and using generally accepted laboratory equipment. To lay a foundation under this Frye test variation, the proponents must show: (1) The underlying scientific theory is considered valid (generally accepted) by the scientific community, and (2) The techniques and equipment used are known to be reliable and are in widespread use (generally accepted) in the scientific community. These are essentially the same requirements contained in part four of the relevancy test, so will be considered below.

Admissibility Under a Relevancy Test

Whether evidence of acoustic analyses of Captain Hazelwood's speech made from audio tapes from the Exxon Valdez should be admissible under the modern relevancy test depends on the answers to five questions: (1) Is a properly qualified expert witness available who can explain the theoretical and practical reliability of the tests? (2) Is the subject-matter one on which expert testimony will assist the jury in understanding the evidence and drawing accurate conclusions? (3) Were all pieces of equipment used in the test in good working order, and all technical personnel adequately trained? (4) For this particular procedure, are the underlying theories reliable, the techniques valid, and the proper procedure followed? (5) Is there an exaggerated popular opinion of the accuracy of the technique so that its use is likely to mislead the jury?

1. Expert witness. Properly qualified experts in speech science must be available to sponsor the evidence. Speech science is a diverse field that draws upon linguistics, psychology, clinical speech and hearing science, physiology, physics and electrical engineering.[41] The tie that binds the field together is the Acoustical Society of America, which includes a speech communication section. A person with a graduate degree in one of the underlying disciplines, with training or experience focusing on speech science, affiliated with a university- or industry-sponsored speech laboratory, who is a member of the ASA, and who is familiar with the literature on alcohol and speech,[42] would qualify as an appropriate witness.

Because the evidence in the Hazelwood case concerns the effects of alcohol on speech, some experience in the field of alcohol research also is desirable. Even among speech scientists, however, few will have the necessary expertise. The effects of alcohol on speech production is tangential to the main body of acoustic-phonetic research.[43] Few speech scientists have conducted research in the area personally, because the research protocols[44] are more complicated than for usual speech research, and because no obvious theoretical issues are involved. However, at least a few qualified experts are available.[45]

[41] See generally Borden and Harris (1984) which is a basic textbook giving an overview of speech science. See also Flanagan (1972) which is a textbook on engineering and physics of speech; Fry (1979) which is a textbook on speech acoustics; Lieberman and Blumstein (1988) which is textbook on the physiology and psychology of speech.

[42] See e.g., Johnson, Pisoni and Bernacki (1990) which summarizes the relevant literature.

[43] See e.g., the detailed index of Borden and Harris (1984) which contains no entries related to research on the effects of alcohol; the table of contents of Hollien (1990) makes no reference to alcohol effects.

[44] Such as obtaining approval from human subject committees.

[45] E.g., the second and third authors of this paper. Dr. Pisoni has a Ph.D. in Cognitive Psychology, is a Professor of Psychology and Cognitive Science at Indiana University and currently the Director of its Speech Research Laboratory, has personally conducted research on the effects of alcohol on speech, such as Pisoni and Martin (1989), and is familiar with the literature (Johnson, Pisoni, & Bernacki, 1990). Dr. Johnson has a Ph.D. in Linguistics from Ohio State University, is currently a post-doctoral fellow in the Phonetics Laboratory of the

2. Proper subject-matter. For the results of acoustic analyses of Captain Hazelwood's speech to be admissible, the evidence must assist the jury in determining whether Hazelwood was intoxicated. People draw conclusions about intoxication all the time,[46] so this is not a situation in which experts are required because the data necessary to draw reliable conclusions are inaccessible to the average person. Neither is this a topic considered so completely within common knowledge that experts are never permitted. Courts routinely admit other forms of expert testimony concerning alcohol impairment, such as the results of breathalyzer and blood tests,[47] and field sobriety tests (see People v. Randolph, 1989; People v. Krueger, 1968). Therefore, our acoustic-phonetic analyses should be evaluated under the middle category in which particular experts will be allowed if their testimony will assist the jury in drawing accurate conclusions.

Empirical research suggests that jurors often can tell whether a speaker is intoxicated. Many phonetic changes that accompany alcohol impairment, such as word substitutions and revisions, are easily detected by naive listeners. Pisoni and Martin found that both college students and Indiana State Troopers could identify intoxicated (0.10% BAL) talkers with about 80% accuracy when they directly compared intoxicated to sober speech samples (Pisoni & Martin, 1989). Even when a speech sample was presented in isolation, Martin and Yuchtman (1986) found that naive listeners were able to identify intoxicated speakers based on listening with about 66% accuracy. Although these detection rates are better than chance, they also show that jurors make mistakes. Can expert testimony by speech scientists assist the jury in minimizing these mistakes? The issue is not whether experts can *more accurately* determine whether a talker has been drinking, but whether they can provide additional data that will assist the jury's common sense determination.

In the Exxon Valdez case, the information that can be provided by experts is nonredundant. Although speech scientists' observations of gross phonetic changes overlap the aural information jurors rely on, their instrumental measurements provide data otherwise unavailable to jurors: objective, unbiased quantitative measurements of segmental and suprasegmental effects, such as pitch, duration, and amplitude (see Figures 3-6 above). Like breathalyzer measurements, expert evidence based on acoustic-phonetic analyses of speech would provide *unbiased* data to supplement jurors' intuitions. This would assist, rather than hinder, the jury in reaching a more accurate conclusion about whether a talker was intoxicated.

3. Equipment in good working order. To satisfy the third part of the foundation for scientific evidence, all equipment involved in acoustic analyses of speech samples must have been in proper working order. This obvious requirement would be an essential part of speech science research (see Borden & Harris, 1984), even if it were not part of the legal foundation. One measure of the true scientific nature of tests and experiments is the degree of care taken to make sure all the equipment was in good working order and operated properly by trained technicians.

Because the analysis procedures used on the Exxon Valdez tapes involved a comparison across speech samples, rather than absolute detection of impairment, the only real requirement for the recordings is that the same recording and playback equipment was used for both samples in the comparison. Most imperfections in equipment would appear on both the control and test recordings and would not affect

Department of Linguistics at UCLA, has personally conducted research on the effects of alcohol on speech, and is familiar with the literature (Johnson, Pisoni, & Bernacki, 1990).

[46] For example New v. State (1970) suggests that lay people commonly make conclusions about intoxication.

[47] For example Shuman v. State (1986) addresses admissibility of blood serum tests for intoxication, and Ballou v. Henri Studios (1981) addresses admissibility of breath tests.

analysis. For example, if a tape recorder runs slightly slowly, the resulting tapes when played back will show increased average pitch compared to the person's *true* pitch. However, we are concerned only with *relative* changes in pitch between samples, a factor not affected by the overall change in speed of the tape.

Still, a tape recorder could suffer from wow and flutter or other erratic speed fluctuations that could affect tape-to-tape comparisons. The possibility of unusual speed variations in this case was minimized by verifying that the speaking rate of a presumably sober Coast Guard radio operator on the same tapes did not vary significantly. In addition, we analyzed background noise across tape samples and found that its frequency did not vary significantly. These two sets of measurements strongly indicate that the equipment was properly functioning to the extent necessary for drawing reliable conclusions based on tape-to-tape comparisons.

In order to minimize the possibility of error in the laboratory, it is common to take multiple measurements using different personnel. We followed such a procedure in analyzing the tapes of Captain Hazelwood's speech. Two well trained post-graduate researchers made independent overlapping measurements. No errors were detected when we compared them. This is all standard operating procedure in speech science laboratories around the country.[48]

4. Scientific reliability. The evidence concerning Captain Hazelwood's possible alcohol impairment is reliable enough to be relevant under Rules 401 and 402.

a. Theory. The general principles of speech science are considered scientifically valid and are not controversial.[49] The specific theoretical assumption underlying our acoustic analyses of Captain Hazelwood's speech samples is that alcohol affects motor function which affects speech control (see Berry & Pentreath, 1980; Johnson et al., 1990; Klingholz et al., 1988). This, too, seems largely undebatable; after all, it is one of the major assumptions underlying field sobriety tests routinely used by law enforcement agencies as preliminary evidence of alcohol impairment. Another indication of the validity of this theory is the fact that three articles describing this research and its theoretical underpinnings have been published in peer-reviewed scientific journals (Pisoni & Martin, 1990; Johnson et al., 1990; Klingholz et al., 1988). No articles have been published that suggest any contrary view. Indeed, the results are consistent with the findings from over forty years of basic research on speech acoustics and speech production.[50]

b. Validation of technique. The general techniques of acoustic-phonetics that we employed are in common use in speech laboratories throughout the country. Previous research has validated them as reliable methods of measuring speech effects (see e.g., Borden & Harris, 1984; Flanagan, 1972).

Previous controlled studies of subjects other than Captain Hazelwood demonstrate that these techniques can reliably measure when a person's speech has been affected by alcohol consumption. Acoustic analyses cannot *prove* that a person was definitely intoxicated, nor specify exactly the blood alcohol level. However, Pisoni and Martin (1989) revealed that alcohol affects speech in predictable ways, producing patterns of measurable effects more consistent with significant alcohol impairment than with any other known condition that affects speech (Pisoni & Martin, 1990). The results of the Pisoni

[48] See Borden and Harris (1984) which describes standard acoustic-phonetic research protocols.

[49] Speech science generally is described in Borden and Harris (1984), and Flanagan (1972). Acoustic-phonetic analysis of speech sounds is also described in Borden and Harris (1984), and by Flanagan (1972).

[50] See Borden and Harris (1984) which reviews theoretical models of speech production. Flanagan (1972) discusses the physiology of speech production and provides a basic description of how speech is produced.

and Martin validating study are further corroborated by other experimental studies of alcohol (Klingholz et al., 1988; Lester & Skousen, 1974; Pisoni et al., 1986; Sobell & Sobell, 1972; Sobell et al., 1982; Trojan & Kryspin-Exner, 1968).

5. Admissibility Under Rule 403. Because the scientific evidence concerning Captain Hazelwood's speech is reliable, it should be admitted unless some Rule 403 danger *substantially* outweighs its probative value. Assuming the evidence is introduced in a case in which the possible intoxication of Captain Hazelwood is an important issue, its probative value is high. It is the only evidence of its kind. The duration and frequency measurements are unique because they are the only physical evidence and the only *unbiased* evidence bearing on the issue of intoxication. The only other evidence available are the opinions of eyewitnesses who worked for Exxon. Nonredundant evidence on a central issue will almost always be admissible (Tanford, 1989), because it would require extreme Rule 403 dangers to significantly outweigh high probative value.

There is no problem here with the usual danger associated with scientific evidence -- misleading the jury into making a *factual* error because of an exaggerated popular opinion of the accuracy of a particular technique. There is little, if any, evidence that the public has ever heard of acoustic-phonetic analysis of speech.

Are there any other possible Rule 403 dangers? The only other likely objection to it is that it may constitute an undue waste of time. This objection is normally unavailing when the evidence goes to the heart of an issue, as this evidence does. It is unlikely to be considered an *undue* waste of time to allow a battle of experts on the central issue--whether Hazelwood was or was not drunk. In any event, battles occur mostly when scientific evidence requires considerable subjective interpretation based on the absence of hard data -- psychiatric diagnosis being the paradigmatic example. Our conclusion that Hazelwood's voice showed effects consistent with alcohol consumption is based in part on objective data (i.e., physical measurements of duration and fundamental frequency) requiring little interpretation. No evidence exists to suggest that there is likely to be any real debate over the conclusions that we draw, only over the methodology. That battle will be fought mostly in front of the judge on the question of admissibility. Therefore, there is no serious risk of prejudice that can *substantially* outweigh the relevancy of the evidence.

Conclusions

Expert testimony that, based on acoustic analyses of audio tapes, Captain Hazelwood probably was intoxicated at the time the Exxon Valdez ran aground, should be admitted. Properly qualified experts are available to sponsor the evidence. The evidence will assist the jury in determining one of the key facts in issue. The analyses of the Hazelwood tapes appear to have been conducted properly. The evidence itself is scientifically reliable. It is based on accepted theories of speech acoustics and uses standard equipment and technology. The accuracy of these techniques has already been demonstrated in controlled laboratory experiments in which subjects were intoxicated to known blood alcohol levels. No particular fact-finding danger is posed by its use, so these analyses should be admissible under the emerging "relevancy test" for scientific evidence.

References

Ballou v. Henri Studios, 656 F2d 1147 (5th Cir 1981).

Bazelon, D.L. (1977). Coping with technology through the legal process. *Cornell Law Review*, **62**, 817-832.

Berry, M.S., & Pentreath, V.W. (1980). The neurophysiology of alcohol. In Sandler, M. (Ed.), *Psychopharmocology of Alcohol*. New York: Raven Press, 43-72.

Borden, G.J., & Harris, K.S. (1984). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Baltimore, MD: Williams and Wilkins.

Brenner, M., & Shipp, T. (1988). Voice stress analysis. *Mental State Estimation, NASA Conference Publication 2504*, 363-376.

Christophersen v. Allied Signal Corporation, 902 F.2d 362 (5th Cir. 1990).

Cleary, E.N. (1984). *McCormick on Evidence*. St.Paul, MN: West Publishing.

Commonwealth v. Fatalo, 346 Mass 266, 191 N.E.2d 479 (1963).

Coppolino v. State, 223 So.2d 68 (Fla. App. 1968; app. dism. 234 So.2d 120 (Fla. 1969).

Expert testimony based on novel scientific techniques: Admissibility under the federal rules of evidence. (1980). *George Washington Law Review*, **48**, 774-790.

Flanagan, J.L. (1972). *Speech Analysis, Synthesis, and Perception*. New York: Academic Press.

Fry, D.B. (1979). *The Physics of Speech*. Cambridge: Cambridge University.

The Frye doctrine and relevancy approach controversy: An empirical evaluation. (1986). *Georgia Law Journal*, **74**, 1769-1791.

Frye v. United States, 293 Fed. 1013 (D.C.Cir. 1923).

Gianelli, P.C. (1980). The admissibility of novel scientific evidence: Frye v. United States, a half-century later. *Columbia Law Review*, **80**, 1197-1250.

Griffin, G.R., & Williams, C.E. (1987). The effects of different levels of task complexity on three vocal measures. *Aviation Space Environmental Medicine*, **58**, 1165-1170.

Hansen, J. (1988). *Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition. Doctoral dissertation, Georgia Institute of Technology.*

Hollien, H. (1990). *The Acoustics of Crime: The New Science of Forensic Phonetics*. New York: Plenum Press.

Ibn-Tamas v. United States, 407 A.2d 626 (D.C. App. 1979).

Jenkins v. State, 156 Ga. App. 387 (1980).

Johnson, K.A., Pisoni, D.B., & Bernacki, R.H. (1990). Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica*, **47**, 215-237.

Klingholtz, F., Penning, R., & Liebhardt, E. (1988). Recognition of Low-Level Alcohol Intoxication from Speech Signals. *Journal of the Acoustical Society of America*, **84**, 924-935.

Kuhn, T. (1970). *The Structure of Scientific Revolution*. Chicago, IL: University of Chicago Press.

Latin, H.A., Tannehill, G.W., & White, R.E. (1976). Remote sensing evidence and environmental law. *California Law Review*, **64**, 1300-1446.

Lester, L., & Skousen, R. (1974). The phonology of drunkenness. In Bruck, A., Fox, R.A., & LaGaly, M.W. (Eds.), *Papers From the Parasession on Natural Phonology*. Chicago: Chicago Linguistics Society.

Lieberman, P., & Blumstein, S.E. (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge: Cambridge University Press.

Maltskos, C.J., & Spielman, S.J. (1967). Introduction of new scientific methods in court. In Yefsky, S.A. (Ed.), *Law Enforcement Science and Technology*. Washington, D.C.: Thompson.

Martin, C.S., & Yuchtman, M. (1986). Using speech as an index of alcohol-intoxication. *Research on Speech Perception Progress Report No. 12*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Moenssens, A.A., Inbau, F.E., & Starr, J.E. (1986). *Scientific Evidence in Criminal Cases*. Mineola, NY: Foundation Press.

Moore, T.J., & Bond, Z.S. (1987). Acoustic-phonetic changes in speech due to environmental stressors: Implications for speech recognition in the cockpit. *Annual Symposium on Aviation Psychology*, **4**, 26-30.

Myers, M.A. (1979). Rule departures and making law: Juries and their verdicts. *Law and Society Review*, **13**, 781-797.

Neufeld, P.J., & Colman, N. (1990). When science takes the witness stand. *Scientific American*, **262**, 46-53.

New v. State, 254 Ind. 307, 259 NE2d 696 (1970).

New York Times v. National Aeronautics and Space Administration, 920 F.2d 1002 (D.C. Cir. 1990).

Nisbett, R., & Ross, L. (1980). Human Inference: Strategies and Shortcomings of Social Judgment. Englewood Cliffs, NJ: Prentice-Hall.

People v. Kelly, 17 Cal.3d 24, 549 P.2d 1240 (1976).

People v. King, 266 Cal. App. 2d 437; 72 Cal.Rptr. 478 (1968).

People v. Krueger, 99 Ill. App. 2d 431 (1968).

People v. Law, 40 Cal. App.3d 69, 114 Cal. Rptr. 708 (1974).

People v. Randolph, 213 Cal. App. 3d Supp. 1, 262 Cal. Rptr. 378 (1989).

People v. Tobey, 401 Mich. 141, 257 N.W.2d 537 (1977).

People v. Williams, 164 Cal. App. 2d Supp. 858, 331 P.2d 251 (Super. Ct. 1958).

Pennsylvania v. Muniz, 110 S.Ct. 2638 (1990).

Petrosinelli, J.G. (1990). The admissibility of DNA typing: A new methodology. *Georgetown Law Journal*, **79**, 313-336.

Pisoni, D.B., Yuchtman, M., & Hathaway, S.N. (1986). Effects of alcohol on the acoustic-phonetic properties of speech. In *Alcohol, Accidents and Injuries*. Warrendale, PA: Society of Automotive Engineers.

Pisoni, D.B., & Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analysis. *Alcoholism*, **13**, 577-587.

Reed v. State, 283 Md 384; 391 A2d 364 (1978).

Richardson, J. (1974). *Modern Scientific Evidence, Civil and Criminal*. Cincinnati, OH: W.H. Anderson.

Risinger, D.M., Denbeaux, M.P., & Saks, M.J. (1989). Exorcism of ignorance as a proxy for rational knowledge: The lessons of handwriting identification "Expertise". *University of Pennsylvania Law Review* **137**, 731-792.

Saks, M., & Kidd, R. (1981). Human information processing and adjudication: Trial by heuristics. *Law and Society Review*, **15**, 123-160.

Shuman v. State, 489 NE2d 126 (Ind. Ct. App. 1986).

Sobell, L.C., & Sobell, M.B. (1972). Effects of alcohol on the speech of alcoholics. *Journal of Speech and Hearing Research*, **15**, 861-868.

Sobell, L.C., Sobell, M.B., & Coleman, R.F. (1982). Alcohol-induced dysfluency in nonalcoholics. *Folia Phoniatrica*, **34**, 316-323.

South Dakota v. Neville, 459 U.S. 553 (1983).

State v. Carlson, 267 N.E.2d 170 (Minn. 1978).

State v. Hall, 297 NW2d 80 (Iowa 1980).

State v. Maudlin, 416 NE2d 477 (Ind. Ct. App. 1981).

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., & Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, **84**, 917-928.

Tanford, J.A. (1989). A political-choice approach to limiting prejudicial evidence. *Indiana Law Journal*, **64**, 831-872.

Tanford, J.A. (1990). The limits of a scientific jurisprudence: The supreme court and psychology. *Indiana Law Journal*, **66**, 137-173.

Tanford, J.A., & Tanford, S. (1988). Better trials through science: A defense of psychologist-lawyer collaboration. *North Carolina Law Review*, **66**, 741-780.

Taslitz, A.E. (1990). Does the cold nose know? The unscientific myth of the dog scent line-up. *Hastings Law Journal*, **42**, 15-134.

Thompson, W.C. (1990). Are juries competent to evaluate statistical evidence. *Law and Contemporary Problems*, **52, 9-41.**

Thompson, W.C., & Ford, S. (1989). DNA typing: Acceptance and weight of the new genetic identification tests. *Virginia Law Review*, **75**, 45-108.

Thompson, W.C., & Schumann, E.L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, **11**, 167-187.

A Trial of Witches. (1665). *Howell's State Trials*, **6**.

Tribe, L.H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, **84**, 1329-1393.

Trojan, F., & Kryspin-Exner, K. (1968). The decay of articulation under the influence of alcohol and paradehyde. *Folia Phoniatrica*, **20**, 217-238.

United States v. Addison, 498 F.2d 741 (D.C. Cir. 1974).

United States v. Baller, 519 F2d 463 (4th Cir 1975).

United States v. Downing, 735 F.2d 1224 (3d Cir 1985).

United States v. Shorter, 809 F.2d 54 (D.C.Cir. 1-16-87).

United States v. Stifel, 433 F.2d 431 (6th Cir. 1970).

Wigmore, J. (1940). *Evidence in Trials at Common Law* § *10*, 293.

Williams, C.E., & Stevens, K.N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, **52**, 1238-1250.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Effects of Alcohol on Speech: Acoustic Analyses of Spondees[1]

Dawn M. Behne and Susan M. Rivera

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

[1] The authors wish to thank David B. Pisoni for his helpful comments. This research is supported by NIH Training Grant #DC-00012-12 to Indiana University.

Abstract

The acoustic effects of alcohol on speech production are not well understood, and have not been widely investigated. In the present study we analyzed recordings of spondees produced while subjects were sober and intoxicated. Acoustic measures were used to investigate the effects of alcohol on phonetic segments and prosody. The results of these analyses revealed that when a talker is intoxicated the speech samples are produced with lower second and third formants, a higher mean amplitude and mean fundamental frequency and greater variability. The findings obtained with spondees are consistent with earlier reports from our laboratory using isolated sentences. The implications of these findings and previous research on alcohol are discussed.

Effects of Alcohol on Speech: Acoustic Analyses of Spondees

Alcohol is known to be a central nervous system depressant which impairs physiological processes, such as nerve impulse transmission and motor control (American Medical Association, 1968). The fine motor coordination and timing of the vocal apparatus makes speech production susceptible to the effects of alcohol. Among other impairments in speech behavior, the changes in speech production resulting from alcohol consumption are characterized by both segmental and prosodic effects manifested in the speech signal.[2]

Segmental Effects

Segmental effects involve the articulation of specific speech sounds. The speech dysfluency resulting from alcohol consumption is often described as "slurred", suggesting that in the production of speech segments, articulatory coordination has been disrupted. Lester & Skousen (1974) examined changes in the segmental properties of speech produced by subjects who read lists of words and engaged in conversations at stages varying between sobriety and intoxication. They found that when talkers were intoxicated, the duration of consonants in unstressed syllables increased, and certain consonants were produced with a more retracted place of articulation. Pisoni and his colleagues (1985, 1986, 1989) conducted segmental analyses of sentences produced in a sober condition and a condition of 0.10% blood alcohol level (BAL) or greater. They reported that intoxicated speakers failed to achieve complete oral closure in the production of stops and affricates, but found no evidence of spectral differences for consonants or vowels.

Prosodic Effects

Prosodic effects refer to the features of speech which extend across speech segments, such as loudness, pitch, and speaking rate. Previous research has suggested that alcohol consumption may affect acoustic correlates of prosody. Sobell, Sobell & Coleman (1982) investigated the effects of alcohol on speech production in a sober, moderately intoxicated (0.05% BAL) and highly intoxicated condition (0.10% BAL). They found that the peak amplitude was lower in their two intoxicated conditions compared with the sober condition but they found no effect on the [mean] fundamental frequency (F0) across the three conditions. Trojan & Kryspin-Exner (1968) recorded subjects naming pictures and speaking spontaneously while in a sober condition and under two levels of intoxication. They found that the effects of alcohol on mean F0 varied among speakers and that alcohol consumption did not appear to have a systematic effect on mean F0. Similarly, Pisoni et al. (1989, 1986) found that F0 decreased with intoxication for some subjects, but not for all. However, they reported that F0 variability was greater in the intoxicated condition than the sober condition.

Previous investigations have also suggested that alcohol consumption may affect speaking rate. Sobell et al. (1982) found that reading rate was slower in the 0.10% BAL condition compared to the sober and 0.05% BAL condition although no difference between the sober and 0.05% BAL condition was obtained. Pisoni, Yuchtman and Hathaway (1986) found overall sentence duration to be greater in the intoxicated condition than in the sober condition.

In summary, previous research on the effects of alcohol on speech production has shown that alcohol can affect acoustic properties of phonetic segments and prosody. Consonant durations may increase, timing of obstruents may be affected, and place of articulation for consonants may, or may not,

[2] In addition to segmental and prosodic effects, gross speech errors, such as lexical omissions and interjections, have been noted (e.g. Sobell et al., 1972; Andrews et al., 1977; Sobell et al., 1982).

be retracted. However, there has been no evidence that vowel production is affected by alcohol. Alcohol has also been found to decrease amplitude, increase F0 variability, and decrease speaking rate. However no systematic effect of alcohol on mean F0 has been found.

The present investigation is part of an on-going project, portions of which have been reported in Pisoni et al. (1985, 1986, 1989). These papers examined sentences produced in both a sober and an intoxicated condition. The present paper reports acoustic analyses of isolated spondees produced in these two conditions. Our analyses address both segmental and prosodic effects of alcohol on speech production.

Method

Subjects

Six male students from Indiana University participated in the study.[3] All of the subjects were at least 21 years old, were native speakers of English, and had no history of a speech, hearing or language disorder. Each subject completed a set of paper and pencil questionnaires used to assess alcohol consumption and risk for alcoholism. These consisted of the short Michigan Alcoholism Screening Test (Selzer, 1975), the MacAndrew Scale (MacAndrew, 1965), the socialization subscale of the California Psychological Inventory (Gough, 1969) and a short alcohol-consumption questionnaire. Only subjects whose scores on these tests showed them to be both moderate social drinkers and at low risk for alcoholism were included in the experiment. Table 1 summarizes the scores obtained by the subjects on these questionnaires.

Insert Table 1 about here

Materials

Thirty-eight spondees were used to elicit speech samples from subjects. A spondee is a word which has two stressed syllables, such as "eggplant" or "horseshoe". The complete list of spondees used in the experiment is presented in Table 2.

Insert Table 2 about here

Procedure

Subjects agreed not to eat or drink for at least four hours prior to the experiment.

Each subject was seen and recorded individually in two counterbalanced conditions; in one condition the subject was sober, and in the other condition the subject consumed enough alcohol to raise his BAL to 0.10% weight/volume. In both conditions, the subject's verbal production of the thirty-eight spondees was recorded.

Subject Preparation. In the sober condition, a Smith and Wesson Breathalyzer (Model 900A) was used to measure the alcohol level of each subject before his speech was recorded. This was done to verify that no alcohol was in his system at the time of testing.

[3] Although data was gathered for nine subjects, difficulties arose in pitch tracking for three subjects in a manner that appears unrelated to the effects of alcohol. Consequently, only the data of six subjects are reported here.

Table 1

Subjects' ages, and BALs at the beginning and end of the recording sessions. Also shown are scores on the MAST, the socialization scale, and scores on the MacAndrew scale. Self-reported total alcohol intake during the 30 days prior to recording (converted to ounces of 200 proof alcohol) are shown in the final column.

Subjects	Age	Initial BAL	Final BAL	MAST	SOC	MAC	Alcohol Intake
1	21	.17%	.10%	5	30	27	16.80
2	21	.13%	.075%	4	29	20	5.15
3	22	.15%	.085%	5	31	27	23.20
4	25	.13%	.15%	7	33	18	26.99
5	26	.10%	.10%	2	35	22	6.15
6	22	.16%	.10%	3	39	23	3.53

Table 2

List of thirty-eight spondees used to elicit speech samples for both the sober and intoxicated condition.

1.	eggplant	20.	duckpond
2.	airplane	21.	toothbrush
3.	washboard	22.	daybreak
4.	scarecrow	23.	backbone
5.	lifeboat	24.	hardware
6.	woodchuck	25.	nutmeg
7.	schoolboy	26.	hotdog
8.	buckwheat	27.	seahorse
9.	oatmeal	28.	wildcat
10.	shipwreck	29.	football
11.	whitewash	30.	platform
12.	cookbook	31.	drawbridge
13.	birthday	32.	starlight
14.	hedgehog	33.	woodwork
15.	bathtub	34.	inkwell
16.	sundown	35.	footstool
17.	windmill	36.	mushroom
18.	earthquake	37.	northwest
19.	headlight	38.	horseshoe

In the intoxicated condition, each subject was weighed and given a breath analysis test before the intoxication process began. Using one gram of alcohol per kilogram of the subject's weight, a mixture of 1 part 80 proof vodka and 3 parts orange juice was prepared. This dose was designed to raise the subject's BAL to 0.10% weight/volume over a 45 minute period. The subject was given a third of the total dose every 15 minutes and was asked to consume the drink gradually over the entire 15 minute period. At the end of 45 minutes, the subject rinsed his mouth several times to remove traces of alcohol from his mouth and was given another breath analysis test. If the subject's BAL was still below 0.10%, he was given another drink containing the same amounts of vodka and orange juice as each of the previous three drinks, and was asked to consume the drink gradually over 15 minutes. The subject then repeated the mouth-rinsing and breath analysis test. When the subject's BAL reached at least 0.10%, recordings of his speech were made. Table 1 gives each subject's BAL prior to the recording session.

Recordings. Subjects sat in a sound-attenuated IAC booth and wore a matched pair of calibrated TDH-39 headphones with an attached EV C090 LO-Z condenser microphone which was adjusted to be four inches away from the front of the subject's mouth.

The spondees were presented visually one at a time on a CRT monitor using a PDP 11/34. The presentation rate of the spondees was self-paced and controlled by the subject; the subject pressed a button on a response box that was interfaced to the computer. Subjects were asked to say the word on the monitor as quickly as possible and press the button for the next item to appear.

Audio recordings of the subject's productions were made using an Ampex-500 tape recorder. The recording level was adjusted at the beginning of the first condition the subject participated in and maintained throughout both the sober and the intoxicated condition.

Recording sessions were the same for the sober and intoxicated conditions. However, in the intoxicated condition, subjects were given a final breath analysis test after the recording session. Table 1 shows each subjects's BAL at the end of the testing.

Measurements

The audio recordings were low-pass filtered and digitized at a rate of 20,000 samples/second. Measurements were made using SRD (Speech Read) which allows an experimenter to mark acoustic events of interest and extract measures of the marked segment to a parameter file. Among the information in the parameter file is the duration of the marked segment; the mean F1, F2 and F3 frequencies within the marked segment; mean fundamental frequency and its variability within the marked segment; and, the mean amplitude and its variability within the marked segment. Using SRD the following measures were extracted for each spondee produced in the sober and intoxicated condition by each subject.

Vowel duration: The duration from the beginning to the end of the periodic portion of the vowel in each syllable.

Intervowel duration: The duration between the end of the first vowel and the beginning of the second vowel of the spondee.

Vowel-to-word duration ratio: The vowel duration divided by the duration of the entire spondee for each syllable.

Intervowel-to-word duration ratio: The intervowel duration divided by the duration of the entire spondee.

First, second and third formant frequencies: The mean frequency of the first (F1), second (F2), and third (F3) formants within the vowel of each syllable.

Mean amplitude: The mean amplitude within the vowel of each syllable.

Amplitude variability: The standard deviation of the amplitude within the vowel of each syllable.

Mean fundamental frequency: The mean fundamental frequency within the vowel of each syllable.

Fundamental frequency variability: The standard deviation of the fundamental frequency within the vowel of each syllable.

Results

The results are divided into sections corresponding to segmental and prosodic effects respectively. For each of the measures, an analysis of variance compared the means of the sober and intoxicated conditions.

Segmental Effects

Formant Frequencies. Formant frequencies are known to reflect relative positions of articulatory constriction in vowel production. If alcohol affects the fine motor coordination of consonant production as previous research has suggested, vowel quality may also be affected. An effect on vowel quality should be realized by changes in the frequencies of the first three formants. The mean frequency of F1, F2, and F3 were determined for V1 and V2. The condition means and F-values for F1, F2, and F3 are shown in Table 3. The mean F1, F2 and F3 for each subject are presented in Figures 1-3 respectively.

Insert Table 3 about here

Insert Figures 1-3 about here

The analyses of variance showed no effect of alcohol on F1, but revealed differences for F2 and F3. Although the mean F2 was lower in the intoxicated than the sober condition for both V1 and V2, this difference was only significant for V1. The difference between F2 of V1 and V2 can be seen in the subject means in Figure 2. Throughout the spondee for Subjects 1-3, F2 is lower in the intoxicated condition than the sober condition. For Subjects 4 and 6, F2 of V1 is lower in the intoxicated condition than the sober condition but F2 of V2 shows the opposite pattern. For Subject 5, both V1 and V2 have a higher F2 in the intoxicated condition than in the sober condition.

Alcohol also had a significant effect on F3. F3 was lower in the intoxicated condition than in the sober condition for both vowels. Figure 3 illustrates this general pattern across subjects. F3 is lower in the intoxicated condition than the sober condition for Subjects 1-4, but not for Subject 5 or 6. For Subject 5, F3 of V1 is higher in the intoxicated condition than in the sober condition whereas F3 of V2

Table 3

Condition means, F-values and probabilities for formant frequencies.

Measure	**Sober (Mean)**	**Intoxicated (Mean)**	***F* Value**	***p***
V1: F1	566 Hz.	576 Hz.	0.31	n.s.
V2: F1	605 Hz.	595 Hz.	0.32	n.s.
V1: F2	1992 Hz.	1913 Hz.	4.27	0.039
V2: F2	1956 Hz.	1885 Hz.	2.50	n.s.
V1: F3	3377 Hz.	3247 Hz.	6.52	0.011
V2: F3	3368 Hz.	3223 Hz.	5.69	0.017

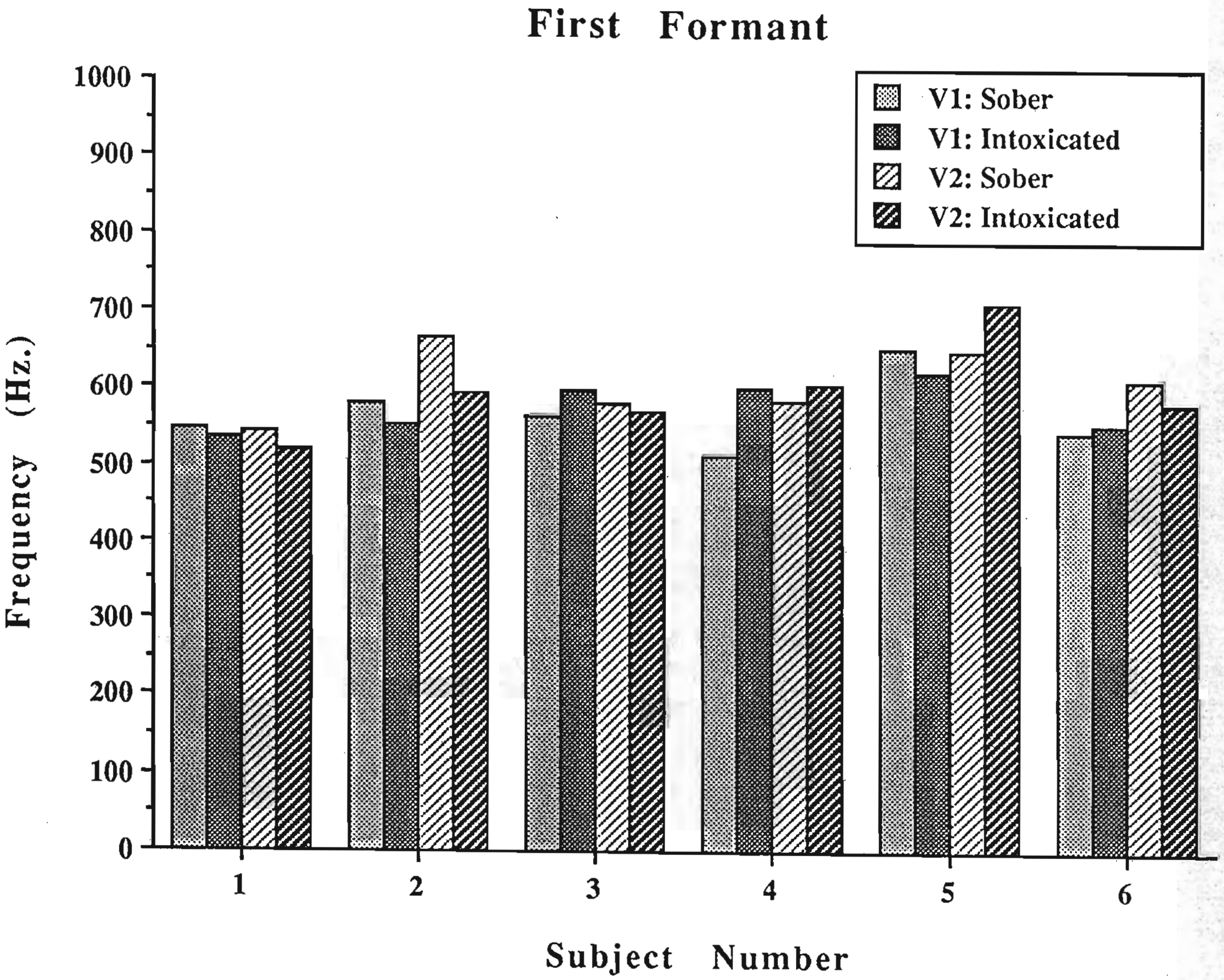


Figure 1. Mean F1 of V1 and V2 produced by six subjects in the sober and intoxicated condition.

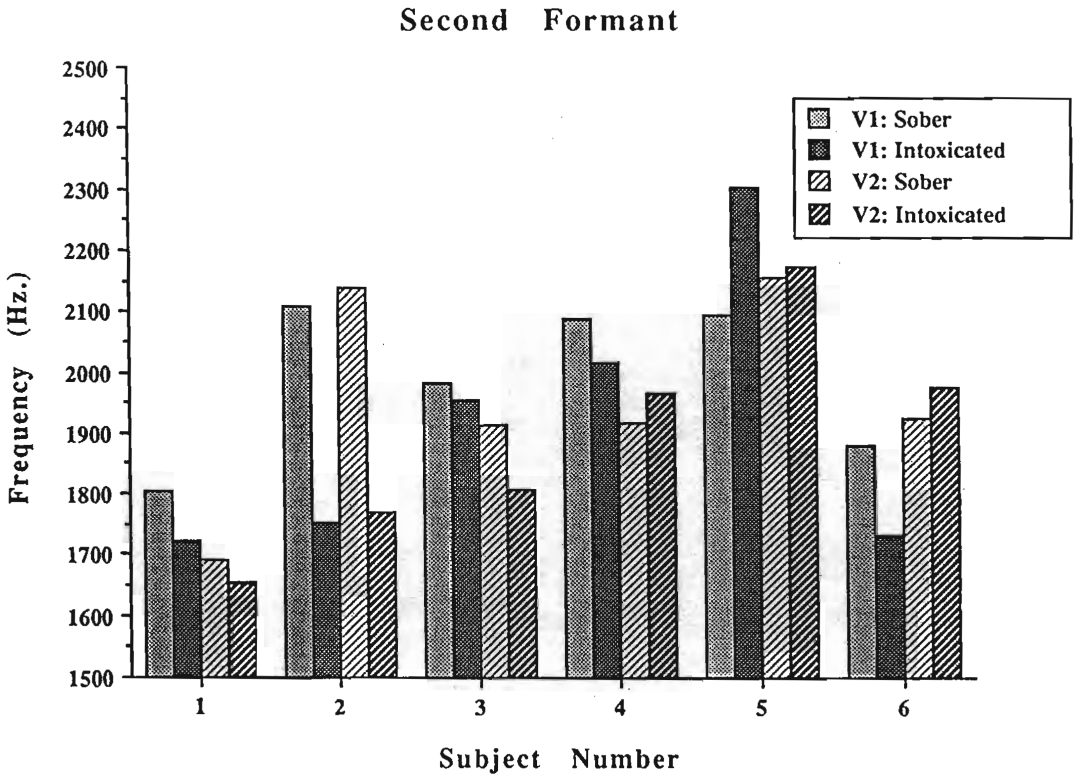


Figure 2. Mean F2 of V1 and V2 produced by six subjects in the sober and intoxicated condition.

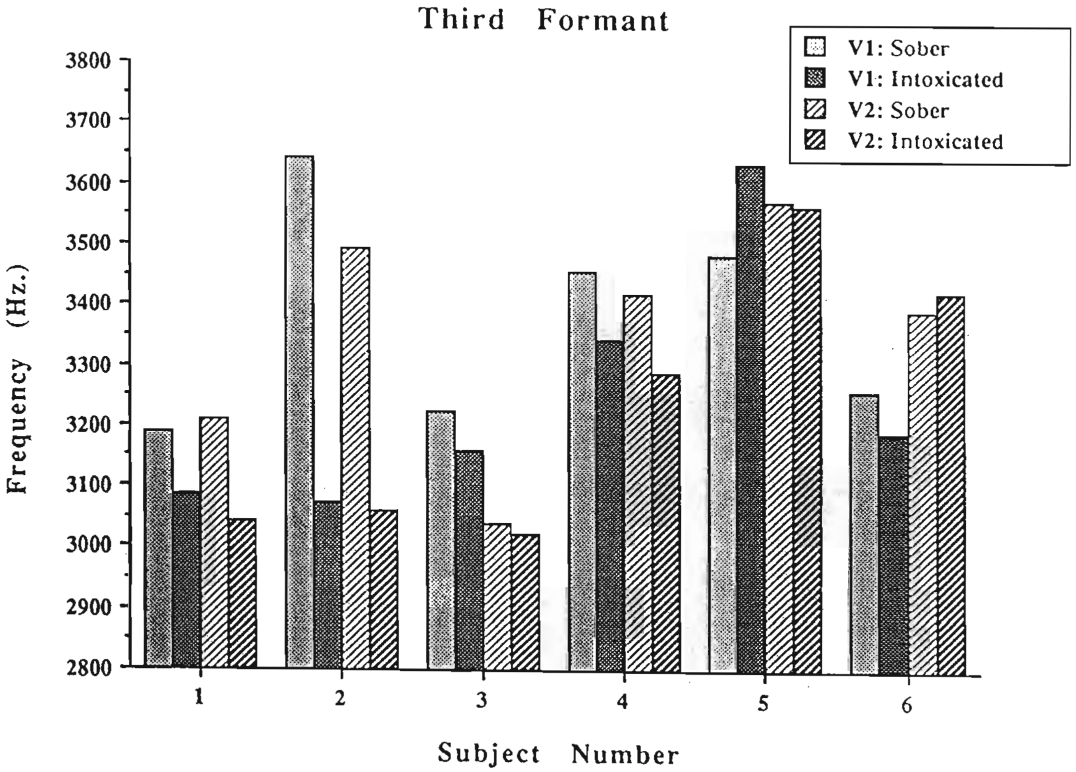


Figure 3. Mean F3 of V1 and V2 produced by six subjects in the sober and intoxicated condition.

is approximately the same for the two conditions. For Subject 6, F3 of V1 is lower in the intoxicated condition than the sober condition, like Subjects 1-4; but F3 of V2 shows the opposite pattern.

Overall, F2 and F3 both tend to be lower in the intoxicated condition relative to the sober condition. However, the results vary across subjects. Subjects 1-3 clearly show this pattern of F2 and F3 for both vowels, and Subject 4's results are similar. However, Subjects 5 and 6 show a distinctly different pattern of results of formant frequencies.

Segment-to-word Duration Ratios. Ratios of segment-to-word duration were calculated in order to examine effects of alcohol on segmental duration, without the possible influences of speaking rate which will be discussed below.[4] Ratios of the vowel-to-word duration for V1 and V2, and intervowel-to-word duration were computed. The condition means and F-values are shown in Table 4. The mean vowel-to-word and intervowel-to-word duration ratios for each subject are displayed in Figures 4 and 5, respectively.

Insert Table 4 about here

Insert Figures 4-5 about here

Although the means shown in Table 4 and Figure 4 suggest an increase in vowel-to-word duration ratio for V1 and V2 in the intoxicated condition, these differences were not significant. The intervowel-to-word duration ratio was affected by alcohol. The ratio of intervowel-to-word duration was significantly reduced in the intoxicated condition, compared with the sober condition. The subject means in Figure 5 illustrate this pattern for Subjects 1, 2, 4 and 5, but not for Subjects 3 or 6.

Overall, the different results for vowel-to-word ratios and intervowel-to-word ratios suggest that effects of alcohol on consonant duration, such as that measured by intervowel duration, is different than the effect of alcohol on vowel segments.

Prosodic Effects

Amplitude. Mean amplitude and amplitude variability were measured for both vowels of each spondee. The condition means and F-values for mean amplitude and amplitude variability are presented in Table 5, and shown for each subject in Figures 6 and 7.

Insert Table 5 about here

Insert Figures 6-7 about here

Alcohol significantly affected mean amplitude; for both syllables of the spondees, mean amplitude was higher for speech produced in the intoxicated condition compared to the sober condition. As Figure 6 shows, the effect on mean amplitude was found for both vowels of the spondees produced by Subjects 1-4. Subject 5 shows the same difference between the sober and intoxicated conditions for the first vowel

[4] Using these duration ratios to extract speaking rate assumes that speaking rate will proportionately increase or decrease segmental durations.

Table 4

Condition means, F-values, and probabilities for segment-to-word ratios.

Measure	**Sober (Mean)**	**Intoxicated (Mean)**	***F* Value**	***p***
V1/Word Duration	0.18 msec.	0.19 msec.	0.85	n.s.
V2/Word Duration	0.27 msec.	0.28 msec.	3.08	n.s.
Intervowel/Word Dur.	0.28 msec.	0.26 msec.	7.76	0.006

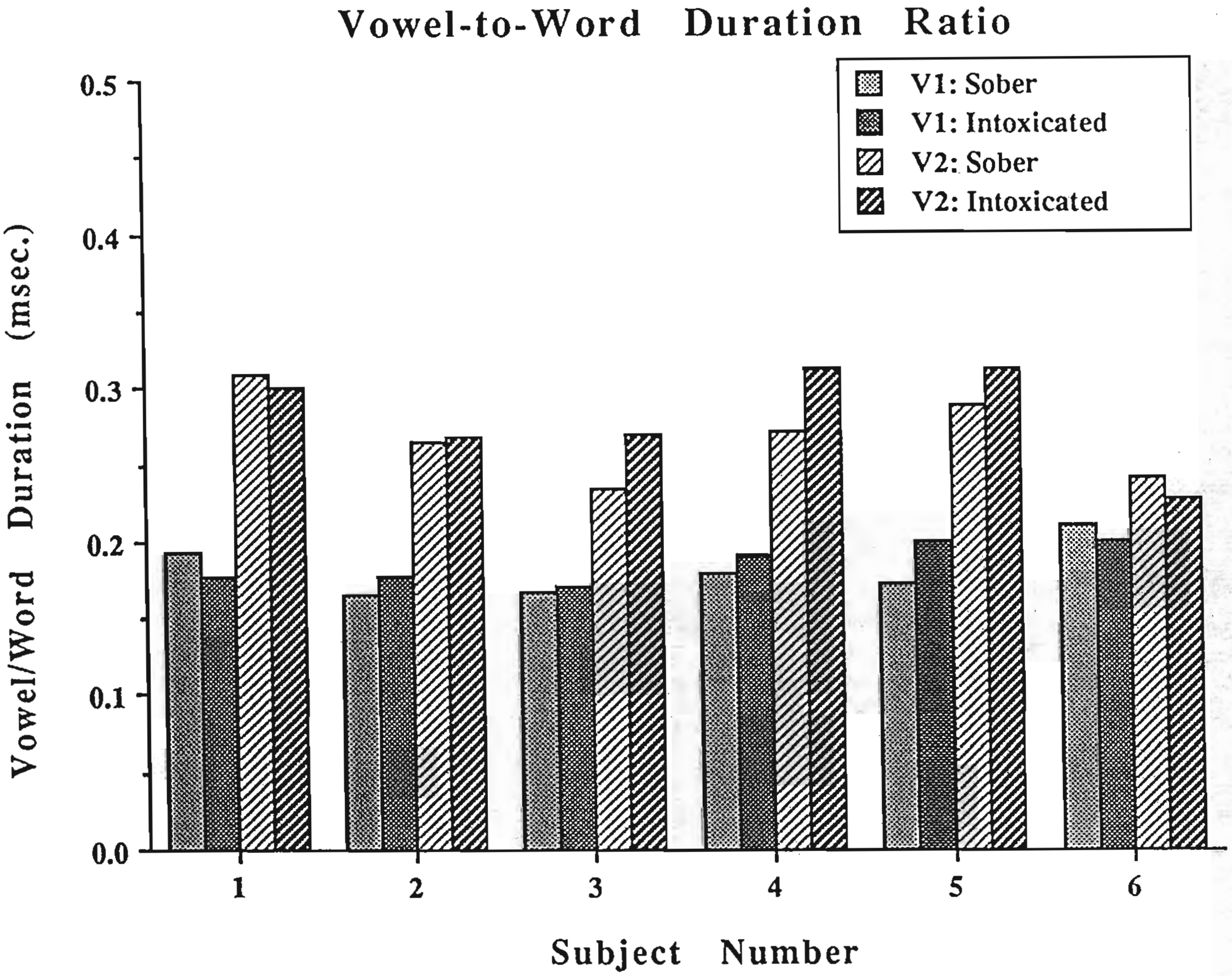


Figure 4. Mean vowel-to-word duration ratio of V1 and V2 produced by six subjects in the sober and intoxicated condition.

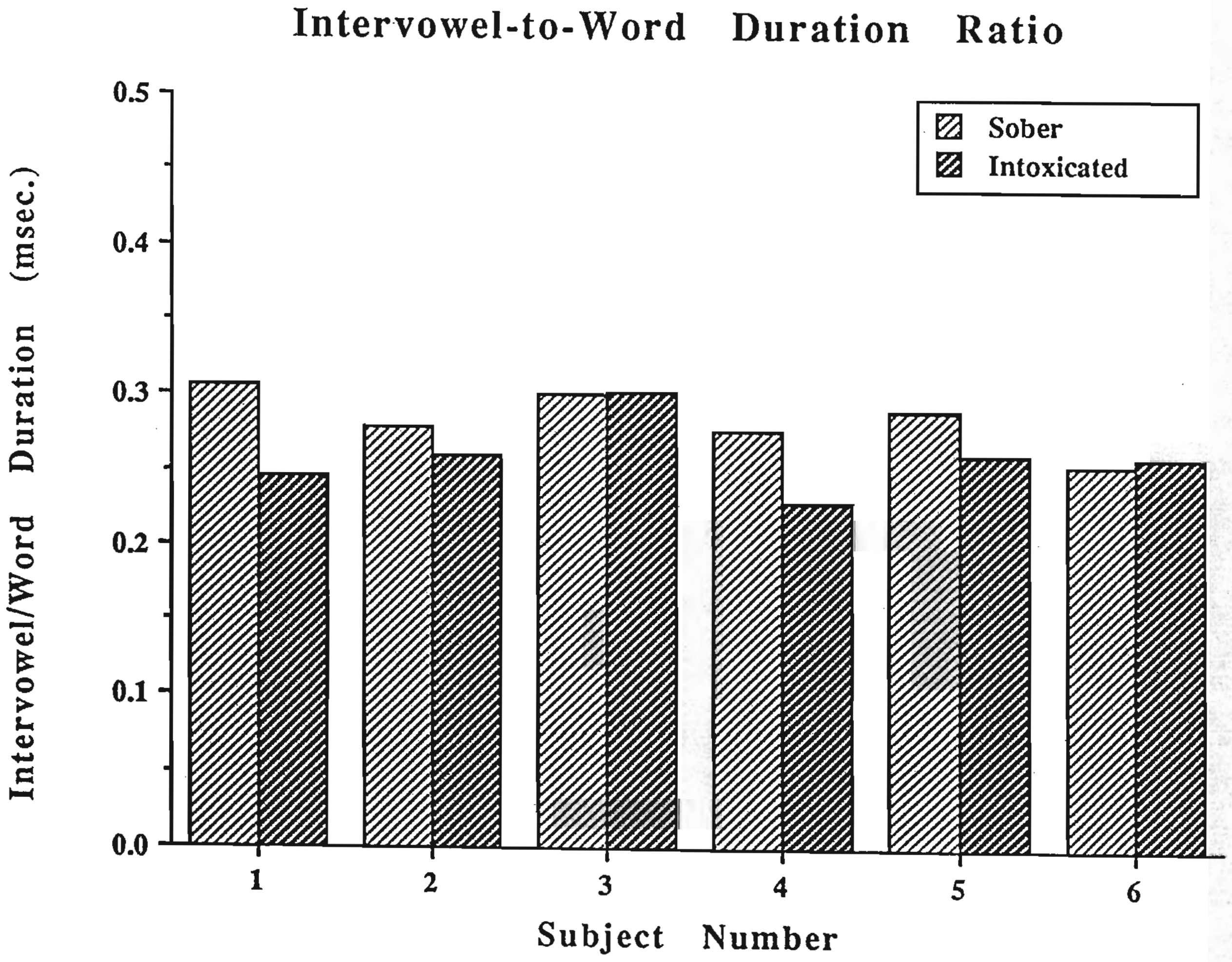


Figure 5. Mean intervowel-to-word duration ratio of V1 and V2 produced by six subjects in the sober and intoxicated condition.

Table 5

Condition means, F-values, and probabilities for amplitude measures.

Measure	Sober (Mean)	Intoxicated (Mean)	*F* Value	*p*
V1: Mean Amplitude	79.5 dB.	80.9 dB.	67.49	0.0001
V2: Mean Amplitude	75.9 dB.	78.2 dB.	173.16	0.0001
V1: Amplitude Var.	74.6 dB.	76.4 dB.	69.38	0.0001
V2: Amplitude Var.	71.6 dB.	74.0 dB.	122.67	0.0001

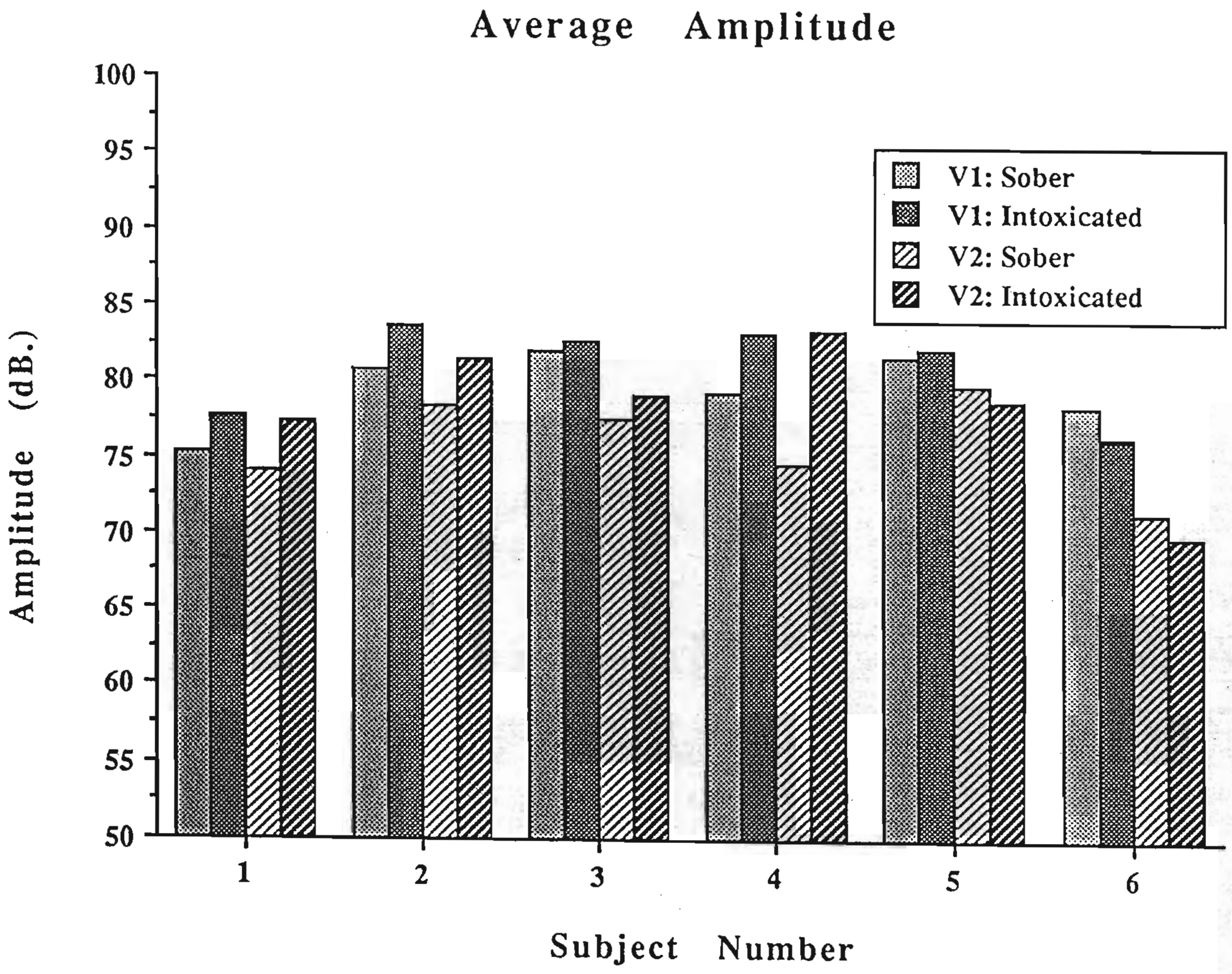


Figure 6. Mean amplitude of V1 and V2 produced by six subjects in the sober and intoxicated condition.

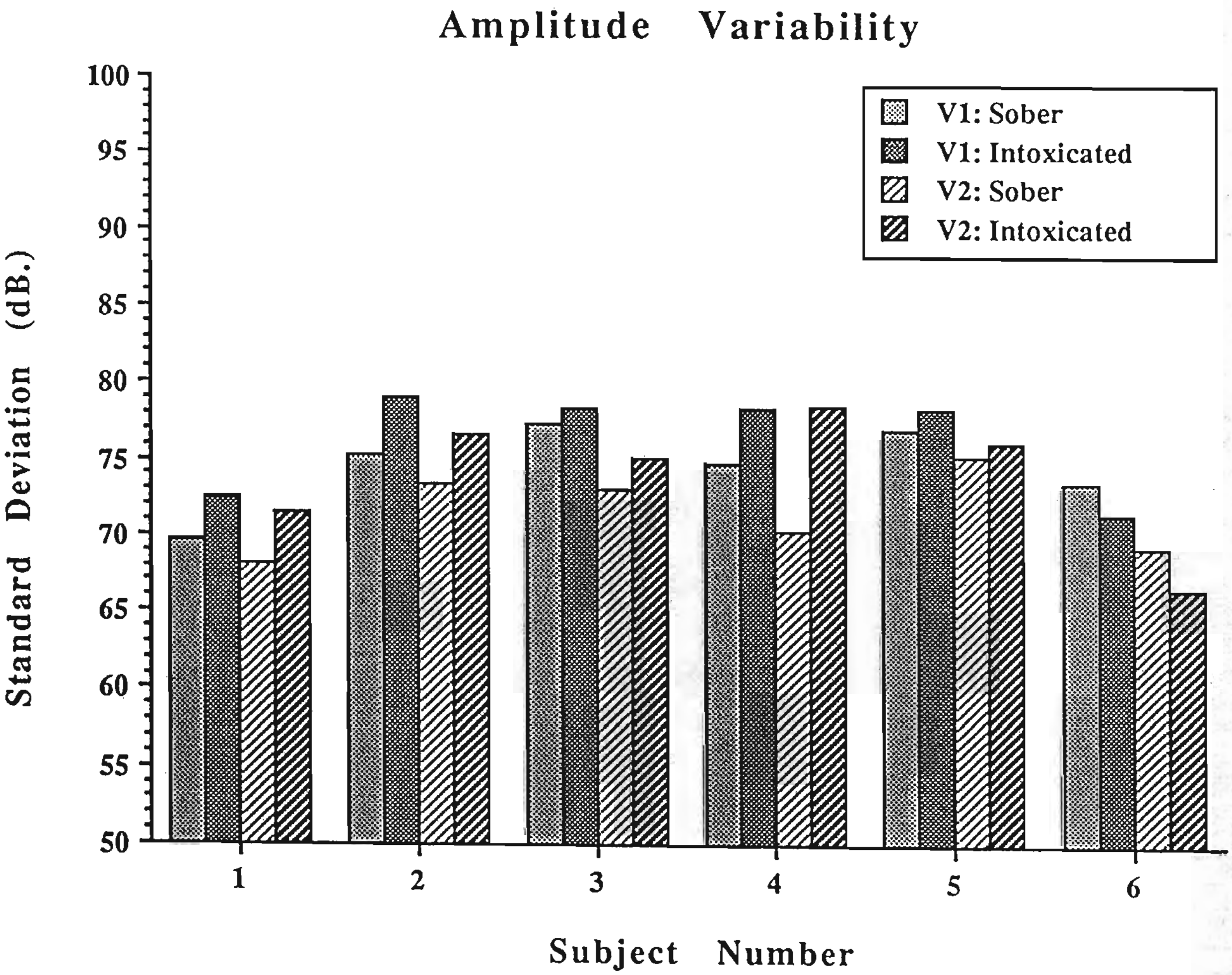


Figure 7. Mean amplitude variability of V1 and V2 produced by six subjects in the sober and intoxicated condition.

Table 6

Condition means, F-values, and probabilities for F0 measures.

Measure	**Sober (Mean)**	**Intoxicated (Mean)**	***F* Value**	***p***
V1: Mean F0	130.6 Hz.	139.4 Hz.	109.94	0.0001
V2: Mean F0	113.9 Hz.	124.3 Hz.	229.57	0.0001
V1: F0 Variability	3.4 Hz.	5.0 Hz.	14.49	0.0001
V2: F0 Variability	4.1 Hz.	5.8 Hz.	13.88	0.0001

of the spondee, but not for the second. For Subject 6, the mean amplitude of both vowels of the spondee decreased with intoxication.

Alcohol also had a significant effect on amplitude variability; amplitude variability was greater when subjects were intoxicated than when they were sober. However, as Figure 7 illustrates, alcohol did not affect all subjects in the same way. Whereas both vowels of the spondees produced by Subjects 1-5 in the intoxicated condition showed an increased amplitude variability, the amplitude variability of both vowels decreased for Subject 6.

The results showed both mean amplitude and amplitude variability to increase with intoxication. However, alcohol affected these measures differently for different subjects. For most subjects, both amplitude measures decreased with alcohol consumption, but the opposite result was found for Subject 6.

Fundamental Frequency. Mean fundamental frequency and fundamental frequency variability were measured for both vowels of the spondees. The condition means and F-values for fundamental frequency and its variability are presented in Table 6. The subject means are shown in Figures 8 and 9 respectively.

Insert Table 6 about here

Insert Figures 8-9 about here

Fundamental frequency was significantly greater in the intoxicated condition compared to the sober condition for both vowels. The general increase in mean F0 associated with intoxication is shown in Figure 8. For both vowels produced by Subjects 1-4, the mean F0 increased. The mean F0 of the V1 produced by Subject 5 increased slightly, whereas the mean F0 decreased in both vowels of the spondees produced by Subject 6.

Alcohol also significantly affected F0 variability; F0 variability was greater in the intoxicated than the sober condition for both vowels of the spondees. F0 variability for each subject's productions is presented in Figure 9. Subjects produced the vowels with greater F0 variability in the intoxicated condition than in the sober condition, with the exception of V2 for Subjects 3 and 6; F0 variability of V2 for Subject 3 showed no clear difference between conditions, but for Subject 6 F0 variability was lower in the intoxicated condition.

Both mean F0 and F0 variability increased with intoxication, but the results varied among subjects. While both F0 measures increased for most subjects, neither mean F0 nor its variability increased for Subject 6.

Duration. Segment durations can serve as a general measure of speaking rate. The duration of both vowels of the spondees and the intervowel duration were measured. The condition means and F-values for these measures are presented in Table 7. The subject means for the vowel and intervowel durations are shown in Figures 10 and 11 respectively.

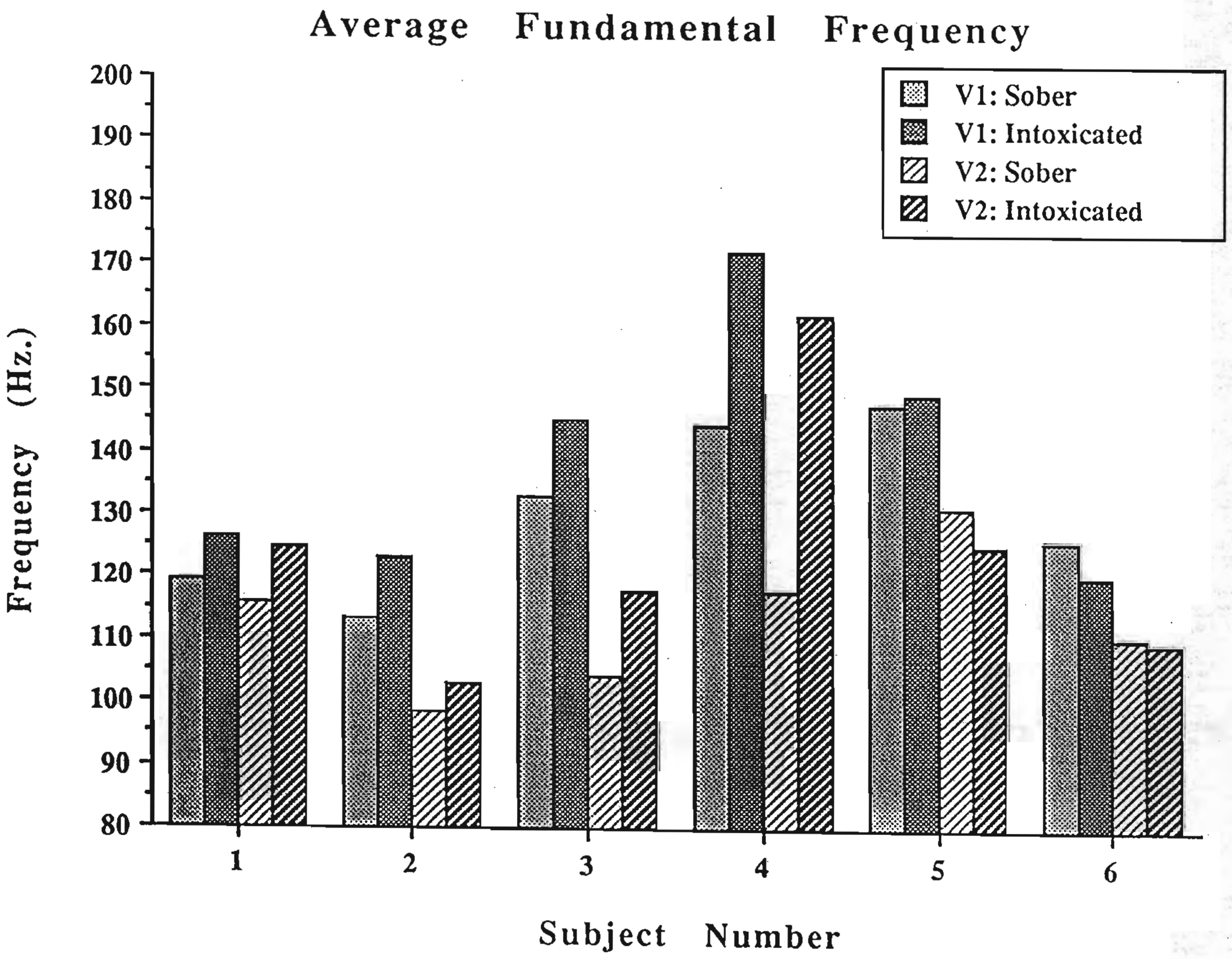


Figure 8. Mean F0 of V1 and V2 produced by six subjects in the sober and intoxicated condition.

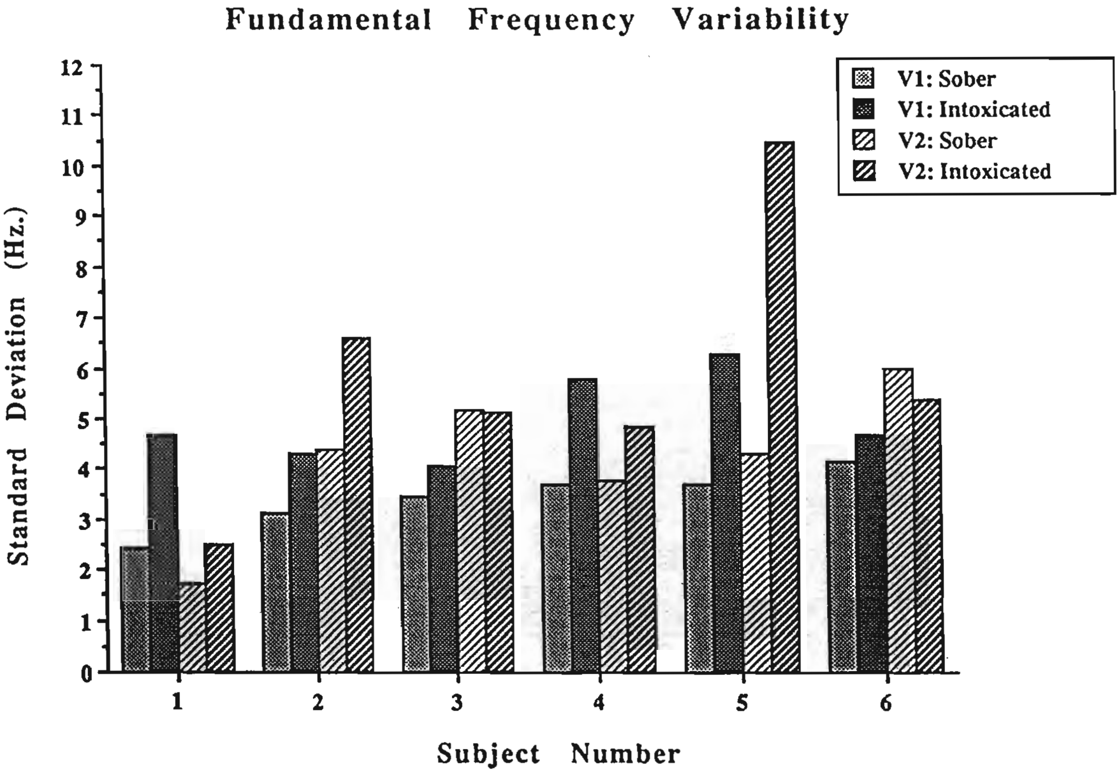


Figure 9. Mean F0 variability of V1 and V2 produced by six subjects in the sober and intoxicated condition.

Insert Table 7 about here

Insert Figures 10-11 about here

Alcohol consumption did not have a significant effect on the vowel of either V1 or V2. However, intervowel duration was significantly reduced in the intoxicated condition, compared with the sober condition. In Figure 11 the subject means show that the intervowel duration was shorter for Subjects 1, 2, 4, and 5 in the intoxicated condition. For Subject 3, the intervowel duration was greater for the intoxicated condition than the sober condition, whereas for Subject 6 the intervowel duration for the two conditions was the same.

Discussion

The goal of this project was to investigate the effects of alcohol on the production of isolated spondees. The acoustic measures investigated revealed that several acoustic parameters are affected by intoxication.

Alcohol decreased the average frequency of both F2 and F3, suggesting that vowel quality is affected. A decrease of F2 is associated with articulations made with the point of articulatory constriction further back toward the glottis. Similarly, F3 tends to decrease slightly as the point of constriction is further back toward the glottis, and as the mouth opening decreases in size and/or becomes more rounded. Given the previous research showing that consonants produced under the influence of alcohol are articulated further back in the vocal tract, it is not surprising that vocal tract constriction in vowel production would also be retracted.

The back of the tongue is known to have less flexibility and less potential for fast movement than the tip of the tongue. If consonant and vowel articulations tend to be produced further back in the vocal tract when a speaker is intoxicated, increased segmental durations and/or decreased speaking rate would be expected. Although this has been reported in previous investigations, the duration analyses in the present study showed that alcohol generally reduced intervowel durations and intervowel-to-word duration ratios, and measures of vowel duration were unaffected by alcohol. The difference between the findings of the present investigation and previous research may be due to the differences in utterance lengths; past studies have examined the effects of alcohol on sentence length utterances, whereas the present investigation analyzed isolated words. The utterance lengths of the spondees produced without carrier sentences may have been too short to be able to clearly reflect speaking rate.

Alcohol was also found to affect measures of amplitude and F0; mean amplitude, amplitude variability, mean F0, and F0 variability all increased when subjects were intoxicated. The results for mean amplitude conflict with previous research which suggests that amplitude decreases with intoxication. However, since a speaker is not likely to lose control of mean amplitude or mean F0 unless the conditions are extreme, the conflicting results are most likely due to individual differences. However, the present results support earlier reports (Pisoni et al., 1986, 1989) that F0 variability increases with intoxication, and further demonstrate an increase in amplitude variability. This increased laryngeal variability may be due to the effect of alcohol on the mucosa of the vocal folds. Klingholz, Penning and Liebhardt (1988) have suggested that alcohol may cause swelling and desensitization of the vocal folds which would disturb normal vocal fold vibration and increase amplitude and F0 variability.

Table 7

Condition means, F-values, and probabilities for duration measures.

Measure	**Sober (Mean)**	**Intoxicated (Mean)**	***F* Value**	***p***
V1 Duration	0.09 msec.	0.09 msec.	0.08	n.s.
V2 Duration	0.14 msec.	0.14 msec.	0.26	n.s.
Intervowel Duration	0.15 msec.	0.13 msec.	9.74	0.002

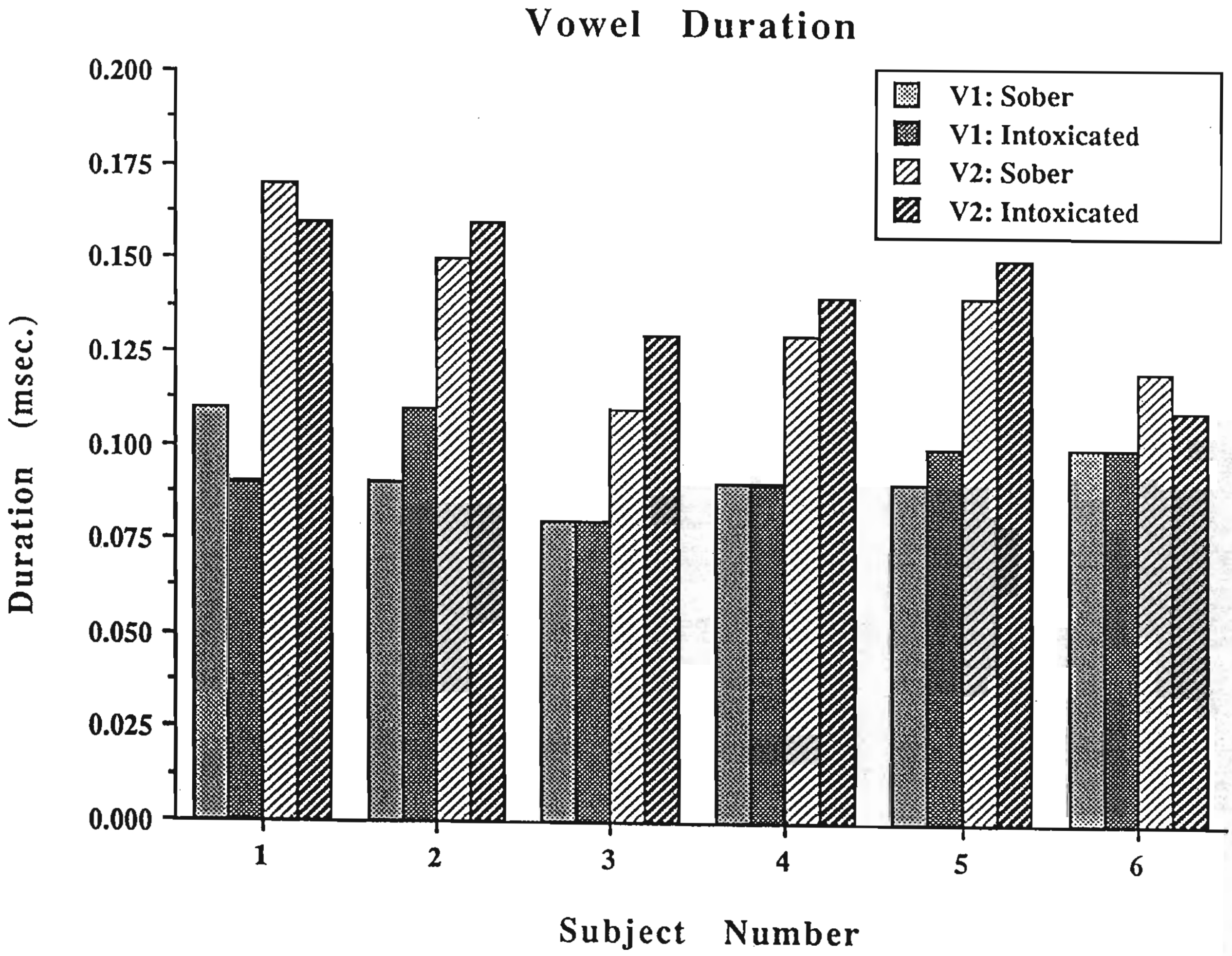


Figure 10. Mean vowel duration of V1 and V2 produced by six subjects in the sober and intoxicated condition.

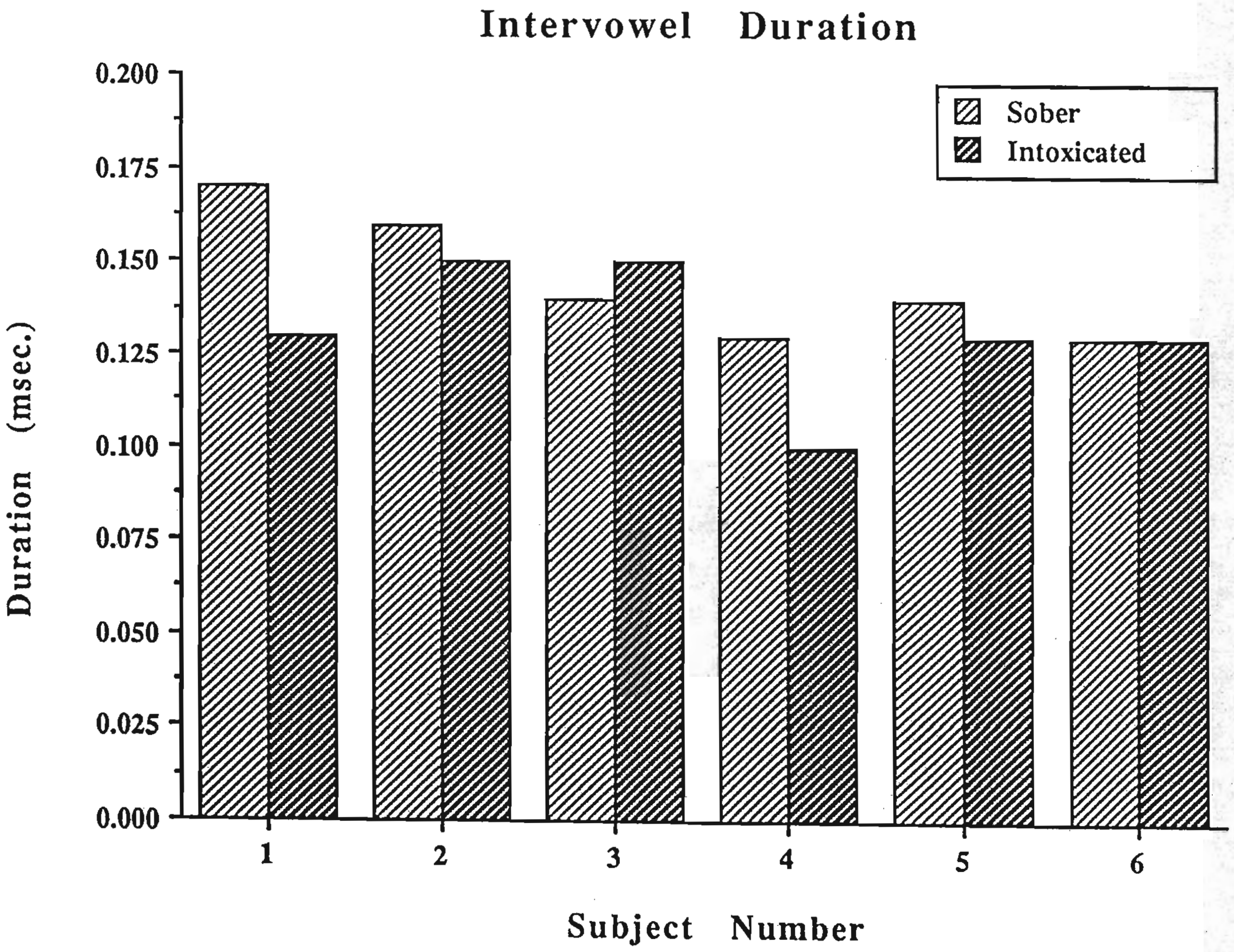


Figure 11. Mean intervowel duration of V1 and V2 produced by six subjects in the sober and intoxicated condition.

Finally, the goal of any study investigating the effects of alcohol on speech is to identify characteristics of speech which can reliably discriminate between speech produced by a sober or intoxicated talker. The subjects in the present study consumed enough alcohol to be considered intoxicated by most legal standards. Yet, despite the acoustic properties generally found for speech produced under the influence of alcohol, there were clear individual differences among the subjects. In particular, the pattern of results for Subject 6 stands out from the other subjects on almost every acoustic measure we examined. At first glance, Subject 6's relatively high BAL (0.16-0.10%) throughout the experiment suggests that his high level of intoxication had a pervasively different effect on his speech; however, Subject 1's BAL was slightly higher (0.17-0.10%) and exemplified the acoustic measures found with other subjects. Although these findings emphasize the regularity of the effects of alcohol on speech found for the other subjects, they also illustrate the need to investigate the effects alcohol over a wider range of experimental conditions with a larger sample of subjects.

In summary, the results of this study demonstrated that in the production of isolated spondees, intoxication affects acoustic measures of phonetic segments and prosody. Although alcohol appears to affect talkers in different ways, speech tends to be produced further back in the vocal tract with a higher amplitude and F0, and with greater variability when the talker is intoxicated.

References

American Medical Association Committee on Medicolegal Problems (1968). *Alcohol and the Impaired Driver.* Chicago: American Medical Association.

Andrews, M.L., Cox, W.M., & Smith, R.G. (1977). Effects of alcohol on the speech of non-alcoholics. *Central States Speech Journal,* **28**, 140-143.

Gough, H.G. (1969). *Manual for the California Psychological Inventory.* Palo Alto, California: Consulting Psychologists Press.

Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from the speech signal. *Journal of the Acoustical Society of America*, **84**, 929-935.

Lester L., & Skousen, R. (1974). The phonology of drunkenness. In Bruck, A., Fox, R.A., & LaGaly, M.W. (Eds.), *Papers from the Parasession on Natural Phonology*. Chicago: Chicago Linguistics Society.

MacAndrew, C. (1965). The differentiation of male alcoholic outpatients from nonalcoholic psychiatric outpatients by means of the MMPI. *Quarterly Journal of Studies on Alcohol*, **26**, 238-246.

Pisoni, D.B., Hathaway, S.N., & Yuchtman, M. (1985). Effects of Alcohol on the acoustic-phonetic properties of speech: Final Report to G.M. Research Laboratories. Bloomington, IN: Speech Research Laboratory, Indiana University.

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**(4), 577-587.

Pisoni, D.B., Yuchtman, M., & Hathaway, S.N. (1986). Effects of alcohol on the acoustic-phonetic properties of speech. In *Alcohol, Accident and Injuries*. Warrendale, PA: Society of Automotive Engineers, 131-150.

Selzer, M.L., Vinokur, A., & Van Rooijen, L. (1975). A self-administered short Michigan Alcoholism Screening Test (SMAST). *Journal of Studies on Alcohol*, **36**, 117-126.

Sobell, L.C., & Sobell, M.B. (1972). Effects of alcohol on the speech of alcoholics. *Journal of Speech and Hearing Research*, **15**, 861-868.

Sobell, L.C., Sobell, M.B., & Coleman, R.F. (1982). Alcohol-induced dysfluency in nonalcoholics. *Folia Phonetica*, **34**, 316-323.

Trojan, F., & Kryspin-Exner, V. (1968). The decay of articulation under the influence of alcohol and paraldehyde. *Folia Phoniatrica*, **20**(4), 217-238.

Zalmov, K. (1969). Die Sprachstorungen als Kriterium der Berwusstseinstrubungern. *Psychiatrie, Neurologie, und Medizinische Psychologie*, **21**, 218-225.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Compensation for Talker Variability and Vowel Variability in the Perception of Fricatives[1]

Keith A. Johnson[2]

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

[1] I appreciate the comments of Van Summers, John Mullennix, John Logan, Steve Goldinger, David Pisoni and other members of the Speech Research Laboratory at Indiana University. Denise Beike helped collect the data. This research was supported by NIH Training Grant #DC-00012-12 to Indiana University.

[2] Now at Department of Linguistics, UCLA, Los Angeles, CA 90024.

Abstract

Two processes of perceptual compensation for variability in speech were compared. Previous research has shown that listeners' identifications of synthetic fricative noises are influenced by both rounding on the adjacent vowel and by the identity of the talker who produced the following vowel. The generalization is that variability which indicates a lengthening of the vocal tract during the production of the fricative (vowel rounding, or male talker) results in fewer "sh" responses to the items on a [s] - [š] continuum; that is, the listeners expect generally lower vocal tract resonances from longer vocal tracts. The experiments reported here indicated that although listeners appear to treat vocal tract lengthening in the same way regardless of its physical cause, the mechanisms underlying these perceptual compensations are different. The relevant experimental manipulations involved presentation type and interstimulus interval. When tokens produced by different talkers were presented blocked by talker, the compensation for talker differences was reduced, and the degree of reduction depended on interstimulus interval. However, neither of these manipulations had an impact on compensation for coarticulation. The data suggest that perceptual compensation for talker differences is an active process.

Compensation for Talker Variability and Vowel Variability in the Perception of Fricatives

It has been demonstrated that the perception of fricatives in American English is affected by coarticulatory rounding on the fricative (Mann & Repp, 1980; Whalen, 1981), and that there is a similar perceptual effect as a result of talker variation (May, 1976; Mann & Repp, 1980). The experiments described in this paper were conducted to test these two seemingly similar perceptual compensatory effects.

The effects of coarticulatory rounding on the spectra of particular utterances of [s] and [š] in American English are shown in Figure 1. The top panel illustrates that one of the major acoustic differences between [s] and [š] is spectral. The spectrum for [š] is displaced down in frequency as compared with [s] (For further discussion, see Strevens, 1960 and Hughes & Halle, 1956). The middle panel shows the effect of coarticulatory rounding on [š], and the bottom panel shows the coarticulatory rounding effect on [s]. In both cases the spectrum of the fricative produced in the environment of the rounded vowel [u] is displaced down in frequency as compared with the spectrum of the fricative produced in the environment of the unrounded vowel [a]. Figure 2 (from Carney & Moll, 1971) shows the positions of the articulators during the production of [s] in three vocalic environments. This figure illustrates that the main difference between [s] in the environment of [a] and [s] in the environment of [u] is almost entirely due to a difference in the position of the lips during the fricative. By narrowing and lengthening the cavity between the lips, the resonance frequencies of the front part of the vocal tract are lowered (Heinz & Stevens, 1961).

Insert Figures 1 and 2 about here

Listeners' identifications of fricative sounds in American English are influenced by the quality of the following vowel. For example, Mann and Repp (1980) found that if the vowel following a fricative is produced with rounded lips, the listener responds as if he/she expects the spectrum of the fricative to be transposed down in frequency as compared to the same segment produced in the context of an unrounded vowel. Mann and Repp (1980) also found that the presence of a silent gap between the fricative and the vowel reduced the effect of the vowel on the perception of the fricative; when the fricative was adjacent to the vowel there was a large effect of vowel quality on the boundary between [s] and [š], but when the two segments were separated by a silent interval, the effect of vowel quality was greatly reduced. This phenomenon can be described by asserting that the listener is compensating for the coarticulatory changes in fricative production which occur when the "same" fricative is produced in different vowel environments. Whalen's (1981) findings tend to support this characterization of the effect of vowel quality on fricative perception. He found that vowels with an acoustic vowel quality between unrounded [i] and rounded [u] tended to be associated with a boundary on an [s] - [š] continuum which was between the boundaries found when the post-consonantal vowels were [i] and [u]. This finding suggests that the listener may compensate for the effects of vowel rounding even when the vowels in question are not normally used in his/her native language, however, this conclusion is only partially supported by Whalen's (1981) data because the results were not consistent (See his Figure 5.). So, Mann and Repp (1980), and Whalen (1981) found that the perceptual boundary between [s] and [š] is shifted down in frequency when a continuum of fricative noises is presented in the context of a rounded vowel as compared with the same continuum presented in the context of an unrounded vowel, and that this effect reflects an apparent perceptual compensation for coarticulatory rounding in fricatives.

Another type of variation which affects the perception of fricatives is talker variation. The same physiological differences which result in acoustic differences between vowels produced by men and

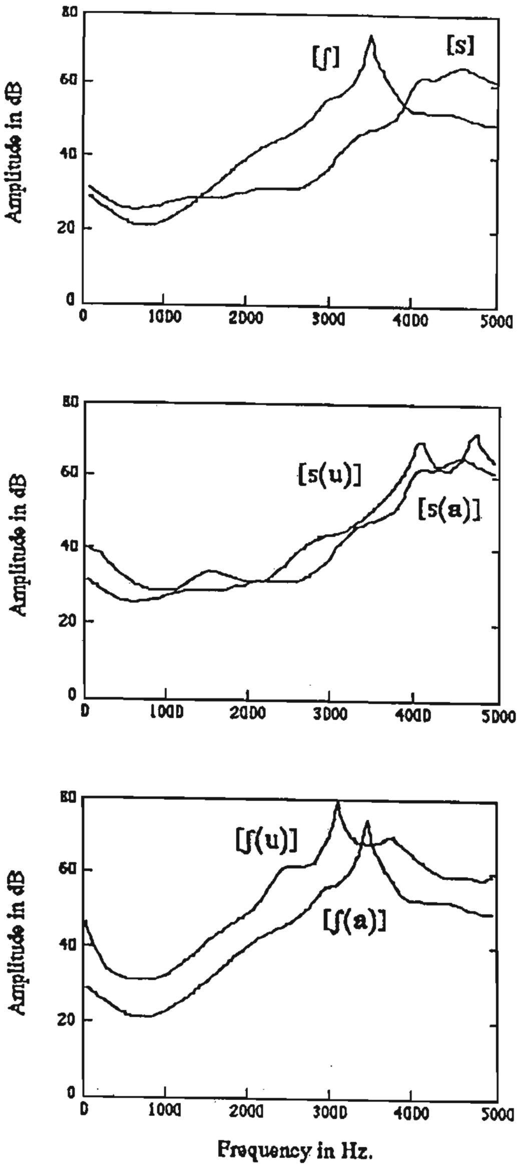


Figure 1. Power spectra of the fricatives [s] and [š] in various vowel environments. All tokens were produced by a male speaker. Top panel: The [s] of 'saw' compared with the [š] of 'shah'. Middle panel: The [s] of 'sue' compared with the [s] of 'saw'. Bottom panel: The [š] of 'shoe' compared with the [š] of 'shah'.

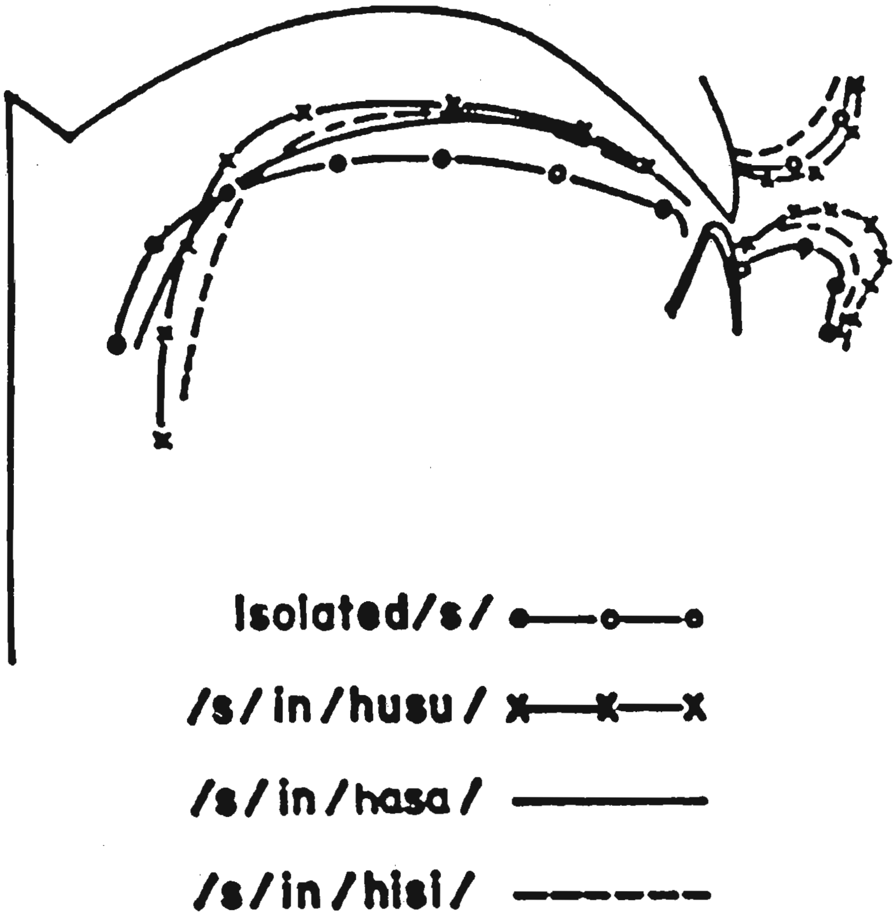


Figure 2. Positions of the articulators during intervocalic [s] in three vowel environments and in isolation. (Taken from Carney & Moll, 1971.)

women (Peterson & Barney, 1952) are the source of gender differences in the acoustics of fricatives. Because men tend to have longer vocal tracts than women, the acoustic resonances of men's vocal tracts tend to be lower in frequency than are those of women's vocal tracts. This is illustrated in Figure 3, in which spectra of [s] and [š] as produced by a male and female talker are compared. The acoustic consequences of the typical difference in vocal tract length parallels the difference seen in Figure 2 for fricatives produced in the environment of rounded and unrounded vowels. The generalization is that the spectra of fricatives produced by a male talker and in the environment of a rounded vowel are shifted down in frequency as compared with the same fricative produced by a female talker or in the environment of an unrounded vowel. Stated more generally, fricatives produced by a longer vocal tract (due to sex differences or lip rounding) are shifted down in frequency as compared with fricatives produced by a shorter vocal tract.

Insert Figure 3 about here

The perceptual consequences of talker variation on vowel perception have been extensively studied (Johnson, 1990a; Nearey, 1989), but there is also evidence that listeners' identifications of fricatives are influenced by perceived talker differences. For example, May (1976) found that when vowel formant frequencies were shifted from values appropriate for a male speaker to values appropriate for a female speaker, listeners' identifications of syllable initial fricatives (from [s] to [š]) were affected. Mann and Repp (1980) also investigated the effect of talker variation on the perception of [s] and [š]. They spliced the members of a synthetic fricative continuum onto the vowels [a] and [u] as produced by a male and female talker. They found a reliable difference in the boundary between [s] and [š] as a function of the gender of the talker, with the boundary occurring at a lower frequency when the fricatives were presented in the environment of vowels produced by a male. Interestingly, when there was a gap between the fricative noise and the following vowel, the perceptual effects of coarticulation were substantially reduced (as was mentioned above), but the effect of talker variability, although reduced, was of the same order of magnitude as it was in the no gap condition. So, the perceptual consequences of talker variability roughly parallel the perceptual consequences of vowel variability. The listener appears to perceptually compensate for these two sources of variability in the same way. If the speech signal presents evidence concerning the length of the vocal tract, that information plays a role in the perceptual identification of the fricative noise.

The experiments reported in this paper further explored these two types of perceptual compensation. In particular, I was interested in whether these two seemingly similar perceptual processes are different in any way.

There are a couple of a priori reasons to expect that compensating for talker variability and compensating for coarticulation differ. First, the effects of the two sources of variation have different degrees of reliability in speech production. Coarticulatory changes always occur when segments are concatenated with each other. While, talker related changes are not so reliable - some people with high pitched voices have long vocal tracts, and some people with low pitched voices have short vocal tracts. Second, these two sources of variation pose different perceptual problems for the listener. In order for communication to take place listeners must treat the linguistic information of speech categorically, while talker variation and the personal information that it conveys does not have to be treated categorically for successful speech communication.

To test the hypothesis that perceptual compensation for talker variability and perceptual compensation for coarticulation differ, I investigated the perception of a synthetic fricative continuum from [s] to [š] in two experiments: the first investigated the effect of talker variation on the perception

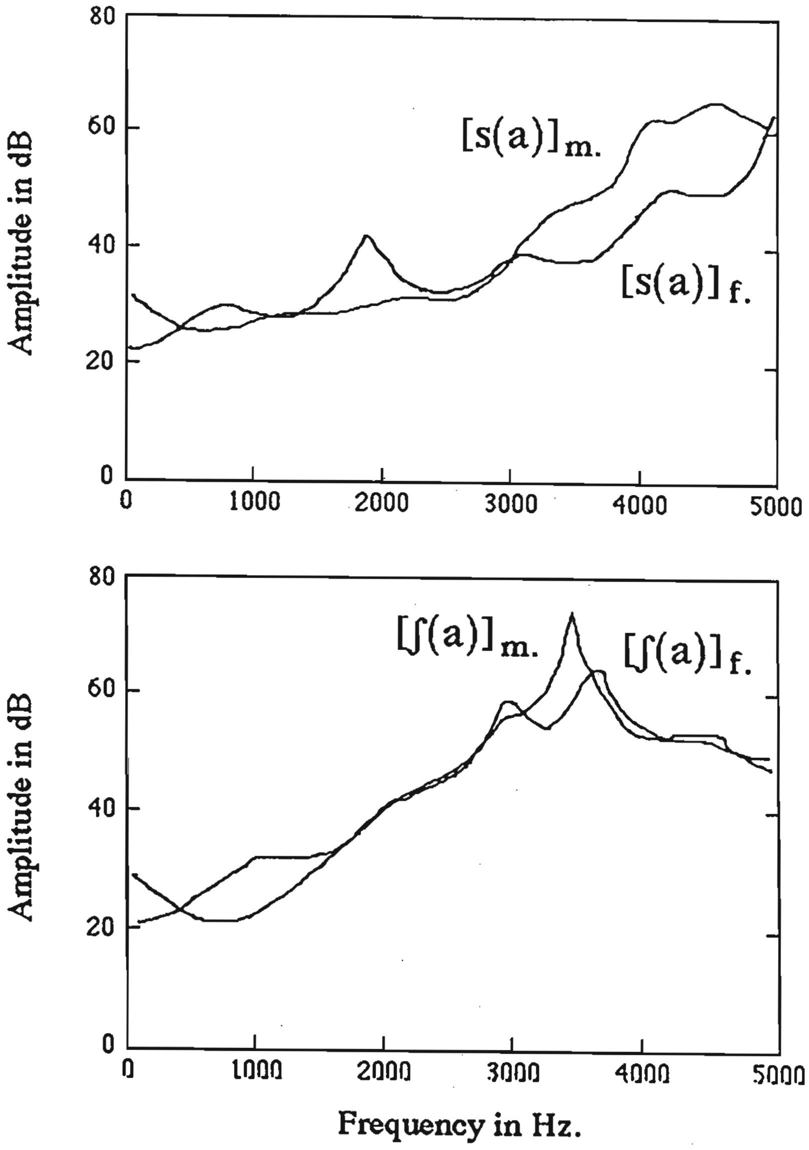


Figure 3. Power spectra of fricatives produced by a male and a female talker. Top panel: Male and female productions of the [s] in 'saw'. Bottom panel: Male and female productions of the [s] in 'shah'.

of the fricative continuum, the second investigated the effect of rounding coarticulation on the perception of the fricative continuum. Both experiments involved two manipulations in addition to talker or vowel variation. Johnson (1990b) found that perceptual vowel normalization is affected by a manipulation of presentation type. When vowels having different fundamental frequencies were randomly intermixed with each other, listeners displayed the type of identification behavior which has been described as F0 normalization; that is, the listeners behaved as if they expected the items synthesized with high F0 to have been produced by a shorter vocal tract than those items synthesized with a low F0. However, when these same tokens were presented blocked by F0, the F0 normalization effect was not seen. I hypothesized (in Johnson,1990b) that this finding reflects the operation of a talker contrast effect and presented the results of an experiment which seemed to support this interpretation. Whatever the correct interpretation of this result, this manipulation may be a useful tool in empirically separating effects of talker variation and vowel variation in the perception of fricatives. The second manipulation involved the inter-stimulus interval (ISI). The results in Johnson (1990b) suggested that the listener may hold a representation of the talker in memory from one token to the next in an identification experiment, so increasing the interval between trials may reduce the listener's ability to hold over talker information from one trial to the next.

EXPERIMENT 1: Talker Variability

Methods

Subjects

Forty undergraduate psychology students from Indiana University participated in the experiment in fulfillment of a course requirement. None of the listeners reported any history of speech or hearing difficulty and all were native speakers of American English. The listeners were randomly divided into four groups of ten as discussed below.

Materials

A nine step synthetic continuum from [š] to [s] was produced using the Klatt software formant synthesizer (Klatt & Klatt, 1989). The control parameters for the synthesizer are shown in Table 1. The fricative continuum is essentially the same as the one used by Mann and Repp (1980) although the step sizes were equated.[3]

Insert Table 1 about here

The synthetic fricatives were concatenated with four different naturally produced vowels. Using a digital waveform editor, the vowels were spliced from the words "shah" and "saw" as produced by a male and a female. Figure 3 shows spectra of the original fricatives in these words. The acoustic properties of these naturally produced vowels are shown in Table 2. The peak RMS amplitudes of the vowels were equated and then each of the fricative noises was concatenated with the vowels, producing 4*9=36 stimuli. Before being concatenated with the vowels, the peak RMS amplitudes of the synthetic fricative noises were equated at a level 15 dB below the level at which the vowels had been set.

Insert Table 2 about here

----

[3] Mann & Repp were limited by the frequency intervals available on the OVE synthesizer.

Table 1

Synthesizer control parameters
for the synthetic [š] - [s] continuum.[4]

	[š] **1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**[s]** **9**
F3	2466	2614	2771	2936	3111	3296	3492	3698	3917
F4	3108	3294	3491	3698	3918	4151	4396	4657	4932

[4] Step sizes for F3 and F4 are in Bark. Delta Z of F3=0.396. Delta Z of F4=0.398. Bandwidth of each resonance=0.1 (resonance center frequency). Total duration=140 ms. Ramp on=70 ms. Steady amplitude=35 ms. Ramp off=35 ms.

Table 2

Acoustic properties of the naturally produced vowels used in Experiments 1 and 2.

Measure	EXPERIMENT 1		EXPERIMENTS 1 & 2		EXPERIMENT 2	
	Female		Male		Male	
	"saw"	"shah"	"saw"	"shah"	"Sue"	"shoe"
Duration	304	324	384	382	334	316
Mean F0	211	214	104	103	108	120

Procedure

The experiment was conducted on-line at the Speech Research Laboratory at Indiana University, Bloomington. Stimulus presentation, randomization and response collection were performed by a PDP 11/43 computer. Subjects were run in groups of four to six listeners. The subjects' task was to identify the synthetic fricative in each stimulus by pushing buttons labeled 's' and 'sh'.

Two groups of ten listeners identified the fricatives in stimuli blocked by talker. The listeners in these groups heard a block of stimuli in which the vowels had been produced by one of the talkers and then another block of stimuli in which vowels had been produced by the other talker. The order of the blocks was counter-balanced across subjects. The interstimulus interval for one group in this blocked condition was between 1400 and 2000 ms and the interstimulus interval for the other group was between 3400 and 4000 ms. The actual length of the interstimulus interval was dependent upon reaction time, with the minimum ISI set to 1000 and 3000 ms for the two groups respectively. Two other groups of ten listeners responded to blocks of trials in which all 36 stimuli had been randomly intermixed. Thus, for these two groups of listeners, the identity of the talker was unpredictable from trial to trial. One of the two groups in this mixed condition had an ISI of at least 1000 ms, while the other group had an ISI of at least 3000 ms.

All listeners responded to each of the 36 stimuli 10 times, the order of the stimuli being separately randomized for each pool of four to six listeners.

Results

The identification data (summed across the continuum) were analyzed in a four way analysis of variance with between-subjects factors ISI (1000 ms versus 3000 ms) and presentation type (blocked versus mixed), and within-subjects factors gender (male versus female speakers) and original consonant context ([s] versus [š]).

As has been found in previous research, the original consonantal context of the vowels influenced subjects' labelling behavior [$F(1,36)=219$, $p<.01$]. When the vowel had been produced in the context of [s] the percent 'sh' responses was 49.9%, while when the vowel had been produced in the context of [š] the percent 'sh' responses was 66.6%. Similarly, as has been reported before, the gender of the speaker had an impact on fricative perception [$F(1,36)=126.69$, $p<.01$]. When the vowel had been produced by the male speaker 52.8% of the responses were 'sh', while when the vowel had been produced by the female speaker 63.8% of the responses were 'sh'. This indicates that the listeners expected the vocal tract resonances to be generally lower in frequency for the male speaker and, thus, that more of the stimuli in the synthetic continuum were acceptable tokens of 's' when the voice was male than when it was female; that is, when the context indicated that the fricative had been produced with a longer vocal tract, due to the gender of the talker, listeners identified fewer of the synthetic stimuli as "sh").

Presentation type had a reliable effect on this gender effect [$F(1,36)=25.02$, $p<.01$]. This interaction is shown in Figure 4. The difference between fricative identification in the environments of vowels produced by a male and a female talker was smaller when the stimuli were presented in the blocked condition as opposed to the same stimuli as they were identified in the mixed condition. This is consistent with the finding reported by Johnson (1990b) in which the F0 normalization effect in vowel perception was reduced when stimuli were presented blocked by F0.

Insert Figure 4 about here

There was a marginally significant interaction between the gender and ISI factors [$F(1,36)=7.29$, $p<.05$]. This interaction is shown in Table 3. Although the difference between the gender effect with an ISI of 1000 ms and the gender effect with an ISI of 3000 ms is small, the trend represented by this interaction is for the gender effect to be larger at the longer ISI. This effect is easier to understand when we consider another marginally significant interaction; namely the three-way interaction between gender, ISI and presentation type [$F(1,36)=3.85$, $p<.06$]. As shown in Figures 5 and 6, the effect of presenting the stimuli blocked by talker varied as a function of ISI. It had been expected that the difference between blocked and mixed presentation might be reduced when the interstimulus interval was increased, and that appears to have happened. However, the particulars are a little surprising. On the analogy with vowel contrast effects, we might expect that the addition of a longer ISI would result in a reduced amount of contrast and thus that the main difference between long and short ISI in this experiment would be seen in the mixed presentation type. As Figure 5 indicates though, the number of "sh" responses in the mixed condition was virtually unchanged as a function of ISI, rather the difference between short and long ISI was to be found almost entirely for the blocked condition. At short ISI the gender of the talker had no effect on fricative identification in the blocked condition, while these same stimuli in the long ISI, blocked condition gave rise to a substantial normalization effect. So, the different effect of gender at the two ISI's (Table 3) appears to be a result of the effect of ISI on the perception of the stimuli in the blocked condition.

Insert Table 3 and Figures 5 and 6 about here

Discussion

The results of Experiment 1 replicate several previous findings: (1) the boundary between [s] and [š] depends on the original consonantal context (Whalen, 1981 showed that this effect is primarily due to the formant transitions on the vowel), (2) the boundary between [s] and [š] depends on the gender of the talker, (3) as would be expected from the results reported by Johnson (1990b), the speaker normalization effect is affected by presentation type, when items are presented blocked by speaker the normalization effect is reduced. An additional finding, which will be discussed in more detail in the general discussion, is that the effect of presentation type is reduced when interstimulus interval is increased.

EXPERIMENT 2: Vowel Variability

Experiment 2 was an analog of Experiment 1, the only difference being that the rounding of the vowel was manipulated, rather than the speaker. The main point of interest has to do with the effects of presentation type and ISI on the process of perceptual compensation for rounding coarticulation which has been previously demonstrated (Whalen, 1981; Mann & Repp, 1980).

Methods

Subjects

Forty undergraduate psychology students from Indiana University participated in the experiment in fulfillment of a course requirement. None of the listeners reported any history of speech or hearing

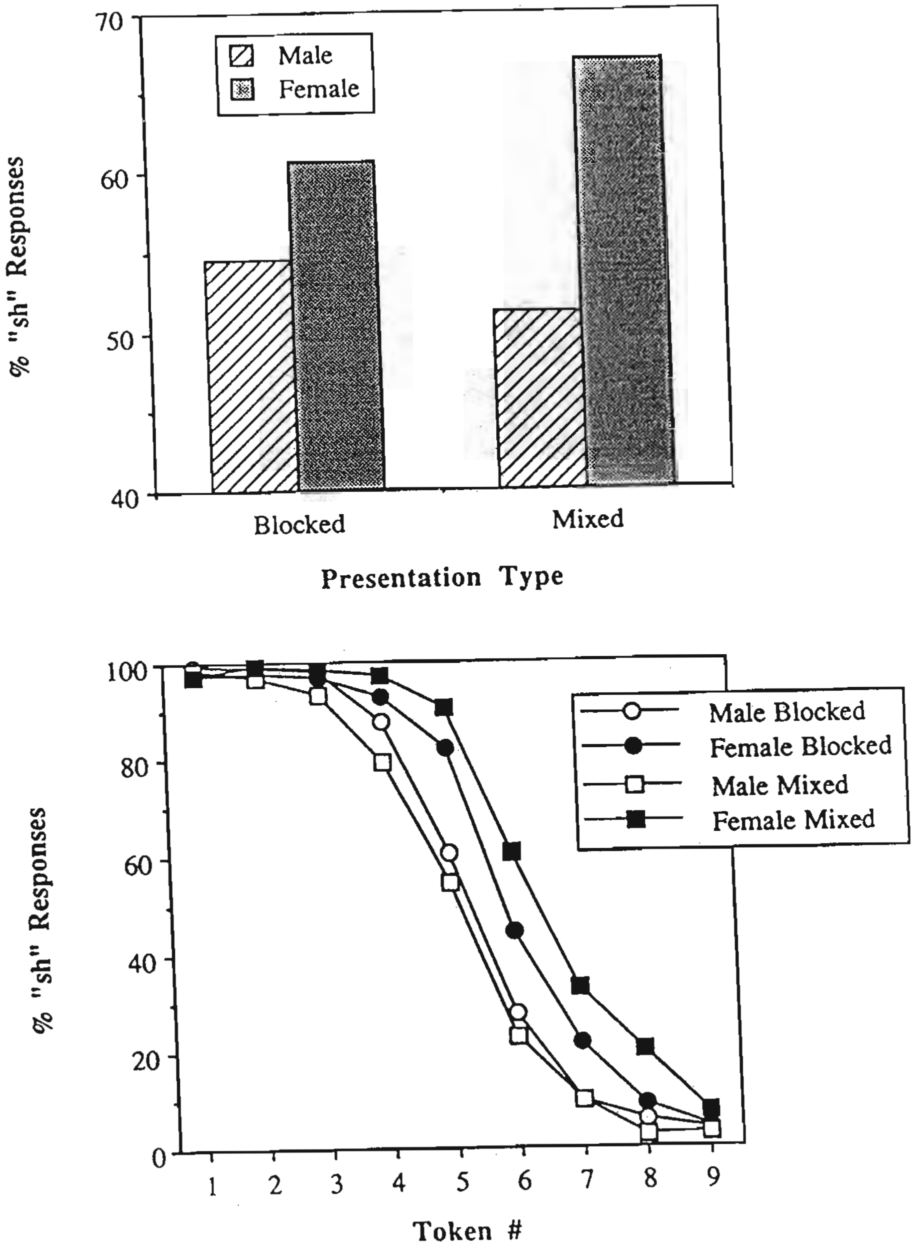


Figure 4. Results of experiment 1. The interaction between the gender and presentation type factors. Top panel: Percent 'sh' responses averaged across the tokens in the [s] - [š] continuum. Bottom panel: Percent 'sh' responses as a function of token number.

Table 3

Gender by ISI Interactions: Percent 'sh' responses.

ISI	**Male**	**Female**	**Difference**
1000	56.2	64.6	8.4
3000	49.3	63.0	13.7

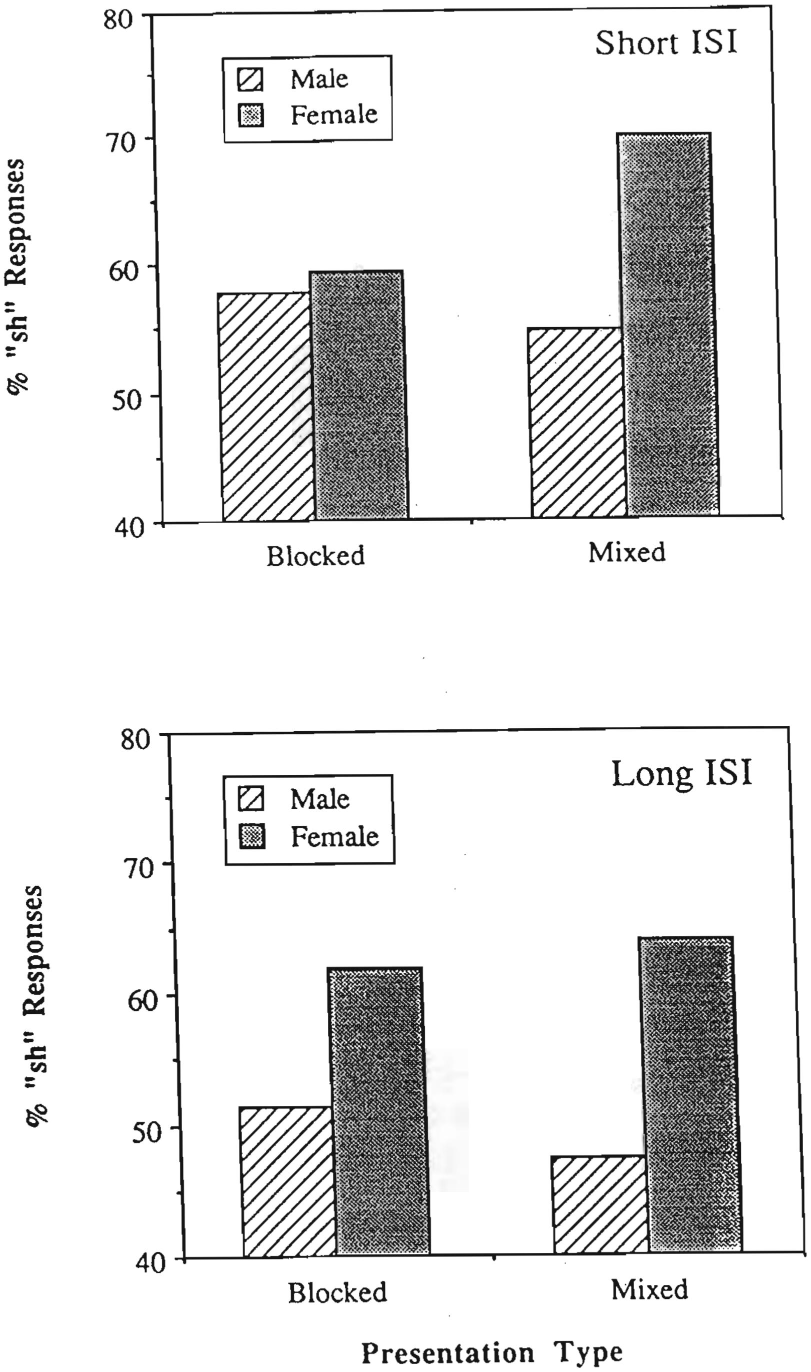


Figure 5. Results of experiment 1. The effect of ISI on the gender by presentation type interaction averaged across tokens. Top panel: short ISI. Bottom panel; long ISI.

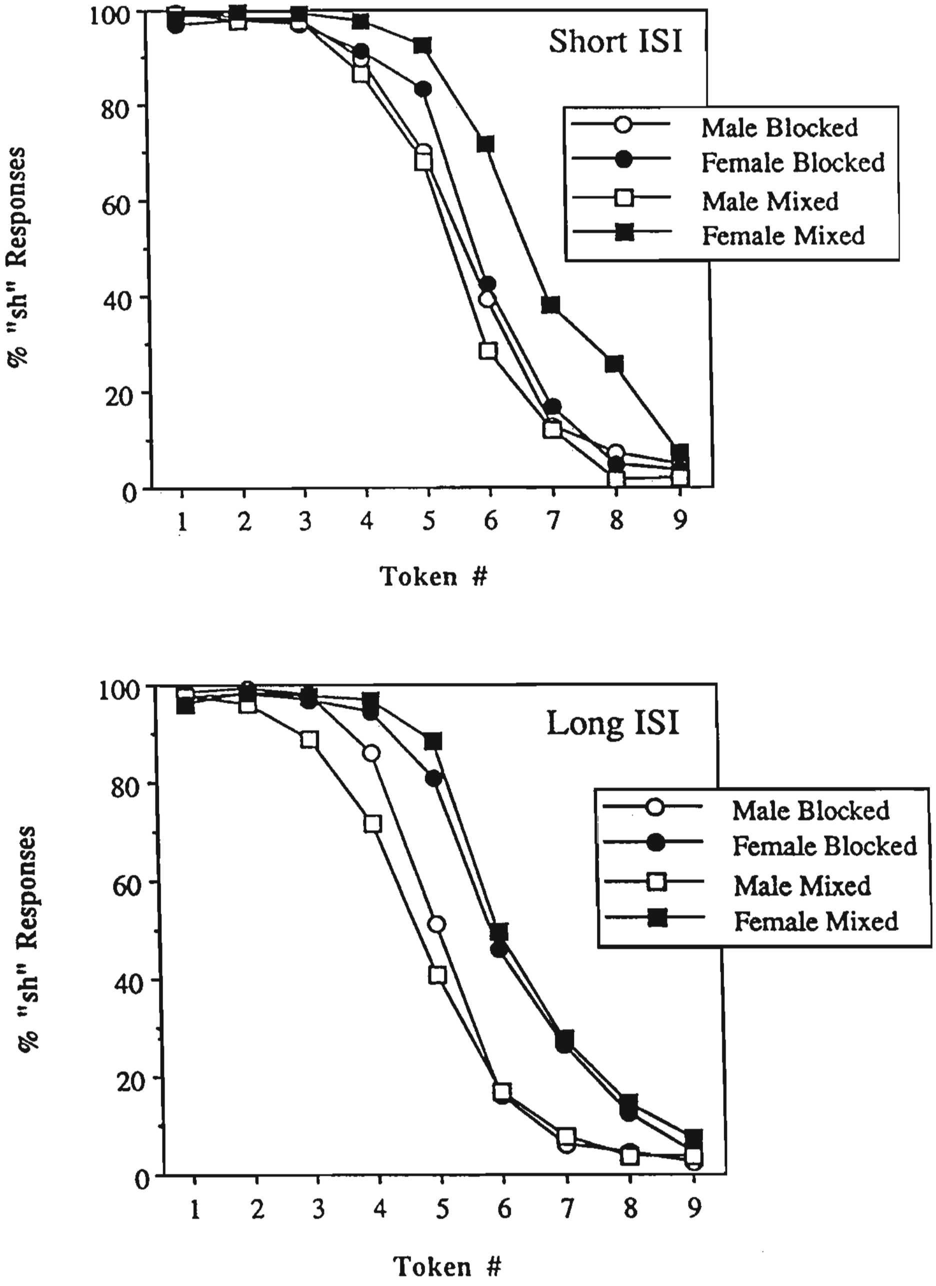


Figure 6. Results of experiment 1. The effect of ISI on the gender by presentation type interaction plotted by tokens. Top panel: short ISI. Bottom panel: long ISI.

difficulty and all were native speakers of American English. The listeners were randomly divided into four groups of ten as discussed below.

Materials

The same nine-step synthetic fricative continuum which had been used in Experiment 1 was used in Experiment 2. These synthetic fricatives were concatenated with four vowels to produce 4x9=36 stimuli. The vowels had been spliced from naturally produced utterances of the words "saw", "shah", "sue" and "shoe", and are described in Table 2. The speaker was a male native speaker of American English. Spectra of the original fricatives in these words are shown in Figure 1. As in Experiment 1, the peak RMS amplitude values of the vowels were equated before concatenating them with the synthetic fricatives, and the fricative noises had a peak RMS amplitude which was 15 dB below the peak level of the vowels.

Procedure

All equipment used in conducting the listening sessions was identical to that used in Experiment 1. The listeners were randomly assigned to one of four groups. Two groups heard the stimuli blocked by vowel (randomized within blocks) at one of two interstimulus intervals (at least 1000 or 3000 ms between successive stimuli). Two other groups heard all 36 stimuli randomly intermixed with each other, and again one group had a short ISI and the other a long ISI.

Results and Discussion

The identification data, summed across tokens in the synthetic continuum, were submitted to a four-way analysis of variance with between-subjects factors ISI (1000 versus 3000 ms) and presentation type (blocked versus mixed), and within-subjects factors original consonant context ([s] versus [š]) and vowel ([a] versus [u]).

There was a main effect for original consonant context [$F(1,36)=222.13$, $p<.01$]. When the vowel had originally been produced in the context of [s] the percent "sh" responses was 41.74%, and when the original consonant context had been [š] the percent "sh" responses was 58.9%. This result is consistent with the original consonant context effect found in Experiment 1 and in previous research (Whalen, 1981; Mann & Repp, 1980).

There was also a main effect for vowel [$F(1,36)=45.89$, $p<.01$]. Synthetic fricatives in the context of [a] were labelled "sh" 54.6% of the time, while those same fricatives in the context of [u] were labelled "sh" only 46.03% of the time. This result is also consistent with previous research and can be described as a perceptual compensation for coarticulation. The only other effect which approached significance was the main effect for ISI [$F(1,36)=7.32$, $p<.02$]. There was a tendency for the fricatives to be identified as "sh" more when the ISI was short (52.9% versus 47.7%, for the short and long ISI's respectively). It is not clear why this difference occurred.

Note that the concept of perceptually compensating for coarticulation entails an assumption about the units of speech perception. Namely, that the hearer must recover phoneme-sized units during speech perception. However, the present data are consistent with the hypothesis that listeners aim toward recovering context-sensitive allophones; that is, that lexical representations are phonetically concrete. In particular, the fact that the perceptual effects of the original consonant context (i.e., vowel formant transitions) and coarticulatory rounding behave similarly under manipulations of presentation type and ISI suggests that the speech perception system treats the two in a very similar way. It seems to be generally true that effects which can be described as involving trading relations are also describable in terms of allophonic representation rather than perceptual compensatory processing. The fact that intrinsic allophony (e.g., vowel formant transitions) and extrinsic allophony (e.g., coarticulatory rounding) are reliably present in the speech signal suggests that the perceptual representations of words may include these regularly occurring properties, rather than being composed of more abstract representations which must be recovered by undoing processes of coproduction.

GENERAL DISCUSSION

The results of these experiments suggest that perceptual compensation for talker variability is influenced by presentation type (blocked versus mixed) and to a lesser extent by interstimulus interval, while perceptual compensation for coarticulation is not.

Johnson (1990b) argued that the effect of presentation type on normalization reflected a talker contrast effect. According to this argument, when items produced by different talkers are randomly intermixed with each other, the relative difference in the perceived identity of the talker is increased by a contrast effect. Thus, it was argued, a normalization effect which makes reference to the perceived identity of the talker will be exaggerated in the mixed presentation. The present data are in general agreement with this account. As in vowel normalization, the process of perceptual normalization of fricatives appears to involve reference to the perceived identity of the talker. The fact that this normalization effect is larger in the mixed condition than it is in the blocked condition suggests (indirectly) that when an experimental manipulation has an impact on the perceived identity of the talker, the normalization effect (which makes reference to talker identity) is impacted. So, these experiments have shown that two seemingly similar perceptual processes (compensation for coarticulation and talker normalization) are, in fact, quite different.

One interesting feature of the experiments reported here is that a basic similarity between normalizing vowels and normalizing fricative noises has been observed. This observation is important to keep in mind because it limits the class of theories of normalization which can adequately account for the data. If it could be demonstrated that vowel normalization and fricative normalization are empirically different in some way, this could be taken as an argument in favor of theories of vowel normalization which involve reference to formant ratios and the ratio between F1 and F0 (Miller, 1989; Syrdal & Gopal, 1986). The data presented here on fricative normalization suggest, however, that a more general theory of normalization is needed; one in which a representation of the speaker can be used in both vowel and fricative normalization.

Finally, the three-way interaction in Experiment 1 of ISI, gender and presentation type (Figures 5 and 6) suggests that the talker contrast account of the presentation type effect must be modified. It had been expected that with increased interstimulus interval the difference between blocked and mixed presentation would decrease, and this was observed, but the data were a little surprising. If we assume that some representation of the talker is held in short-term memory, it is reasonable to suppose that, if the interval of time between presentations is short, the representations of successive talkers may contrast with each other to produce greater perceived differences between talkers than would occur otherwise. Therefore, an explanation of the presentation type effect (both in Experiment 1 and in Johnson, 1990b) in terms of talker contrast predicts that the normalization effect in the mixed condition will, if anything, be reduced as the interstimulus interval is increased. This was not the observed pattern of data in Experiment 1. The magnitude of the normalization effect did not decrease in the long ISI mixed condition, rather the normalization effect was stronger in the long ISI blocked condition. It now appears that the presentation type effect may not be so much a result of talker contrast as of hypernormalization; the tendency of listeners to overcompensate for speaker characteristics when they encounter a new speaker (see also Mann & Repp, 1980, p. 222). The effect of long ISI, on this account, is to force the listener to treat each successive stimulus as if it were produced by a new speaker, and hence to hypernormalize even when the identity of the speaker was held constant from trial to trial. Thus, the results of Experiment 1 (and also of the studies reported in Johnson, 1990a, 1990b) suggest that when the listener encounters a new speaker, he/she must set up some sort of speaker representation which can then be used as a point of reference for processes of speech perception, and the three-way interaction in Experiment 1 can be accounted for if we assume (1) that upon encountering a new speaker listeners hypernormalize, and (2) that the listener's perceptual representation of the speaker is a short term representation.

References

Carney, P.J., & Moll, K.L. (1971). A cinefluorographic investigation of fricative consonant-vowel coarticulation. *Phonetica*, **23**, 193-202.

Heinz, J.M., & Stevens, K.N. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, **33**, 589-596.

Hughes, G.W., & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, **28**, 303-310.

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

Johnson, Keith. (1990b). Contrast and normalization in vowel perception. *Journal of Phonetics*, **18**, 229-254.

Klatt, D., & 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.

Mann, V.A., & Repp, B.H. (1980). Influence of vocalic context on perception of the [š]-[s] distinction. *Perception and Psychophysics*, **23**, 213-228.

May, J. (1976). Vocal tract normalization for /s/ and /sh/. *Haskins Laboratories Status Report on Speech Research, SR-48*. New Haven, CT: Haskins Laboratories, 67-73.

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

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

Peterson, G., & Barney, H. (1952). Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech*, **3**, 32-49.

Syrdal, A., & Gopal, H. (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.

Whalen, D.H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America*, **69**, 275-282.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Effects of Talker Variability in Self-Paced Serial Recall[1]

Stephen D. Goldinger

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

[1] Thanks to David Pisoni and Scott Lively for discussions of this research, and to Luis Hernandez for technical assistance. This research was supported by NIH Research Grant #DC-00111-14 to Indiana University.

Abstract

Goldinger, Pisoni, and Logan (1991) examined serial recall of several types of spoken word lists; words within a list were produced by either a single talker or by multiple talkers, and lists consisted of either "easy" or "hard" words, as determined by a metric of word frequency and similarity neighborhood characteristics. Goldinger et al. manipulated presentation rate and found that talker variability interacted with presentation rate but word confusability did not. We concluded that talker variability affected the rehearsal processes involved in serial recall of spoken word lists. The present experiment complemented the earlier experiment by examining recall of the same spoken word lists using a *self-paced listening* procedure. In this procedure, subjects were able to control the presentation rate of successive items within a list. Both the listening time and the eventual recall were examined. The data showed that both stimulus dimensions affected rehearsal time, but rehearsal time differences were only predictive of recall differences between single- and multiple-talker lists. Rehearsal time differences were not predictive of differences between recall of lists of "easy" vs. "hard" words. The findings both confirm and qualify the earlier conclusions of Goldinger et al. (1991).

Effects of Talker Variability in Self-Paced Serial Recall

Recent experiments have demonstrated that speech perception, spoken word recognition, and word recall are all affected by talker variability. If talker identity changes from trial to trial in an experiment, word recognition is slower and less accurate (Mullennix, Pisoni, & Martin, 1988), selective attention to phonetic detail is impaired (Mullennix & Pisoni, 1990), and accuracy of word recall is reduced (Martin, Mullennix, Pisoni, & Summers, 1989). All of these findings reveal the perceptual consequences of talker variability, but are ambiguous with respect to the underlying processes or mechanisms affected by changing voices. One possibility is that each spoken item must be "normalized" with respect to voice-specific information (e.g., Joos, 1948). Such an explanation would imply that voice-specific information is selectively ignored and is absent from higher lexical representation. Another possibility is that voice-specific information is selectively attended to and is integrated into memory representation. Whatever the explanation, both sorts of processes would be expected to require some portion of the listener's total processing capacity and both processes would be expected to yield the sort of data that have been reported. However, the two sorts of processes are qualitatively different, one implying a capacity-demanding *perceptual* operation, the other implying a capacity-demanding *storage* operation.

Martin et al. (1989) suggested that talker variability impairs serial recall of spoken word lists by imposing extra rehearsal demands, relative to the rehearsal demands imposed by single-talker lists. Although the Martin et al. suggestion may be accurate, it does not address the nature of the extra rehearsal processing; is the extra processing used for "normalization" or integration? An experiment was conducted by Goldinger, Pisoni, and Logan (1991) to examine the role of talker variability in serial recall of spoken word lists more closely. Goldinger et al. (1991) compared recall of lists of words spoken by either a single talker or by ten different talkers. At the same time, they compared recall of lists of "easy" words and lists of "hard" words, with the division along the *word confusability* dimension determined by a combined metric of word frequency and similarity neighborhood characteristics. Recall of these different kinds of spoken word lists was compared across several presentation rates of the spoken lists. Lists were presented with one word spoken every 250 msec, 500 msec, 1000 msec, 2000 msec, or 4000 msec.

The findings from the Goldinger et al. (1991) experiment can be easily summarized: At faster presentation rates, words from early positions of single-talker lists were recalled more accurately than words from early positions of multiple-talker lists, as Martin et al. (1989) reported earlier. Also, at faster presentation rates, "easy" words from early list positions were recalled more accurately than "hard" words from early list positions. As the presentation rate was slowed down, however, recall of early list items from multiple-talker lists was found to improve, and eventually surpass recall of the same items from the single-talker lists. In contrast, no interaction was observed between presentation rate and word confusability; "easy" words were recalled more accurately than "hard" words at all presentation rates.

On the basis of these findings, Goldinger et al. (1991) concluded that talker variability and word confusability affected recall in qualitatively different ways. We suggested that word confusability affects recall primarily by adding time to the perceptual encoding of the list items (thereby *indirectly* affecting rehearsal), and by confusions at retrieval. Conversely, we also suggested that talker variability slows perceptual encoding, but also interferes with rehearsal more directly. Specifically, we assumed that listeners attempt to either ignore or use voice information, depending on the time available for list processing. Accordingly, at fast rates, talker variability adds perceptual "noise" to the list, but at slow rates, talker variability actually enhances the distinctiveness of the items and helps the subject encode the temporal structure of the list.

The present experiment was intended to complement and to scrutinize the earlier experiment of Goldinger et al. (1991), by examining the rehearsal demands of the various sorts of spoken word lists more directly. The present experiment employed a new experimental procedure, called a *self-paced listening* procedure to examine rehearsal. In the earlier experiment summarized above, the presentation rate for any given list was determined a priori by the experimenter, and was held constant throughout the experimental session for an entire group of subjects. Then, as usual, the accuracy of serial-ordered recall was examined as a function of presentation rate. In the self-paced listening procedure used in the present experiment, subjects were allowed to control the presentation rates throughout the experiment for themselves. The basic procedure was simple: each subject listened to spoken word lists over headphones, one word at a time. The subject was provided with a button on a response box that controlled list presentation. Whenever the subject was ready for the next word in a list, he or she pressed the button and a new word was output by the computer. Once the list was complete, the subject recalled the words in serial order, as in the earlier experiment. Two dependent measures were examined-- the duration of the inter-word intervals during list encoding, which indicated the rehearsal demands of the various lists, and the recall data, which indicated the relation of rehearsal time to accuracy of recall.

The basic motivation for conducting an experiment using self-paced rehearsal and serial recall was to provide a more direct observation of the rehearsal requirements imposed by varying the stimulus dimensions of talker variability and word confusability. Several predicted outcomes can be derived from the arguments made by Goldinger et al. (1991). First, if talker variability directly affects rehearsal processes and word confusability does not, then the rehearsal times should differ for single- and multiple-talker lists, but not for "easy" and "hard" word lists. Second, providing listeners with all the time they need to rehearse should improve recall of multiple-talker lists relative to single-talker lists, but should not affect the relative recall of "easy" and "hard" word lists.

Method

Subjects

Forty-two students enrolled in introductory psychology courses at Indiana University served as subjects. Subjects received course credit for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimulus Materials

The stimulus materials were obtained from a large digitized database of spoken monosyllabic English words recorded by several different talkers. This database was the same source of stimulus materials used by Goldinger et al. (1991). The original words came from the vocabulary used in the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965). In the present experiment, only a subset of the original 300 words were used. The words selected for the present experiment satisfied several constraints: First, the words were ranked according to their frequency of occurrence, according to the Kucera and Francis (1967) norms. Second, the words were ranked according to their neighborhood densities, as determined by a one-phoneme substitution, addition, and deletion metric (Luce, 1986). Third, the words were ranked according to their neighborhood frequencies, a measure of the average frequency of the words' neighbors. Using these three criteria, two sets of words were selected for use in the present experiment. One set, the "easy" words, consisted of high frequency words from low density, low frequency neighborhoods. The other set of words, the "hard" words, consisted of low frequency words from high density, high frequency neighborhoods. A final criterion used in selection was subjective familiarity; all of the words chosen for use in the experiment were rated as highly familiar by subjects in an earlier experiment conducted by Nusbaum, Pisoni, and Davis (1984). After the words were divided into "easy" and "hard" sets according to these four criteria, each condition contained 50 items. These 100 words were then used to generate 10 lists of ten words each. Five of the lists contained all "easy" words and five contained all "hard" words. (For a list of all words, see Goldinger et al., 1991, pg. 162.)

Once the words had been selected, digitized files containing tokens of each word were obtained from the database. One set of tokens consisted of utterances that were all produced by a single male talker; these tokens were used for the single-talker conditions of the experiment. Another set of tokens was selected from the database so that each word in each list was spoken by a different talker; these tokens were used for the multiple-talker conditions. In the multiple-talker conditions, the same ten talkers, five males and five females, were used for all ten lists of words. All of the stimuli were originally recorded on audio tape and digitized with a 12-bit analog-to-digital converter using a PDP 11/34 computer. The mean RMS amplitude of all stimulus tokens was equated using a signal processing package. All stimulus tokens employed in the present experiment were tested for intelligibility and were found to be highly intelligible.

Procedure

Subjects were tested individually in a quiet testing room used for speech perception experiments. Stimuli were presented over matched and calibrated TDH-39 headphones at 75 dB (SPL). A PDP 11/34 computer presented the stimuli and controlled the experimental procedure in real-time. The digitized stimuli were reproduced using a 12-bit digital-to-analog converter and were low-passed filtered at 4.8 kHz.

All subjects were tested under the same conditions. Subjects first heard a 500 ms, 1000 Hz warning tone indicating that a list of words was about to be presented. Then, a list of ten words was presented at a self-determined rate of presentation. After each word was presented, the subject pressed either button on a two-button response box to indicate that he or she was ready for the following word. In this manner, subjects controlled the rate of item presentation throughout the experiment. The PDP 11/34 computer recorded the inter-word pause durations in milliseconds for later analysis. If a pause between any two words in a list reached ten seconds, the computer automatically recorded a ten-second pause duration and presented the next word of the list.[2] After the tenth word of each list, the subject pressed the button again to sound a second tone and begin the recall period. Once the recall period was initiated by the second tone, subjects had 60 seconds to recall all the words they could. Subjects were instructed to recall the words in the exact order of their presentation in the lists. Subjects wrote their responses on answer sheets using a pen or pencil. The end of the recall period was indicated by the presentation of a third tone.

Talker variability was a between-subjects variable; word confusability was a within-subjects variable. Half of the subjects were presented single-talker lists and half were presented multiple-talker lists. The same words were heard by all subjects; only the number of talkers and the self-determined presentation rates varied between subjects. The order of presentation of words within each list varied randomly across sessions. The lists themselves were presented in the same order for all subjects, with lists alternating between those lists containing "easy" words and those containing "hard" words.

[2] The ten-second deadline used in this procedure was imposed to limit the variability of the pause durations. In a pilot experiment, the self-pacing procedure was employed with no deadline imposed, so pause durations were determined entirely by the vigilance of the individual subjects. Across the 20 subjects in the pilot experiment, the pause duration data suggested that vigilance was a widely variable personality trait. The ten-second deadline was found sufficient to keep variation across subjects within a reasonable range.

All subjects received the same 12 lists of words. The first two lists were practice lists to allow subjects to learn the procedure; only data from the remaining ten lists were included in the final data analysis.

Results

The data analyses are discussed first for the inter-word pause duration data, and then for the recall data.

Pause Durations

Figure 1 displays the data for the inter-word pause durations as a function of serial position, talker condition, and word confusability. The upper panel shows pause durations as a function of serial position and talker condition, collapsed across word confusability. The lower panel shows pause durations as a function of serial position and word confusability, collapsed across talker condition.

Insert Figure 1 about here

A three-way analysis of variance (talker X word confusability X serial position) was conducted on the inter-word pause durations. In the ANOVA, talker condition was treated as a between-subjects variable; word confusability and serial position were treated as within-subjects variables. A significant main effect of talker was observed [$F(1,41)=21.31$, $p<.0001$]. Subjects took more time to listen to multiple-talker lists than single-talker lists. A significant main effect of word confusability was also observed [$F(1,41)=77.37$, $p<.0001$]. Subjects took more time to listen to lists of "hard" words than lists of "easy" words. In addition to the main effects of talker and word confusability, both panels of Figure 1 show a strong main effect of serial position [$F(9,369)=89.88$, $p<.0001$], reflecting the inverted U-shaped pause duration functions.

A significant two-way interaction of talker and serial position was also observed [$F(9,369)=2.21$, $p<.05$]. The differences in pause durations between the single-talker and multiple-talker lists were larger at earlier list positions. Post-hoc Tukey's HSD analyses indicated that the pause duration functions for single- and multiple-talker lists were significantly different at all intervals except 8-9, 9-10, and 10-tone. A similar two-way interaction of word confusability and serial position was also observed [$F(9,369)=11.21$, $p<.0001$]. The differences in pause durations between lists of "easy" and "hard" words were larger at earlier list positions. Post-hoc Tukey's HSD analyses indicated that the pause duration functions significantly differed at intervals 1-2, 2-3, 3-4, and 4-5. Most critically, the three-way interaction of talker, word confusability, and serial position was also significant [$F(9,369)=2.48$, $p<.01$]. The three-way interaction indicated that the differences between the pause duration functions for the single- and multiple-talker lists were larger than the differences between the pause duration functions for the "easy" and "hard" word lists.

Word Recall

A given response was scored as correct only if the presented word or some phonetically equivalent spelling of the presented word was recalled in the same serial position as the word presented in the list. Figure 2 displays the recall data as a function of serial position, talker condition, and word confusability. The upper panel shows recall as a function of serial position and talker condition, collapsed across word confusability. The lower panel shows recall as a function of serial position and word confusability, collapsed across talker condition.

Insert Figure 2 about here

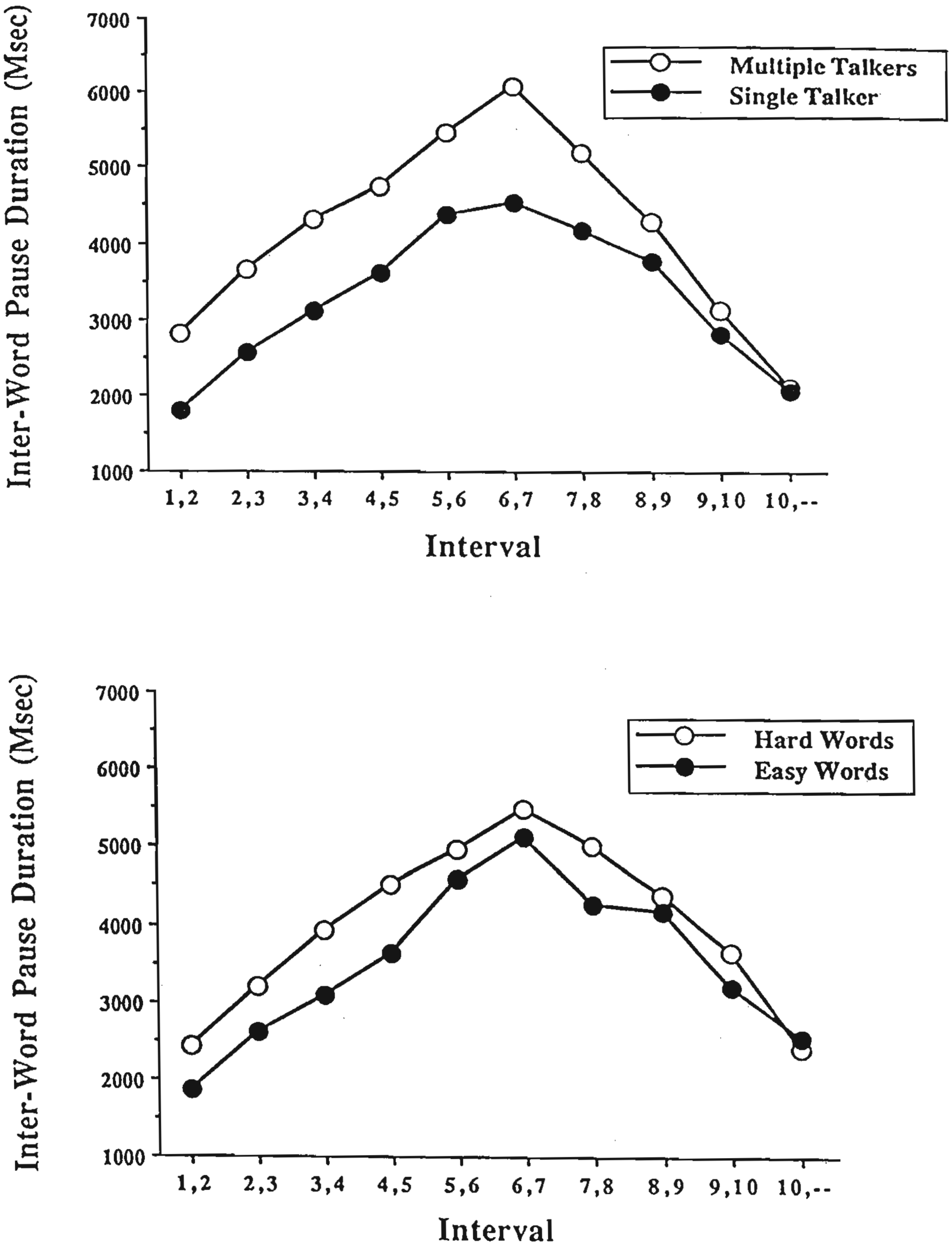


Figure 1. Inter-word pause durations shown as a function of serial position, talker condition, and word confusability. The upper panel shows pause durations as a function of serial position and talker condition, collapsed across word confusability. The lower panel shows pause durations as a function of serial position and word confusability, collapsed across talker condition.

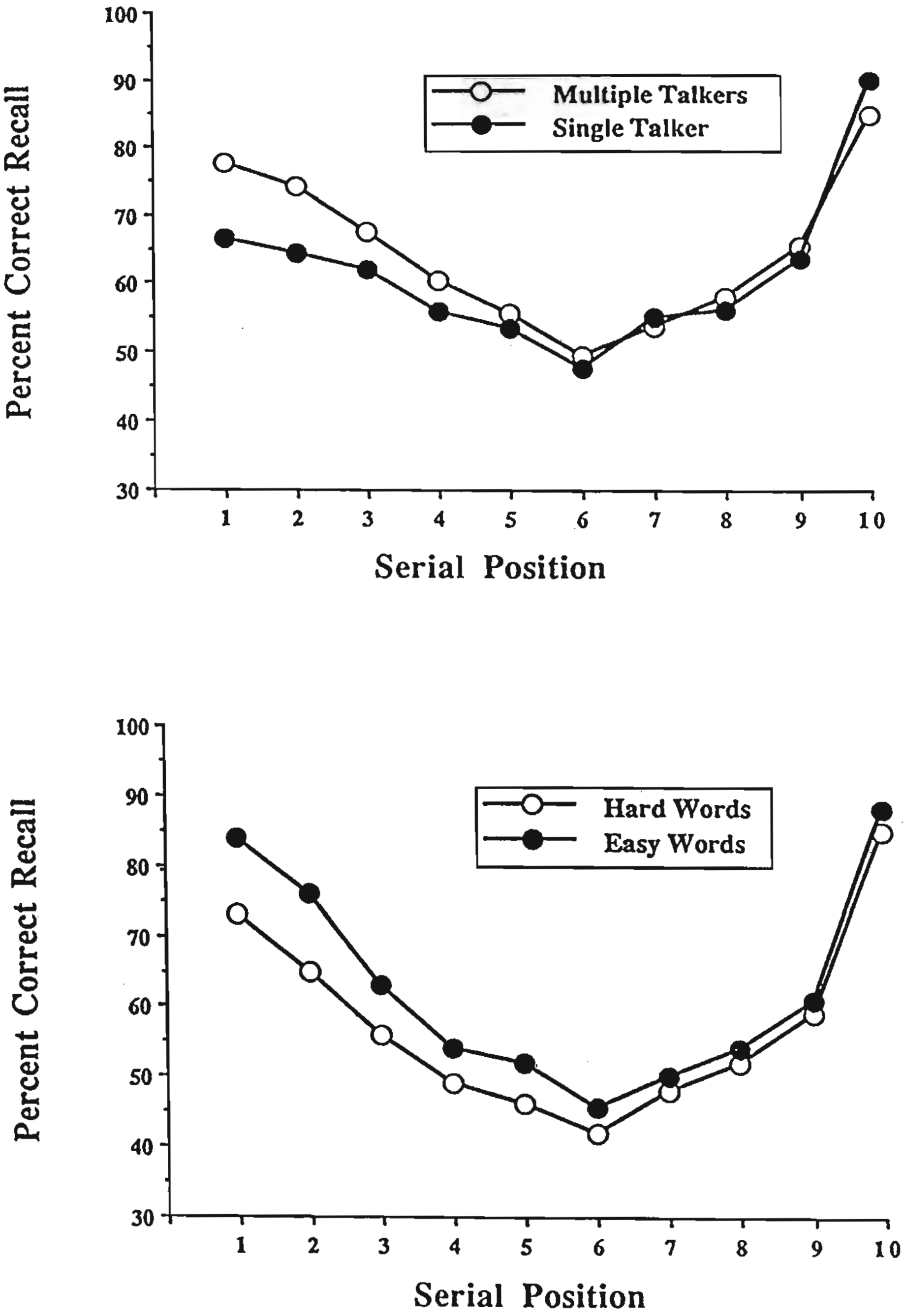


Figure 2. Recall data shown as a function of serial position, talker condition, and word confusability. The upper panel shows recall as a function of serial position and talker condition, collapsed across word confusability. The lower panel shows recall as a function of serial position and word confusability, collapsed across talker condition.

A three-way analysis of variance (talker X word confusability X serial position) was conducted on the percentages of correctly recalled words. A significant main effect of talker was observed [$F(1,41)=6.18, p<.05$]. Words from multiple-talker lists were recalled more accurately than words from single-talker lists. A significant main effect of word confusability was also observed [$F(1,41)=7.14$, $p<.05$]. Words from "easy" lists were recalled more accurately than words from "hard" lists. A strong main effect of serial position was observed as well [$F(9,369)=59.08$, $p<.0001$], reflecting the usual U-shaped function obtained in recall tasks.

A significant two-way interaction of talker and serial position was observed [$F(9,369)=2.01$, $p<.05$]. The differences in recall of words from the single-talker and multiple-talker lists were larger at earlier list positions. Post-hoc Tukey's HSD analyses indicated that the recall functions for single- and multiple-talker lists significantly differed at serial positions 1, 2, 3, and 4. A similar two-way interaction of word confusability and serial position was observed [$F(9,369)=14.27$, $p<.001$]. The differences in recall between lists of "easy" and "hard" words were larger at earlier list positions. Post-hoc Tukey's HSD analyses indicated that the recall functions significantly differed at serial positions 1, 2, 3, 4, and 5. The three-way interaction of talker, word confusability, and serial position was not significant [$F(9,369)=1.03$, $p>.10$].

A natural question that arises with respect to the present data involves the relation of recall obtained under conditions of self-pacing to recall obtained under a controlled presentation rate. Figure 3 displays the recall functions obtained in the present experiment using the self-pacing procedure and the recall functions obtained by Goldinger et al. (1991) using a 4-second presentation rate. The self-pacing data is shown in the left panels; the controlled rate data is shown in the right panels. As before, the upper panels compare single- and multiple-talker list recall; the lower panels compare "easy" and "hard" list recall.

Insert Figure 3 about here

As Figure 3 shows, with both self-pacing and the relatively slow 4-second presentation rate, words from early positions of multiple-talker lists were recalled more accurately than words from early list positions of single-talker lists. Similarly, in both experiments, words from early positions of "easy" lists were recalled better than words from early list positions of "hard" lists. A four-way analysis of variance (experiment X serial position X talker X word confusability) was conducted to compare the recall data across experiments. In this analysis, the difference in recall between single- and multiple-talker lists did not differ across experiments; the two-way interaction of experiment X talker was not significant [$F(1,73)=0.88$, $p>.10$]. The two-way interaction of experiment X word confusability was significant, however [$F(1,73)=6.64, p<.02$]. This interaction reflects the larger differences in recall between "easy" and "hard" words in the controlled-rate experiment than in the self-pacing experiment. Finally, a significant two-way interaction of experiment X serial position was obtained [$F(2,576)=15.43$, $p<.0001$]. This interaction reflects the respective slopes of the serial position functions across the experiments-- the serial position curves were more bowed in the controlled-rate experiment than in the self-paced experiment.

Discussion

The main findings of this experiment can be easily summarized: First, listeners took more time to process spoken word lists for later recall when the words were either spoken by multiple talkers or when the words themselves were low frequency words that were highly confusable with other words in the language. However, the amount of time taken to compensate for talker variability was greater than

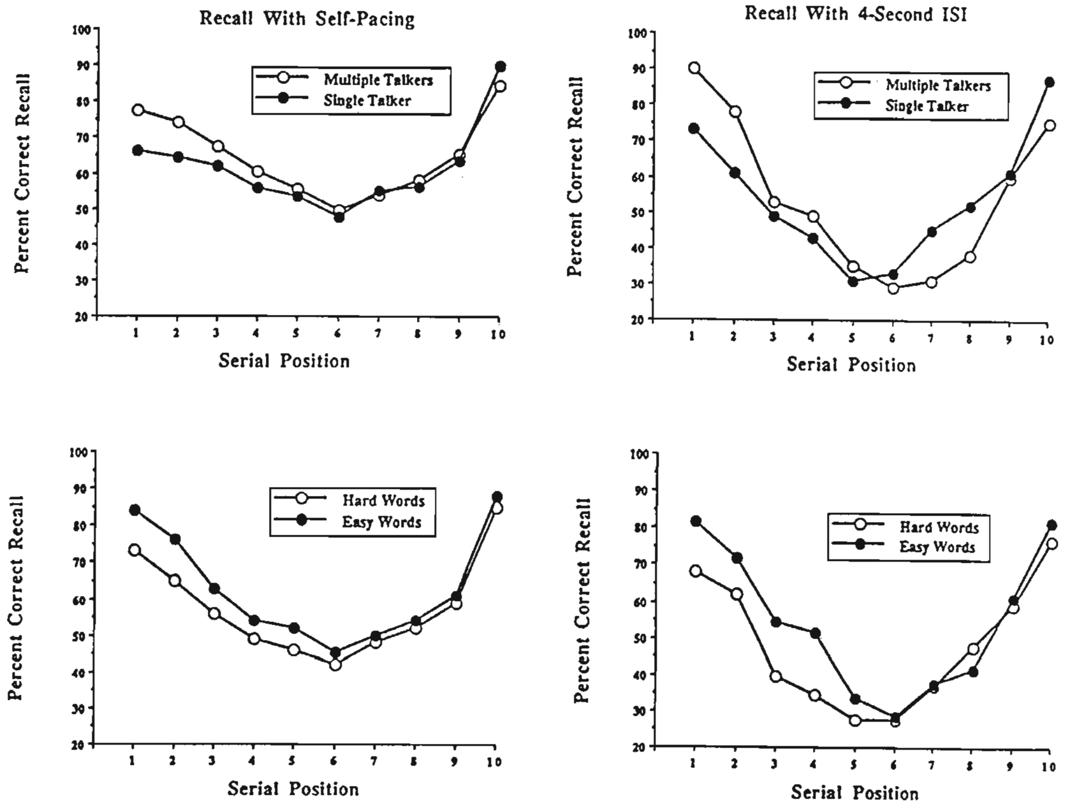


Figure 3. Recall functions obtained using the self-pacing procedure and recall functions obtained using a 4-second presentation rate. The left panels show the self-pacing data; the right panels show the controlled rate data. The upper panels compare single- and multiple-talker list recall; the lower panels compare "easy" and "hard" list recall.

the amount of time taken to compensate for word confusability.[3] Second, the extra time taken to compensate for talker variability resulted in superior recall of multiple-talker lists, relative to single-talker lists. The extra time taken to compensate for "hard" lists, however, did not result in superior recall of "hard" lists, relative to "easy" lists. Third, the time taken to rehearse words in a spoken list for later recall was partially determined by the serial position of the words in their lists.

The patterns observed in both the pause duration data and the serial recall data provide converging evidence in partial support of the claims made by Goldinger et al. (1991). In our earlier study, we argued that the differences found between recall of single- and multiple-talker lists were due to differing rehearsal demands associated with the two sorts of lists. Conversely, we suggested that the differences between recall of "easy" and "hard" lists were due to perceptual confusions either at encoding or retrieval. The pause duration data partially confirmed these claims-- subjects provided themselves with more time to compensate for talker variability than to compensate for "harder" words. This difference suggests that talker variability places greater rehearsal demands on the listener than word confusability does. However, it should also be noted that there were reliable differences in the inter-word pause durations between lists of "easy" and "hard" words. Moreover, the differences in the duration functions interacted with serial position. These data suggest that word confusability does exert some influence on rehearsal processes, contrary to the earlier strong claim of Goldinger et al. (1991).

In our earlier study, we also suggested that talker variability interacted with presentation rate because voice information cannot be ignored; the listener must dedicate processing capacity to either intentionally ignoring voice information or to integrating voice information into some representation of the spoken list as a whole. We assumed that voice information cannot be ignored because of the perceptual salience and communicative importance of voice-specific information. Moreover, experimental testing had demonstrated that listeners cannot selectively ignore variations in talker identity while selectively attending to the phonetic content of the speech signal (Mullennix & Pisoni, 1990). Conversely, we argued that word confusability had neither perceptual salience nor communicative importance. Accordingly, word confusability was not found to interact with presentation rate.

Given these earlier findings, an important aspect of the present data concerns the relation of rehearsal time to recall accuracy, which complements the findings of Goldinger et al. (1991) summarized above. At faster presentation rates, Goldinger et al. (and also Martin et al., 1989) found that single-talker lists were recalled more accurately than multiple-talker lists. When the presentation rates were slower, however, recall of multiple-talker lists improved and eventually surpassed recall of single-talker lists. In the present experiment, it was again found that words in the early portions of multiple-talker lists were recalled more accurately than words in the early portions of single-talker lists. Moreover, listeners provided themselves with more time to rehearse multiple-talker lists than single-talker lists. This isomorphism of rehearsal time to accuracy was *not* observed for the word confusability dimension. Although listeners provided themselves with more time to rehearse lists of "hard" words than lists of "easy" words, the extra time taken was ineffectual in changing their relative recall accuracy. "Easy" words from early list positions were still recalled more accurately than "hard" words from early list positions, despite their rehearsal time disadvantage.

[3] Given this finding, it is important to note that Goldinger et al. (1991, footnote 3) found that the perceptual deficits incurred by talker variability and word confusability for these stimuli were nearly equivalent. Therefore, the differences in pause durations should not be attributed to differing degrees of perceptual difficulty associated with the two dimensions of stimulus variation.

Because it was found that listeners take more time to process lists of "hard" words than lists of "easy" words, it is not reasonable to claim that there are no differences in rehearsal demands between such lists. However, the finding that extra rehearsal time has a strong effect on recall of single- vs. multiple-talker lists but has little effect on recall of "easy" vs. "hard" lists does suggest that the extra processing time is used differently in the two conditions. The most plausible account of these data, as well as the Goldinger et al. (1991) data, is that voice information is available to strategic, attentional processing and memory coding whereas the abstract, non-perceptual properties of words (e.g., frequency or neighborhood density) are not. The wider implication of this finding concerns the nature of voice information in lexical memory-- the powerful effects of voice information in this task suggest that voice information may not be "normalized" out of the speech signal or discarded after early stages of perceptual analysis. Further research is planned to examine the role of voice information in lexical memory over periods longer than the one-hour sessions used in the present experiment.

In summary, the present experiment used a self-paced listening procedure to examine the differences in recall between single- and multiple-talker lists of spoken words and lists of "easy" and "hard" words. The self-pacing procedure provided a unique opportunity to examine the rehearsal demands imposed by the different sorts of lists of spoken words. Rather than having the experimenter control the presentation rate and measure the effect on recall, the self-pacing procedure allowed the subject to control the presentation rate dynamically throughout the lists, yielding valuable information regarding rehearsal time as well as recall (see also Rundus, 1971). The rehearsal time and recall data obtained in the present experiment complement and extend the findings of Goldinger et al. (1991)-- the locus of the effect of talker variability on recall of spoken word lists appears to be in the rehearsal stage of list processing for eventual recall.

References

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

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.

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

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.

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 and Psychophysics*, **47**, 379-390.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1988). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

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.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, **89**, 63-77.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Some Lexical Effects in Phoneme Categorization: A First Report[1]

Scott E. Lively and David B. Pisoni

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

[1] This research has been supported by NIH Training Grant #DC-00012-12 and NIH Research Grant #DC-00111-14 to Indiana University.

Abstract

Two experiments using a phoneme categorization task examined the role of word frequency and lexical density in lexical access. In Experiment 1, subjects monitored spoken words for either two or four word-initial target phonemes. No word frequency or lexical density effects were observed in the reaction time data, indicating that subjects were responding on the basis of prelexical information. In Experiment 2, we used Frauenfelder and Segui's (1989) generalized phoneme monitoring task in which the position of the target phoneme varied randomly from trial to trial. The results showed that high frequency words were responded to more rapidly than low frequency words and that high density words were responded to more rapidly than low density words. The results of Experiment 2 suggest that subjects responded on the basis of information derived from a lexical representation that may not have been fully specified. The results are discussed in terms of an interactive-activation model that is similar to the Race Model proposed by Cutler, Mehler, Norris, and Segui (1987).

Some Lexical Effects in Phoneme Categorization: A First Report

The phoneme monitoring task has been used for many years as a measure of on-line word recognition and comprehension. The simple speeded-classification task requires subjects to respond whenever a prespecified target phoneme occurs in a spoken utterance. Since Foss first used the task, the phoneme monitoring task has provided a mixed set of results (Foss, 1970; Foss & Swinney, 1973). Segui and Frauenfelder (1986) point to three contrasting patterns of results in the phoneme monitoring literature. One set of results indicates that subjects respond strictly in the basis of lexically-derived information. Reliable lexicality effects are observed in this case; responses to words are faster than to nonwords (Rubin, Turvey, & van Gelder, 1978). A second set of results suggests that subjects respond strictly on the basis of prelexical information (Segui, Frauenfelder, & Mehler, 1981). By this account, no effects of word frequency or lexical status are predicted. The final pattern of results relates to a processing assumption generated from a combination of the first two observations. Prelexical and postlexical information may compete for a single response outlet. Thus, lexical effects may be observed under certain conditions (Eimas, Hornstein, & Payton, 1990). The present set of experiments was conducted to investigate some of the conditions under which the phoneme monitoring task produces word frequency and lexical density effects for isolated spoken words.

Foss and colleagues (Foss, 1970; Foss & Swinney, 1973) introduced the phoneme monitoring task as a means of investigating on-line processes involved in spoken language comprehension. In the phoneme monitoring task, subjects are given a target phoneme or set of target phonemes to monitor for, prior to hearing an utterance. Subjects respond whenever one of the target items is present. Under certain conditions, the task can be sensitive to lexical variables. Foss and Swinney (1973), Morton and Long (1976), and Dell and Newman (1980), for example, all demonstrated that subjects responded more rapidly to phonemes in target words that occurred in highly predictable sentential contexts. In general, response time increased as predictability decreased.

Foss and Blank (1980) hypothesized that words in predictable contexts were activated more rapidly than words in neutral or unpredictable contexts. They suggested that rapid activation of lexical information was responsible for the observed predictability effects. This set of results, indicating that prior sentential context facilitates a phoneme monitoring response, is consistent with at least the first observation offered by Segui and Frauenfelder (1986). However, the results may also be consistent with the third observation, which suggested that prelexical and lexically-derived information compete for a single response outlet. Based on the available results, however, it is not possible to determine when phonetic information became activated relative to lexical information. One alternative is that information about the target phoneme became available only after lexical access occurred. This would indicate that the data is consistent only with the first of Segui and Frauenfelder's observations. A second alternative is that prelexical and lexically-derived information were both available and that subjects selected one of the two alternatives.

The studies by Foss and Swinney (1973), Morton and Long (1976), and Dell and Newman (1980) demonstrated lexical effects using a phoneme monitoring task in sentence contexts. Several studies have also used the phoneme monitoring task to study the access and recognition of isolated words (Rubin, Turvey, & Van Gelder, 1978; Marslen-Wilson, 1984). Rubin et al., for example, required subjects to monitor for phonemes occurring at the beginning of monosyllabic words and nonwords. A significant lexicality effect was observed in their study; words were responded to significantly more rapidly than nonwords. Rubin et al. provided two interpretations for their data. Their first explanation suggested that larger units of perceptual analysis became available prior to smaller units. By this explanation, lexical access takes place prior to a phoneme detection response. This explanation is consistent with Segui and

Frauenfelder's (1986) first hypothesis. Rubin et al. point out, however, that this is an unrealistic hypothesis given that the stimuli were equated for length and that words were still responded to more rapidly than nonwords. The stronger interpretation of their data is based on speed of access rather than on the length of the analysis unit. Rubin et al. suggested that words become available prior to nonwords because of their increased familiarity to the listener and their semantic representation. This explanation predicts that lexical characteristics of words may influence phoneme monitoring reaction times. This pattern is consistent with the first and third explanations outlined by Segui and Frauenfelder (1986); listeners may respond strictly on the basis of a lexical code (observation 1) or prelexical and postlexical information may compete for a single response outlet (observation 3). If the two codes race for a single outlet, then Rubin et al.'s data indicate that the postlexical code may have become available prior to the prelexical code or that the postlexical code was given preference over the prelexical code.

Marslen-Wilson (1984) replicated Rubin et al.'s result by showing that subjects may respond on the basis of information derived from lexical access when responding to isolated stimuli. Subjects were asked to monitor for phoneme targets in words and nonwords. The position of the target phonemes varied from trial to trial. Reaction times were measured from the onset of the target phoneme. Marslen-Wilson found that reaction times for phoneme targets in words varied as a function of the recognition point of the stimulus.[2] In the nonword condition, no significant differences in response times were observed as a function of the position of the target phoneme; reaction times were constant across target positions in the nonwords. This pattern of results suggests that monitoring responses may have been facilitated by the use of lexical information. When no lexical information was available, as in the case of nonwords, subjects responded on the basis of acoustic-phonetic information. Predictive information from the lexicon could not be used to facilitate a response. Marslen-Wilson's results, like those of Rubin et al. (1976), are consistent with the first and third accounts suggested by Segui and Frauenfelder (1986); subjects respond strictly on the basis of lexical information, unless there is no lexical information to be found. However, because no compelling evidence has been offered to exclude the availability of prelexical information, Segui and Frauenfelder's third suggestion may be preferred to the first. Recall that their third suggestion is that prelexical and lexically-derived sources of information compete for a single response outlet.

Much of the previously described research has been consistent with the first and third hypotheses offered by Segui and Frauenfelder (1986). There is, however, also a corpus of data consistent with their second hypothesis which suggested that subjects respond on the basis of prelexical information. For example, Foss and Blank (1980) demonstrated that subjects responded on the basis of prelexical information when monitoring sentences for phoneme targets that occurred at the beginning of words and nonwords. No significant difference in reaction times was observed for responses to words and nonwords.[3] Foss and Blank claimed that this pattern of results was evidence that listeners compute a prelexical representation from the acoustic-phonetic input that can be used to rapidly output a response. Because of the speed of its computation, this low-level, acoustic-phonetic information becomes available prior to lexical information.

[2] The recognition point of a word is the minimal amount of auditory information a subject needs in order to reliably identify the word. Marslen-Wilson (1987) claims that recognition points occur very early in most words and that this helps to account for the speed of the word recognition process.

[3] It should be noted, however, that subjects also demonstrated postlexical responses in the same experiment. When targets occurred in words that followed a word, reaction times were faster than when the target phoneme followed a nonword. The result can be explained by a failure to properly segment the end of the preceding nonword.

Foss and Gernsbacher (1983) extended Foss and Blank's (1980) results. They failed to find a main effect for lexical status using a phoneme monitoring task in sentence contexts. Average reaction times to words were not significantly different than reaction times to nonwords. In their second and third studies, Foss and Gernsbacher manipulated processing load in order to induce subjects to use a postlexical code. Processing load was manipulated by varying the phonetic similarity of the initial phoneme of the word preceding the target-bearing word to the target phoneme. For example, subjects might monitor for word-initial /b/'s. In a high similarity or high processing load condition, the phonetically similar "dear" might immediately precede the target-bearing word "boy", in a sentence context. The phoneme /d/ was assumed to be highly similar to /b/ because they differ only in their place of articulation. No clear pattern of lexicality effects was observed. When subjects responded to /b/ targets, responses to words were faster than responses to nonwords. In contrast, responses to /d/ targets were faster when the target occurred in nonwords. These effects became stronger when the target-bearing word was preceded by a word with a similar initial phoneme. Foss and Gernsbacher concluded that subjects did not consistently adopt a postlexical response code.

In a final set of experiments, Foss and Gernsbacher (1983) demonstrated that reaction times were positively correlated with the vowel duration of the target bearing item. This correlation was used to explain the faster reaction times to nonwords containing /d/ targets in their previous experiment. Based on the data collected in this series of experiments, Foss and Gernsbacher suggested that subjects compute a low-level acoustic-phonetic code and a high-level, lexically-derived code. When the lexically accessed code becomes available, a phonological representation for the target phoneme is specified and a monitoring decision can be made. However, this process is relatively slow in comparison to the low-level acoustic-phonetic computations. Thus, subjects may respond on the basis of prelexically computed information. The pattern of results is therefore consistent with Segui and Frauenfelder's second and third observations.

Segui, Frauenfelder, and Mehler (1982) also found evidence for the use of a prelexical code. Their experiment was similar in many respects to Rubin et al.'s (1976) study. Subjects were required to monitor for phonemes occurring at the beginnings of isolated words and nonwords. In addition to monitoring for word initial phonemes, Segui et al. also had their subjects monitor for the initial syllable of bisyllabic words. Thus, the monitoring unit (i.e., phoneme or syllable) and the length of the stimuli (i.e., bisyllabic stimuli) varied in comparison to Rubin et al.'s experiment. The results also differed between the two studies. Whereas Rubin et al. found differences in reaction times to words and nonwords, Segui et al. found no differences. The discrepancy between the two studies was attributed to differences in the size of the processing units. Segui et al. suggested that the recognition of a syllabic unit preceded the recognition of a phonemic unit. In their view recognition of phonemes occurred only after the syllables had been recognized. Rubin et al.'s use of monosyllabic words may have masked this effect. This is similar to the first explanation offered by Rubin et al. The failure to find lexical effects in Segui et al.'s experiment indicates that monitoring responses may be executed on the basis of information available prior to lexical access, consistent with the second observation made by Segui and Frauenfelder (1986).

Given the conflicting results reviewed above, several general observations can be made about the phoneme monitoring task. First, Segui and Frauenfelder's (1986) observations that the phoneme monitoring decisions are made strictly on the basis of prelexical, acoustic-phonetic information or strictly on the basis of information derived through lexical access lacks strong empirical support. Their third observation appears to provide the best account of the available data; prelexical and postlexical information may both become active and compete for a single response outlet. Competition between the two alternatives may be biased toward one alternative by the specific experimental conditions. If both response codes do become active and compete for a single response outlet during the phoneme monitoring task, then a model is needed to explain how both codes become activated, what the time course of their activation is, and how the two codes compete for the single response outlet.

At least two models have been proposed to explain the parallel activation of prelexical and lexically-derived codes. Cutler and Norris (1979) and Cutler, Mehler, Norris, and Segui (1987) have offered their Race Model as one account for the conflicting pattern of data. They assume that subjects compute two codes in parallel. The first code is based on prelexical, acoustic-phonetic input whereas the second code is based on information derived from lexical access. Once the lexical code has been activated, a phonological form of the word becomes available for use in making a phoneme categorization. The two codes are computed in parallel and are in competition with each other. When the prelexical code is specified prior to lexical access, subjects respond on the basis of acoustic-phonetic information. In this case lexical variables, such as lexical status or word frequency, should not affect reaction time. If lexical access occurs prior to a detection response, then subjects should respond on the basis of the phonological code derived from the lexicon in an obligatory manner. In this case, lexical variables should have observable, facilitatory influences on monitoring response times.

Cutler et al.'s model predicts many of the prelexically and lexically based effects observed in the previous studies. For example, the Race Model predicts that, in the absence of a biasing context, the prelexical code becomes available first under most conditions and provides the basis for the categorization response. This is a reasonable prediction when the temporal nature of spoken language is considered. In the absence of any context, prelexical, acoustic-phonetic information is available to the listener for processing prior to any candidate set of words for recognition. Thus, the prelexical code has a "head start" in the race and the lexically computed code may be unable to recover. However, the lexical code can also be given preference over the prelexical under certain circumstances. For example, a preceding sentence context or a semantically related prime may preactivate a set of consistent word candidates prior to the realization of the target-bearing word (Foss & Blank, 1980; McClelland & Rumelhart, 1981). Thus, lexical processing may begin prior to the computation of a prelexical code. Cutler et al. also predict that the postlexical code can be made available prior to a prelexical code in the absence of a preactivating context. For example, they suggest that if the target-bearing stimuli are short words, such as monosyllabic words, the postlexical code may be rapidly activated, able to overcome its initial processing disadvantage, and therefore provide the basis for the categorization response. This prediction is consistent with Rubin et al.'s and Segui et al.'s results.

The Dual Code Model of Foss and Blank (1980) provides a similar account for the conflicting pattern of results. The Dual Code Model differs from the Race Model in several of its assumptions. Cutler et al. (1987) claim that the computation of a prelexical code is a task-induced strategy. They suggest that the syllable, not the phoneme, may be the early unit of representation (Segui et al., 1982). Foss and Blank, in contrast, suggest that the prelexical, acoustic-phonetic code is a naturally computed, quickly decaying representation. The two models also differ with respect to assumptions made about competition between the response codes. While prelexical and lexical codes compete with each other in the Race Model, the Dual Code model assumes that the two codes are computed independently, in parallel, and in a noncompetitive manner. The first available code is not necessarily used as the basis for a response. This assumption implies that subjects may use a lexically-derived code in making their responses, even though the prelexical code may already be fully specified. The difference in assumptions concerning the competition between response codes can be used to distinguish the two models. For example, the two models make different predictions about the facilitatory nature of lexically-derived information. In the case of the Race Model, all postlexical responses must be facilitative.[4] In contrast, the Dual Code predicts either facilitation or inhibition. Inhibition is predicted when the prelexical code becomes available first but defers its decision to the postlexical code (Eimas et al., 1990), whereas facilitation is predicted when the first available code is used to provide a response alternative.

Recently, Eimas et al. (1990) have extended the Race Model by adding an attentional component. An attentional component was proposed in order to account for how subjects change processing strategies in response to varying task demands. In a series of 16 experiments, Eimas et al. confirmed some of the predictions of the Race Model by demonstrating that the prelexical code was typically available to subjects prior to the postlexical code. This was particularly true when only a single phoneme categorization response was required. In the cases where prelexical information was available first, no response time advantage was found for words over nonwords. However, when a lexically-oriented, secondary task was added to the primary, phoneme monitoring decision, a clear lexical advantage was observed. For example, when subjects were asked to make a lexical decision response following a phoneme detection response, phoneme monitoring responses were significantly faster to targets in words. Eimas et al. argued that the change in experimental conditions from a single task to a dual task caused a shift in the subjects' attentional allocation which was responsible for the lexical effects. Because the prelexical code is assumed to be available prior to the lexical code, responses in the single phoneme monitoring task can be facilitated by focusing attention at a prelexical level. However if subjects attempt to maintain a prelexical attentional focus when they are performing the dual monitoring and decision tasks, they must shift their attention between prelexical and postlexical codes on each trial in order to accurately perform the two tasks. A shifting attentional strategy within trials might lead to increased response times. Thus, Eimas et al. suggested that subjects focus their attention on lexically-based information in response to the competing demands of the phoneme monitoring task and the lexical decision task.

Frauenfelder and Segui (1989) also found evidence to support an attentional focus account using a generalized phoneme monitoring procedure (GPM). In the GPM, subjects were presented prime-target pairs and asked to monitor for a prespecified phoneme within the target words. When the prime-target pairs were blocked so that target phonemes always occurred word initially, no priming effects were observed. However, when the position of the target phoneme varied randomly from trial to trial, large facilitatory semantic priming effects were observed. Based on these results, Frauenfelder and Segui concluded that subjects reallocated attention to lexically-derived information when uncertainty is introduced by varying the phoneme target position. When the position of the target phoneme occurred late in the word, or when its position was uncertain from trial to trial, subjects shifted attention toward lexically computed output. Marslen-Wilson (1984) also found lexically-based effects by varying target position on a trial-by-trial basis.

Given the previous mixed set of results and Frauenfelder and Segui's success with the GPM, there were two principle sources of motivation for the current experiments. First, we were interested in examining the conditions under which the phoneme monitoring task yields lexically-based effects. Once conditions were found that yielded reliable lexical results, we were interested in examining the effects of word frequency and neighborhood density on phoneme monitoring reaction times. Several previous phoneme monitoring studies have demonstrated word frequency effects (Eimas et al., 1990; Segui & Frauenfelder, 1986; Rubin et al., 1976). As in the previous studies, high frequency words are predicted to be responded to faster than low frequency words. High frequency words require the summation of less acoustic-phonetic information for activation and may send more activation per unit of time to a

[4] Recall that in the Race Model the prelexical code was typically made available prior to the postlexical code.

decision unit. Frequency effects have been observed numerous times across numerous different tasks, such as lexical decision, naming, and perceptual identification (Balota & Chumbley, 1985; Luce, 1986). Similar explanations have been given for frequency effects in different tasks.

The more interesting results, in our view, deal with the effects of lexical density. Little attention has been paid to the role of lexical density in the phoneme monitoring task. For purposes of this report, lexical density is defined as the size of a neighborhood for a particular target word (Coltheart, Davelaar, Jonasson, & Besner, 1977; Luce, 1986). A phonological neighbor is operationally defined as any word which differs by a one phoneme substitution from a particular word. For example, "cat" and "bat" are phonological neighbors. A neighborhood is defined as the collection of words that vary from a particular word by one phoneme. Thus, "cat" resides in a "dense" neighborhood because many words can be derived from it by a one phoneme substitution rule. In contrast a word such as "chrysanthemum" is a "sparse" word because it has no phonological neighbors.

Lexical density can have three possible effects on phoneme monitoring latencies. First, density may have no effect at all on reaction times. In this case, the size of the lexical response set activated prior to a phoneme monitoring response is orthogonal to the response the subject makes. This prediction implies that inhibitory competition among neighbors observed in perceptual identification (e.g., Luce, 1986; Marslen-Wilson, 1987) will not affect phoneme classification reaction times.

A second prediction is that lexical density will affect phoneme monitoring responses in the same way it affects perceptual identification responses. In a perceptual identification task subjects are required to uniquely identify a word from a set of possible response alternatives derived from the perceptual processing of the acoustic-phonetic input. Accuracy in this task is inversely related to the size of the generated confusion set (Luce, 1986; Marslen-Wilson, 1987). Competition among neighbors in a dense neighborhood is assumed to inhibit activation strengths and make words from dense neighborhoods more difficult to identify. Many highly similar words compete for recognition and must be disambiguated prior to executing a response. This competition and disambiguation process predicts slow responses to words from dense neighborhoods in the phoneme monitoring task.

The third possible prediction is that words from dense neighborhoods will be responded to more rapidly than words from sparse neighborhoods. The phoneme monitoring task may be dissimilar to the perceptual identification task because unique identification of the stimulus word may not be required to initiate a classification response. Phoneme monitoring requires only a gross categorization response. A response may be facilitated by the presence of partially activated, phonetically similar, words competing for recognition. The interactive activation model of McClelland and Rumelhart predicts this "gang effect" (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). Similar items tend to share similar features. Thus, while items may inhibit one another directly, they may also indirectly activate each other by activating lower-level and higher-level features that are common to both items. In the case of phoneme monitoring, words from the same lexical neighborhood may inhibit their neighbors directly, but the same neighbors also receive excitatory activation from the same low-level acoustic-phonetic source. This leads to a large amount of activation within a particular neighborhood and therefore may facilitate a phoneme monitoring response based on lexical information. Thus, words from dense neighborhoods may be predicted to be responded to more rapidly than words from sparse neighborhoods.

We investigated the effects of frequency and density in two experiments. In Experiment 1, we tested the claim that single syllable words would be responded to on a postlexical basis (Cutler et al., 1987; Segui & Frauenfelder, 1986). Subjects monitored for two or four response alternatives in a between-subjects design. We predicted that high frequency words would be responded to more rapidly than low frequency words and that high density words would be responded to more rapidly than low density words, as outlined above. In Experiment 2, we investigated the role of word frequency and neighborhood density using a variation of Frauenfelder and Segui's generalized phoneme monitoring task. In one condition, subjects were required to monitor for phonemes in lists which were blocked by target position. In a second condition subjects monitored for phonemes in a list of isolated words in which the target position varied randomly from trial to trial. In order to respond prelexically in this condition, a subject would have to know before each trial where the target was to occur and would have to shift their attention to that position. Frequency and density effects were expected in this condition. We predicted that responses would be faster for high frequency and high density words compared to low frequency and low density words.

EXPERIMENT 1: Two vs. Four Target Phonemes

Method

Subjects

Subjects in Experiment 1 were 17 undergraduate students enrolled in an introductory psychology course at Indiana University. All subjects were native speakers of English and reported no history of speech or hearing problems at the time of testing. Subjects were given partial course credit for their participation in this experiment.

Stimuli

One hundred-twenty monosyllabic words were chosen from a computerized version of Webster's Pocket Dictionary. The initial phoneme of each stimulus word was either /b/, /d/, /p/ or /t/. The words were selected on the basis of their frequency and lexical density. High frequency words had a mean frequency of 135.58 occurrences per million words while low frequency words had a mean frequency of 3.67 (Kucera & Francis, 1967). Lexical density was computed using the one phoneme substitution rule (Luce, 1986). High density words had a mean of 25.48 phonological neighbors; medium density words had a mean of 18.45 neighbors; low density words had a mean 12.60 neighbors. Ten words were assigned to each of the six conditions that resulted from crossing both variables.

Words were produced by a native speaker of English and were recorded and low pass filtered at 4.8 kHz on an Ampex reel-to-reel tape deck at 15 ips using a DO45 Electro Voice microphone. The audio tape was then digitally sampled at a rate of 10 kHz using a 12 bit A/D converter interfaced to a PDP 11/34 laboratory computer. The resulting digital file was segmented into individual files by WAVES, a digitally controlled waveform editor (Luce & Carrell, 1981). All stimulus files were equated for RMS amplitude and were tested for intelligibility by a separate group of native speakers of English. Any stimulus item that was not correctly identified by all listeners was rerecorded.

Procedure

Subjects in Experiment 1 participated in groups of five or fewer. Each subject sat in a small, sound-attenuated cubicle that was designed for experiments in speech perception. Subjects wore TDH-39 matched and calibrated headphones and were seated in front a small video display monitor (GBC Standard CRT Model MV-10A). All stimuli were presented at a comfortable listening level between 75 and 80 dB SPL. Each subject categorized the stimuli according to their initial phonemes via a two- or four-button response box that was interfaced to the computer.

The number of response alternatives was manipulated as a between-subjects variable. Nine randomly assigned subjects responded to words beginning with one of four target phonemes: /b/, /p/, /d/, or /t/. The remaining subjects responded to stimuli beginning only with /b/ or /t/. One-hundred-twenty trials were presented in the four response alternative condition, while 60 trials were presented in the two response alternative condition. On each trial, a "Get Ready" prompt appeared on the video monitor for 1000 ms prior to stimulus presentation. Subjects responded by pressing the appropriately labeled response button. A maximum of 4 s was allowed for each response. Subjects were encouraged to respond as quickly and as accurately as possible. No feedback was given. Prior to collecting experimental data subjects were given between 6 and 12 practice trials to familiarize them with the task. Although reaction times and accuracy data were collected during all phases of the experiment, only the data from the experimental session was analyzed.

Results

Analyses of variance were conducted separately on the accuracy and the response time data. Word frequency and neighborhood density were treated as within-subjects variables. Data from the two-target and four-target conditions were analyzed separately due to the large difference in the number of trials between the two conditions. For neither condition in the accuracy data were significant differences observed between high and low frequency words or between high and low density words. The interaction also failed to reach significance. Error rates were below 10% in both the two alternative and the four alternative conditions.

Because the stimuli were presented at a high signal-to-noise ratio and the task was relatively easy, accuracy measures may not be particularly informative. Instead, differences in reaction times may reveal information about the roles of frequency and density in lexical access. Therefore, an analysis of variance was calculated for the reaction time data. As in the previous analysis, no main effects of frequency or density were observed in either condition. The frequency by density interaction also failed to reach significance. Table 1 shows mean reaction times to high and low frequency words in the two and four response alternative conditions. Table 2 shows mean reaction times as a function of density. No significant main effects or interactions were observed.

Insert Tables 1 and 2 about here

Discussion

No significant effects of word frequency or lexical density were observed in the present experiment. The null results are predicted by both the Race Model and the Dual Code model and are consistent with Segui and Frauenfelder's second observation: phoneme monitoring can be performed on the basis of prelexical information. In this experiment, all target phonemes occurred word initially. Thus, subjects needed to attend only to the initial segment of each word. Once this portion of the word was identified, subjects could respond without processing the remainder of the stimulus pattern. These experimental conditions apparently did not force subjects to engage in lexically-based processing to execute their response. Task monotony (Cutler et al., 1987), caused by the invariant position of the target phoneme, may have also encouraged subjects to attend to a prelexical code. According to the Race Model, the prelexical code began with a computational head start over the lexical code and maintained that lead through the completion of the trial. Thus, frequency and density information did not influence monitoring responses.

The null results of this experiment are similar to the findings reported by Segui et al. (1981), who observed no effects of lexical status when subjects were required to monitor bisyllabic words and

Table 1

Experiment 1: Mean Reaction Times

Condition	**High Frequency**	**Low Frequency**
(1) /b/-/d/	605.92	616.14
(2) /b/-/d/ /p/-/t/	803.54	800.71

Table 2

Experiment 1: Mean Reaction Times

Condition	**Density**		
	High	**Medium**	**Low**
(1) /b/-/d/	600.99	616.24	615.87
(2) /b/-/d/ /p/-/t/	790.20	816.45	799.72

nonwords for initial phonemes and syllables. Responses to Segui et al.'s stimuli were also made on the basis of a prelexical code. Whether the prelexical code was made available to subjects prior to a lexically-derived code is a debatable point between the Dual Code and the Race Models.

While the results of this first experiment are similar to those of Segui et al. (1981), they are at odds with the findings reported by Rubin et al. (1976). Recall that Rubin et al. found that subjects were faster to monitor for phoneme targets in monosyllabic words than in monosyllabic nonwords. No significant frequency or density effects were observed in Experiment 1. The difference in the results between the two experiments may be accounted for by a difference in stimulus materials. The stimuli in our experiment were monosyllabic English words. In contrast, Rubin et al. used a combination of monosyllabic words and nonwords. One possible explanation for the differences in the results is that subjects in Rubin et al.'s experiment may have made an implicit lexical decision prior to a phoneme monitoring decision. This implicit decision may have increased response times when subjects were asked to monitor for a target phoneme in nonwords because they failed to retrieve the item from the lexicon. The two types of stimuli used by Rubin et al. may have encouraged subjects to respond on the basis of lexical information. Since all of the stimuli in our experiment were words, no lexical decision had to be made prior to a categorization response.

Given the dominance of the prelexical code in Experiment 1, we were interested in finding a set of conditions which might lead subjects to respond on the basis of lexical information. In order to induce lexical effects, Eimas et al. (1990) for example, had subjects perform a dual task which required a lexical decision or noun-verb categorization response after the initial phoneme monitoring response. The authors reasoned that the lexically-oriented demands of the secondary task might induce subjects to respond with postlexical information. In contrast, Frauenfelder and Segui (1989) induced lexical effects by varying the position of the target phoneme from trial to trial. When subjects could not predict the position of the target phoneme on a trial by trial basis, they tended to respond on the basis of lexically-derived information. In the next experiment, we used a version of Frauenfelder and Segui's generalized phoneme monitoring technique to study the effects of word frequency and lexical density on lexical access. As in Experiment 1, we expected subjects to respond faster to high frequency and high density words.

EXPERIMENT 2: Generalized Phoneme Monitoring

Method

Subjects

Subjects in this experiment were 50 undergraduate students who were enrolled in an introductory psychology class at Indiana University. All subjects were native speakers of English and claimed to be audiologically normal at the time of testing. Subjects were given partial course credit for their participation in this experiment.

Materials

Eighty words were chosen from a computerized version of Webster's Pocket Dictionary. All words were monosyllabic and contained a /b/ or /d/ in either initial position or final position. No words contained both phonemes. The phonemes /b/ and /d/ occurred equally often in initial and final positions. Word frequency and lexical density were factorially combined with target position. High frequency words had a mean frequency of 467.625 occurrences per million words, while low frequency words had a mean frequency of 1.78. High density words had a mean of 22.175 neighbors and low density words had a mean of 3.95 neighbors. Stimuli were prepared in the same manner as described in Experiment 1.

Procedure

The procedure in Experiment 2 was almost identical to that of Experiment 1. The major difference in the two experiments was in the blocking of the conditions by target position. In Experiment 2, subjects participated in one of three conditions. Thirty-two subjects responded to stimuli that were blocked by phoneme position. Word initial targets were presented before word final targets to 16 of the subjects. The list of 40 initial position target-bearing words was presented twice in different random orders before the list of final-position target bearing words was presented. The list of words with final position targets was also presented in two random orders. Counter balancing was achieved by presenting final position targets to a group of 18 listeners in the first two blocks. This was followed by two blocks of final position trials. A final group of 16 subjects participated in a version of Frauenfelder and Segui's generalized phoneme monitoring task (1989). In this condition, target position varied randomly between initial position and final position on a trial-by-trial basis.

Subjects participated in two blocks of 80 trials each. Accuracy and latency measures were collected in all conditions on each trial. Response latencies were measured from the onset of the stimulus in all conditions.

Results

An analysis of variance was carried out on the latency data. Condition (mixed vs. blocked) served as a between-subject variable, while target position, block number, word frequency, lexical density, and target phoneme served as within-subject variables. Tukey tests were computed to analyze interactions.

The overall analysis of variance revealed a main effect of condition [$F(2,47)=8.70$, $p<.01$]. Subjects in the final-initial blocked condition had the fastest reaction times. Latencies were longest for subjects in the mixed condition. Mean reaction times for the three conditions are displayed in Table 3.

Insert Table 3 about here

In addition to a main effect of condition, a main effect of target position was observed in the overall analysis [$F(1,47)=380.91$, $p<.01$]. Reaction times to word initial targets were faster (920.97 ms) than reaction times to word final targets (1185.16 ms). The interaction between condition and target position was also significant [$F(2,47)=16.32$, $p<.01$]. Figure 1 shows that the difference in response times between target positions was smaller in the mixed condition than it was in either of the blocked conditions. Responses to initial position phonemes were significantly faster in the blocked condition in which final targets were responded to first.

Insert Figure 1 about here

Main effects of word frequency and lexical density were observed in the overall analysis. Subjects responded more rapidly to high frequency words than to low frequency words [$F(1,47)=153.62$, $p<.01$]. The mean response latency to high frequency words was 1008.44 ms, while the response time to low frequency words was 1097.70 ms. Subjects also responded more rapidly to high density words than to low density words [$F(1,47)=107.85$, $p<.01$]. High density words were responded to with a mean latency of 1023.53 ms. Mean reaction time to low density was 1082.60 ms. A significant interaction also occurred between the two variables [$F(1,47)=20.10$, $p<.01$]. Figure 2 displays the

Table 3

Experiment 2: Overall Analysis

Condition	**Mean RT**
Mixed	1144.08
Blocked (Initial-Final)	1059.31
Blocked (Final-Initial)	966.62

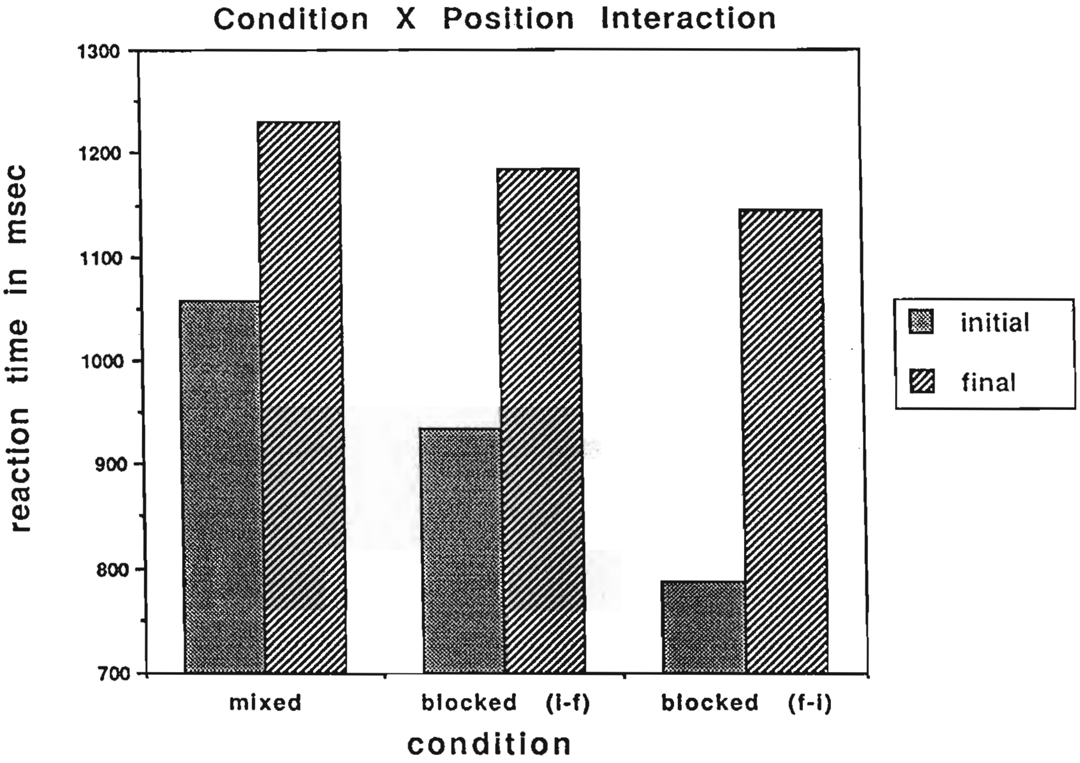


Figure 1. The interaction observed in Experiment 2 between phoneme target position and condition.

interaction of frequency and density, collapsed across target position, session, and condition. The difference in reaction times for high and low frequency words was smaller for high density words than for low density words. However, differences in response latencies between high and low frequency words were significant for both levels of density.

Insert Figure 2 about here

In addition to interactions with each other, frequency and density also independently entered into interactions with target position. The frequency by target position interaction revealed a small difference in reaction times for targets occurring in initial position but a significantly larger difference for targets occurring in word final position [$F(1,47)=65.97$, $p<.01$]. High frequency words were responded to more rapidly than low frequency word in both target positions. The density by target position interaction took a similar form [$F(1,47)=78.56$, $p<.01$]. The difference in reaction times between high and low density words was smaller for word initial targets but significantly larger for word final targets. Tables 4 and 5 display the mean reaction time data as a function of target position and frequency, and target position and density, respectively. In addition to the two two-way interactions, the three-way interaction of target position by frequency by density was significant in the overall analysis [$F(1,47)=4.90$, $p<.05$]. Figure 3 displays this three-way interaction. The figure shows that differences in reaction time for the frequency and density combinations were attenuated for targets in initial position but were significantly increased for targets occurring in final position.

Insert Tables 4 and 5 about here

Insert Figure 3 about here

Word frequency and lexical density behaved differently in their interactions with condition. The interaction between condition and frequency was not significant. In contrast, density interacted significantly with condition [$F(2,47)=5.23$, $p<.01$]. Density effects were significantly larger in the mixed condition than in the blocked conditions. In all cases, high density words were responded to faster than low density words. Figure 4 presents mean reaction time as a function of lexical density and condition.

Insert Figure 4 about here

Mixed Condition Results

The results of primary interest in the present experiment were those obtained in the mixed condition. Subjects in this condition participated in a version of Frauenfelder and Segui's generalized phoneme monitoring task which has previously revealed reliable, lexically-based results. The results described below show the same general trends observed in the overall analysis. All results reported in this section are based on an analysis of variance with subjects as the random factor, an analysis of variance with items as the random factor and a *min F'* analysis on the reaction time data. *Min F'* was computed in accordance with suggestions made by Clark (1973). Only results significant by the *min F'*

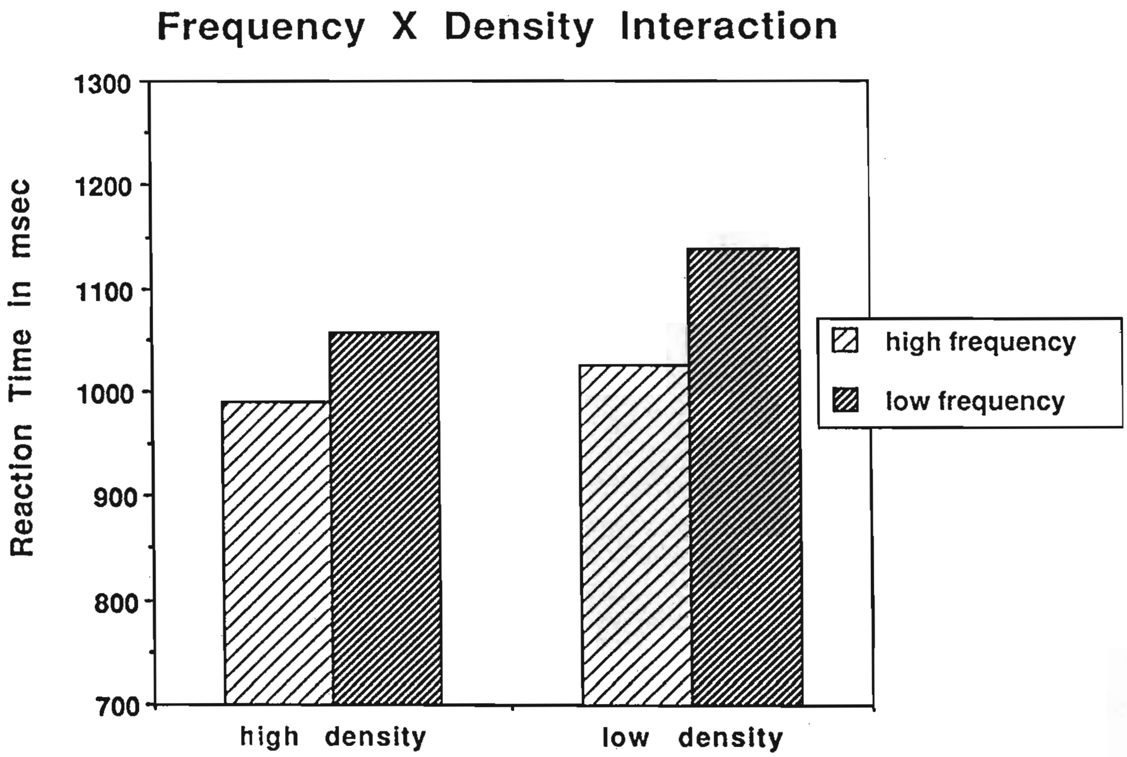


Figure 2. The frequency by density interaction observed in Experiment 2.

Table 4

Experiment 2: Frequency x Position Interaction

	Initial Position	**Final Position**
High Frequency	904.50	1112.37
Low Frequency	937.44	1257.96

Table 5

Experiment 2: Density x Position Interaction

	Initial Position	**Final Position**
High Density	915.62	1131.45
Low Density	926.32	1238.88

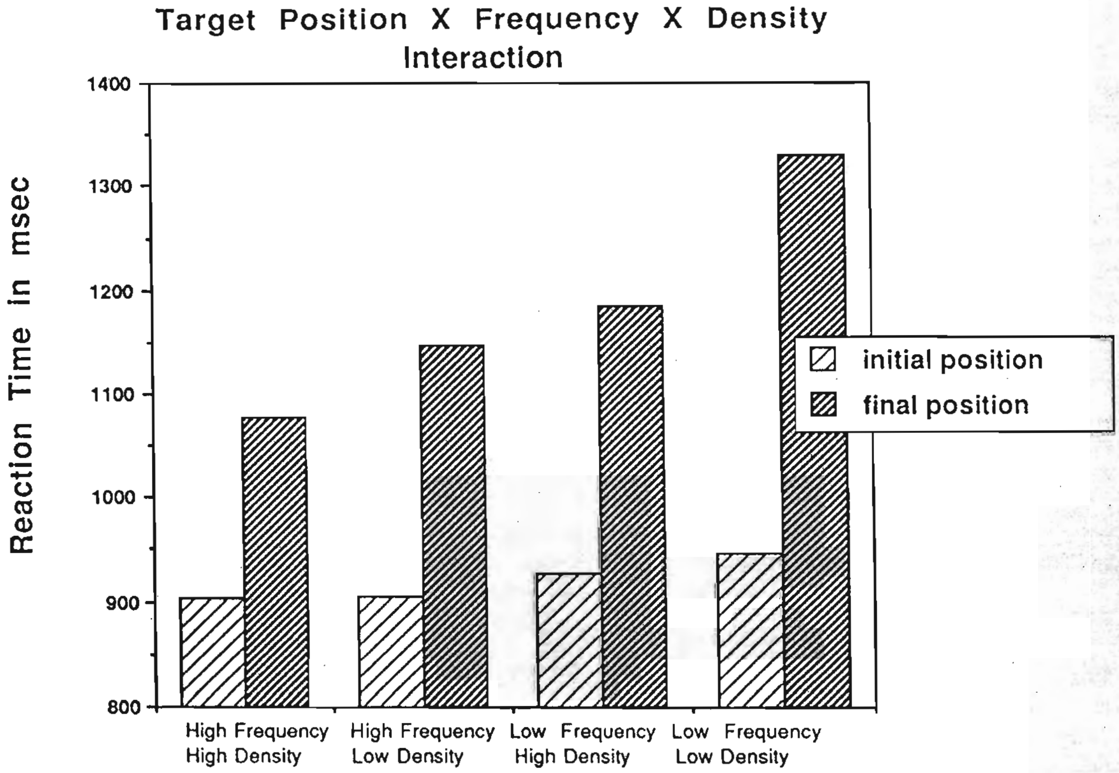


Figure 3. The three-way interaction of frequency, density, and condition in Experiment 2.

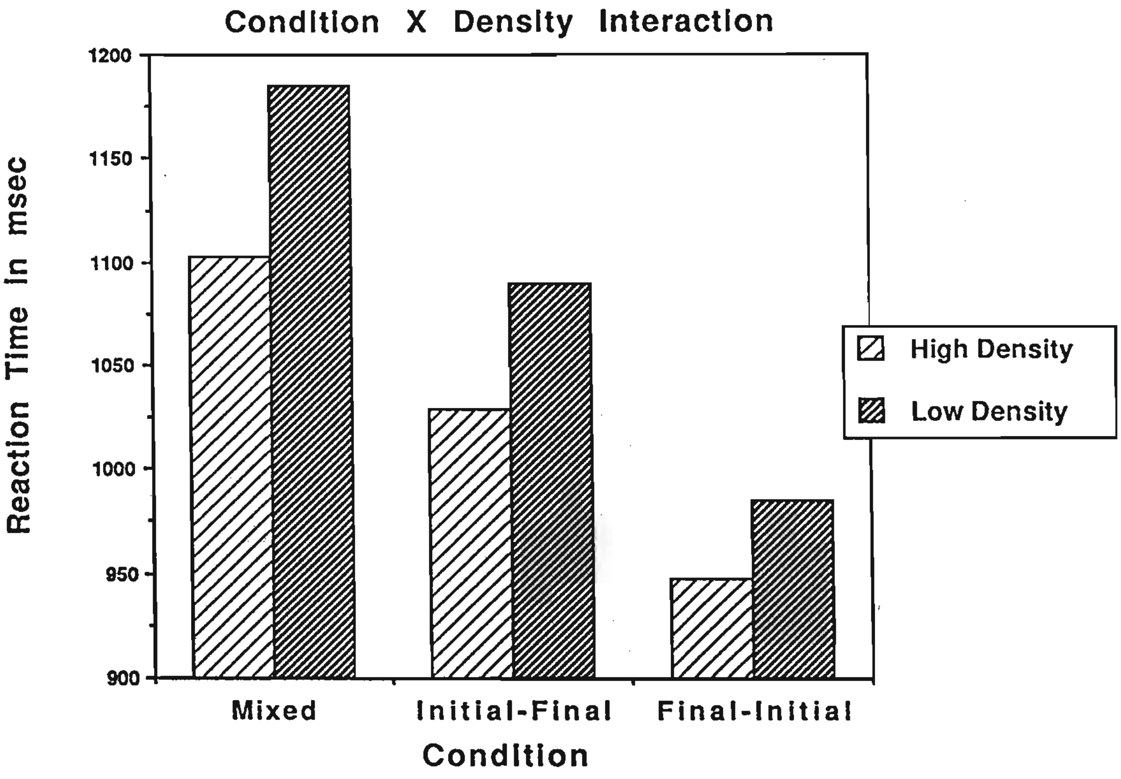


Figure 4. The interaction between density and condition observed in Experiment 2.

analysis will be reported here.[5] Tukey tests were used as a post hoc test to analyze significant interactions. All results reported in this section are significant at the .05 level and beyond.

As in the overall analysis, the main effect of target position was significant [*min F'*(1,88)=89.35, $p<.01$]. Targets in initial position were responded to more rapidly than targets in final position. Mean reaction time to initial position targets was 1057.27 ms, while mean reaction time to final position targets was 1230.88 ms. A main effect of word frequency was also observed [*min F'*(1,86)=24.93, $p<.01$]. High frequency words were responded to more rapidly than low frequency words (1097.06 ms vs. 1191.09 ms, respectively). The main effect of density in the mixed condition was similar to the finding observed on the overall analysis [*min F'*(1,98)=14.59, $p<.01$]. High density words were responded to more rapidly (1102.84 ms) than low density words (1185.31 ms).

Target position by frequency interactions and target position by lexical density interactions were also significant in the mixed condition analysis [*min* $F'_{\text{position x frequency}}$(1,98)=5.45, p<.05; *min* $F'_{\text{position x density}}$(1,98)=9.09, $p<.01$]. The interactions followed the same form as those described in the overall analysis. Differences in reaction times to high and low frequency words and high and low density words were significantly attenuated in initial position relative to final position. Table 6 displays mean reaction times as a function of frequency and target position while Table 7 shows mean latencies as a function of density and target position. Frequency and density also entered into a significant interaction with each other [*min F'*(1, 98)=4.50, $p<.01$]. Table 8 shows that the differences in reaction time were larger between high and low frequency words for low density words than for high density words. This interaction was similar to the one observed in the overall analysis.

Insert Tables 6, 7 and 8 about here

A more detailed analysis of the mixed condition data was performed to determine whether subjects engaged in lexical processing across both target positions. A *min F'* analysis was carried out separately for each half of the data. A main effect of word frequency was observed for word initial targets [*min F'*(1,99)=4.35, $p<.05$]. High frequency words were responded to more rapidly than low frequency words (984.58 ms vs. 1020.81 ms). However, the density effect was not significant for word initial targets. Both frequency and density effects were observed in the word final analysis [*min* $F'_{\text{frequency}}$ (1,81)=21.97, $p<.01$; *min* F'_{density}(1,99)=17.03, $p<.01$). High frequency words were responded to more rapidly than low frequency words (1101.63 ms vs. 1214.90 ms, respectively) while high density items were responded to more rapidly than low density items (1105.25 ms vs. 1211.29 ms, respectively). This analysis indicated that subjects responded with information derived from lexical access in both target positions in the mixed condition.

Discussion

Several general trends were observed in the data collected in the second experiment. First, phoneme targets occurring word initially were responded to more rapidly than targets occurring word finally. One simple interpretation of the position effect is that information about the identity of word initial consonants becomes available to subjects prior to information about the identity of word final consonants and this informational asynchrony facilitates a response. This result, taken by itself, does not implicate lexical processing, although it is consistent with Foss and Blank's assumption (1980) that a

[5] A density by phoneme identity interaction was significant by subjects, but not by items or min F'. No further attention will be given to this result.

Table 6

Experiment 2 (Mixed Condition Only): Frequency x Position Interaction

	Initial Position	**Final Position**
High Frequency	1028.95	1165.18
Low Frequency	1085.60	1296.59

Table 7

Experiment 2 (Mixed Condition Only): Density x Position Interaction

	Initial Position	**Final Position**
High Density	1045.13	1160.56
Low Density	1069.41	1301.21

Table 8

Experiment 2 (Mixed Condition Only): Frequency x Density Interaction

	High Density	**Low Density**
High Frequency	1069.61	1124.51
Low Frequency	1136.07	1251.10

phonetic code is naturally computed. A second trend, observed in the generalized phoneme monitoring task, does implicate lexical processing. The main effects of frequency and density are consistent with a lexical processing strategy. High frequency words were responded to more rapidly than low frequency words, while high density words were responded to more rapidly than low density words. These results are consistent with Cutler et al.'s (1987) Race Model. Prelexical and postlexical information compete for the same outlet. The code that is computed first provides the response. If the prelexical code had provided information sufficient for a response, then word frequency and lexical density should not have affected response times, as was observed in the first experiment.

Given that lexical effects were observed, it is not surprising that high frequency words were responded to more rapidly than low frequency words. Several previous phoneme monitoring studies have found similar results (Rubin et al., 1976; Frauenfelder & Segui, 1989; Eimas et al., 1990). The word frequency effect is a pervasive effect that is observed throughout a number of psycholinguistic tasks. For example Luce (1986) found an advantage for high frequency words in naming, auditory lexical decision and perceptual identification tasks. Balota and Chumbley (1985) have found frequency effects in naming and lexical decision tasks using visually presented stimuli. One simple explanation for frequency effects is based on resting activation levels or response threshold levels. High frequency words are assumed to have higher resting activations or lower response thresholds (Morton, 1969; Marslen-Wilson, 1985; McClelland & Rumelhart, 1981). Thus, less acoustic-phonetic information is needed to recognize high frequency words. This leads to faster reaction times for high frequency words across a variety of psycholinguistic tasks, including the phoneme monitoring task.

The explanation of the lexical density effect in this task is less obvious. Recall that high density words were responded to more rapidly than low density words. Typically, a different pattern has been observed; low density words are responded to more rapidly and more accurately than high density words. For example Luce (1986) found an advantage for low density words in naming, perceptual identification and auditory lexical decision tasks.[6] One way to account for the advantage typically observed for low density words is to consider the results within an activation, competition-based framework. Acoustic-phonetic information activates a set of possible recognition candidates in the lexicon (Luce, 1986; Marslen-Wilson, 1987). Words within a lexical neighborhood compete with each other for recognition. High density words receive more inhibition from their neighbors than low density words because high density words, by definition, have more neighbors. This inhibition slows recognition processes and causes responses to become slower and less accurate (Luce, 1986).

An analysis of the demands of the phoneme monitoring task can account for the seemingly anomalous lexical density effects. In the phoneme monitoring task subjects are required to classify words into a small set of categories based on the presence of a prespecified set of phonemes. Subjects are not required to make a unique identification of the stimulus word since they are in auditory naming or perceptual identification tasks. In the generalized phoneme monitoring task subjects may activate a set of candidate words that send activation to an output unit with a response threshold. When this unit has collected a criterial amount of information from prelexical and lexically-derived sources, a response can be made. Because a large number of words are initially activated, dense neighborhoods may send a sufficient amount of activation to the threshold response unit early in the lexical access process. Due to their lack of neighbors, sparse words are unable to initially send a large amount of activation to the output

[6] Andrews (1989) found a contrasting pattern of results in a visual lexical decision task and a visual naming task. High density words, determined by a one letter substitution rule, were responded to more rapidly than low density items.

unit. Enough information from the initial neighborhood activation may be specified for a classification response, so that detailed analysis of the stimulus pattern may not be necessary to carry out the task. Thus, the output threshold may be exceeded prior to the point at which competition among neighbors begins to severely inhibit or eliminate possible candidates.[7]

The main effects of target position, word frequency, lexical density and their interactions can be modeled within an interactive-activation framework (Elman & McClelland, 1984). The proposed model is composed of three types of units. The most basic level of the network accepts information that is translated into an acoustic-phonetic representation. Lexical units lie at the middle level of representational complexity and are assumed to be arranged in inhibitory phonological neighborhoods. Output threshold units lie at the highest level of the network. These units correspond to the categorization alternatives and are assumed to accumulate activation from both acoustic-phonetic and lexical sources. When enough activation has accumulated to exceed a categorization unit's response threshold, sufficient information is available for a phoneme monitoring response. No assumptions about the form of the threshold function are made here. A schematic diagram of the proposed network is shown in Figure 5.

Insert Figure 5 about here

In keeping with the assumptions of an interactive-activation model, connections within the network are assumed to be bidirectional (McClelland & Rumelhart, 1982). For simplicity bidirectional connections are assumed to be symmetric. Thus, the connection weight from lexical unit *ui* to output unit *uj* is assumed to be the same as the weight from *uj* to *ui*. The bidirectionality of the weights allows information propagated up from lower levels of representation to influence higher levels of representation, and vice versa. Also, keeping with the assumptions of a competition-based network, connections between levels are assumed to be facilitatory while connections within phonological neighborhoods are assumed to be inhibitory (Luce, 1986). Thus, lexical units can send activation to output units, propagate activation back to a developing acoustic-phonetic representation, and inhibit other units in their phonological neighborhood. Similarly, units at other levels of representation are assumed to inhibit other units with that level.

Input units are assumed to be connected both directly and indirectly to output categorization units. Indirect connections are mediated by lexical-level units. The weights on the direct and indirect connections do not have to be the same. An attentional component is added to the output units to attenuate the strength of activation sent directly from the inputs and from the lexicon. The attentional component is sensitive to the sources of activation coming to the output unit and proportionally attenuates and increases connection strengths. This mechanism is assumed to operate across trials rather than within

[7] A different pattern of results might be anticipated if a delayed phoneme monitoring response was required. When the classification response is delayed, recognition processes would proceed and inhibition would increase among phonological neighbors. This leads to the prediction that sparse words would be classified as fast as or faster than dense words, while high frequency words would still retain a response time advantage over low frequency words. Some support for this prediction can be marshalled by examining Eimas et al.'s results. When phoneme monitoring responses were delayed up to 900 ms, word frequency effects were still observed. Additional support can be mustered from Balota and Chumbley (1985), who observed a word frequency effect in a delayed naming task with delays over 1000 ms. However, the most direct support for this prediction comes from a study by Andrews (1989), who found an advantage for high frequency words but no effect of density in a delayed naming paradigm using visual stimuli.

Network Implementation of Cutler et al.'s (1987) Race Model

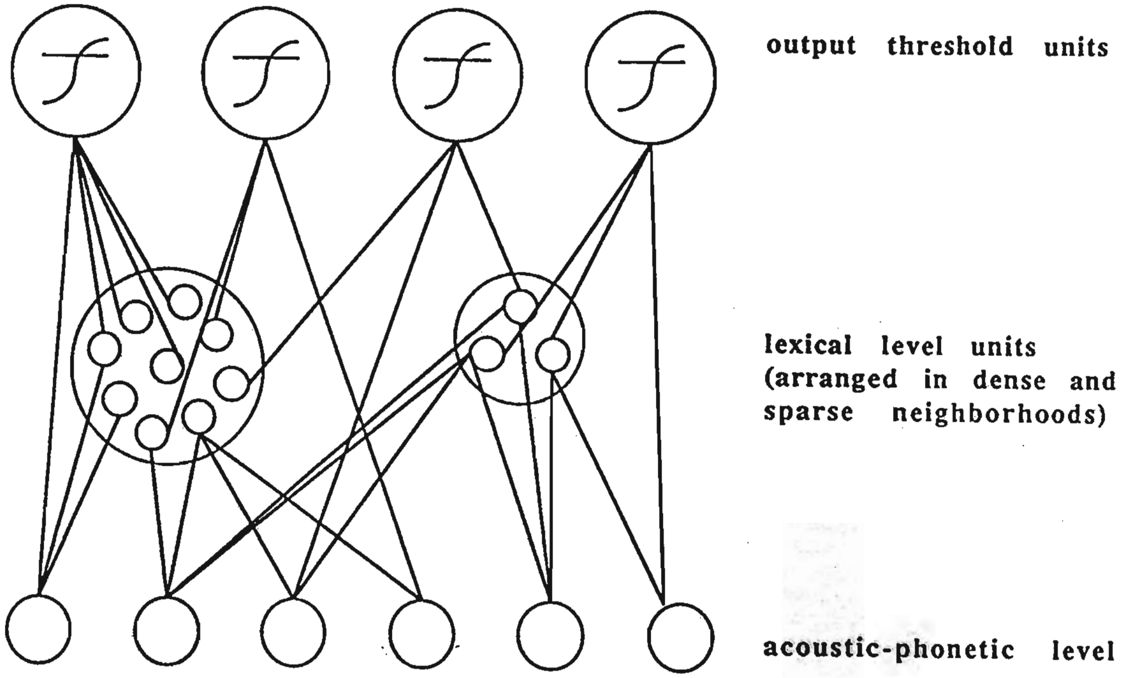

output threshold units
lexical level units (arranged in dense and sparse neighborhoods)
acoustic-phonetic level

Figure 5. A schematic diagram of the proposed interactive-activation instantiation of Cutler et al.'s (1987) Race Model.

trials.[8] Eimas et al. (1990) make a similar suggestion for Cutler et al.'s Race Model. An alternative function of the attention parameter would be to shift the output unit's response threshold while leaving connection weights constant. Thus, more or less information would be needed to activate a response.

Asymmetric direct and indirect (i.e., lexically mediated) connections could account for prelexical and postlexical effects observed here. Strong direct connections between the input nodes and the output categorization units would be responsible for prelexical effects. To account for postlexical effects, attention weights are shifted across trials toward lexically-derived information. A corresponding attenuation of direct connections would occur. Alternatively, prelexical and lexically-derived effects could be obtained by allowing the model to shift the threshold of the output units. Lower thresholds would lead to more prelexical responses. However, this might have the undesirable effect of increasing the system's error rate.

In order to derive predictions about the effects of word frequency and lexical density, additional assumptions need to be made about the values of the connection weights between the lexicon and the output levels. Weights from high frequency words to the output decision units are assumed to be larger than weights from low frequency words. The greater weights permit a larger proportion of the lexical unit's activation to be transmitted to the output units at each time step. This would have the effect of requiring fewer time steps to activate high frequency words. Lexical density effects could be accounted for by assuming that more nonzero connections exist between the members of dense neighborhoods and output units than is the case for sparse neighborhoods. This network structure allows more information to be transferred from units in dense neighborhoods to the output level on each time cycle. Fewer time cycles would be required to surpass an output threshold when responding to a dense word. This is an example of a "gang effect" (McClelland & Rumelhart, 1981). Nodes that share the same features tend to support each other and activation tends to accumulate rapidly among nodes that respond to similar information. By definition dense words have many words that share common phonemes, all of which tend to be activated by the same acoustic-phonetic input. The feedback connections from the output nodes to the lexical units serve to enhance the lexical effects because lexical units receive information both from the low-level acoustic-phonetic units and from the high level output units. Feedback from higher levels would help to rapidly activate lexical units. The higher activation of lexical units would, in turn, lead to higher activation of the output units. Thus, output units would be rapidly activated by dense neighborhoods.

Target position effects can be accounted for by the direct connections between the low-level acoustic-phonetic units and the output units. Information from early portions of the word that are consistent with the output categories directly excites the output decision nodes. In addition to activation propagated up from low-level acoustic-phonetic nodes, information is also sent to the output nodes via the lexical units. This tandem influx of activation rapidly increases the level of excitation in the output nodes and allows for fast responses for phonemes in initial position. The attenuated frequency and density effects in initial position can be accounted for with this architecture. Because a large amount of activation is sent very early in the categorization process from consistent input nodes, only a small

[8] If the attentional component changed the weights on the connections between the input and output units and the lexical and output units within a trial, it is possible that frequency effect observed for targets in initial position of the mixed condition might be eliminated. When the target is in initial position, the attentional component would severely attenuate the weights from the lexical units to the output units. There would be a corresponding increase in the weights on the direct connections between the input and output units. Because of the attenuation to the lexically-based weights, output decisions would consider only prelexical information and frequency effects would not be observed.

amount of activation is built up in the lexical units. Thus, the lexicon makes only a small contribution to the total activation of the output units in this case. Large lexical effects for phonemes in final position can be accounted for with a similar line of reasoning.

While the model proposed here may be able to account for a number of the observed effects in the present experiments, it should be noted that this model, in its simplest form, is just an interactive network instantiation of Cutler et al.'s Race Model (1987). The model is also consistent with the attentional constraint Eimas et al. (1991) added to the Race Model.[9] Currently, we are working on developing a computer simulation of the model to test its predictions. One short-coming of the model is immediately obvious; the model has no initial learning component, so the asymmetric connection weights are imposed on the system "programmer ex machina." This is clearly an unnatural way of deriving frequency and density effects. An improved version of the model would learn to correct mappings from input units to output units by adjusting weights and biases at the lexical and output levels. A learning algorithm applied to the network would allow it to develop its own representation of word frequency information, for example. An additional problem with the present model is that by relaxing the requirement that all connection weights be equal, the number of free parameters in the model expands with the addition of each new input, lexical or output unit. Thus, the model may be able to make any prediction based on fine tuning of the connection weights. A learning algorithm might alleviate this problem to some extent.

In addition to developing a working simulation of the model proposed above, we are extending the current research to examine effects of word length on lexical processing. Cutler et al. (1987) suggest that short words with word initial targets have a higher probability of being responded to via a postlexical code. Segui and Frauenfelder (1986) make a similar claim. Given the failure to find lexical effects with monosyllabic words in Experiment 1, we are interested in testing these claims more thoroughly.

In summary, we have presented preliminary evidence that the generalized phoneme monitoring task is sensitive not only to word frequency, but also to lexical density. High frequency words were responded to faster than low frequency words. High density words were responded to more rapidly than low density words. This was an example of a "gang effect." We suggested that subjects were responding on the basis of information derived from lexical access, but that words were not uniquely identified at the time of the categorization decision. A network instantiation of Cutler et al.'s Race Model was proposed to account for the observed effects. The model took the form of an interactive-activation architecture with an added attentional component. The attentional component made the network sensitive to the relative contributions of prelexical and lexically-derived sources of information that are used in spoken word recognition.

[9] Stemberger, Elman, and Haden (1985) have proposed a similar interactive activation type of model to account for interference effects observed when a target phoneme is preceded by a highly similar phoneme. They found that a simplified version of TRACE (Elman & McClelland, 1984; McClelland & Elman, 1986) was adequate to account for the pattern of interference results they observed.

References

Andrews, S. (1989). Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**(5), 802-814.

Balota, D., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335-359.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI.* Hillsdale, NJ: Erlbaum, 535-555.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, **19**, 141-177.

Cutler, A., & Norris, D. (1979). Monitoring sentence comprehension. In W.E. Cooper & E.C.T. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett.* Hillsdale, NJ: Erlbaum.

Dell, G.S., & Newman, J.E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, **19**, 607-623.

Eimas, P.D., Marcovitz-Hornstein, S.B., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language*, **29**, 160-180.

Elman, J.L., & McClelland, J.L. (1984). Speech perception as a cognitive process: The interactive activation model. In N. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice, Vol. 10.* New York: Academic Press, 337-374.

Foss, D.J. (1970). Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, **9**, 699-706.

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

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

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

Frauenfelder, U.H., & Segui, J. (1989). Phoneme monitoring and lexical processing: Evidence for associative context effects. *Memory and Cognition*, **17**(2), 134-140.

Kucera, H., & Francis, W.N. (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., & Carrell, T.D. (1981). Creating and editing waveforms using WAVES. *Research on Speech Perception Progress Report No. 7.* Bloomington, IN: Speech Research Laboratory, Indiana University.

Marslen-Wilson, W.D. (1984). Function and processes in spoken and word recognition. In H. Bouma, & D.G. Bouwhuis (Eds.), *Attention and Performance: Control of Language Processes.* Hillsdale, NJ: Erlbaum, 125-150.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. In U.H. Frauenfelder & L. Komisarjevsky Tyler (Eds.), *Spoken Word Recognition: A Cognition Special Issue.* Cambridge, MA: MIT Press, 71-102.

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

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive-activation model of context effects in letter perception: Part 1. An account of the basic findings. *Psychological Review*, **88**, 375-407.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, **76**, 165-178.

Morton, J., & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, **15**, 43-51.

Rubin, P., Turvey, M.T., & Van Gelder, P. (1976). Initial phonemes are detected faster in words than in non-words. *Perception and Psychophysics*, **19**, 394-398.

Rumelhart, D.E., & McClelland, J.L. (1982). An interactive-activation model of context effects in letter perception, Part II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, **89**, 60-94.

Segui, J., & Frauenfelder, U.H. (1986). The effect of lexical constraints upon speech perception. In F. Klix & H. Hagendorf (Eds.), *Human Memory and Cognitive Capabilities: Mechanisms and Performances.* North-Holland: Elsevier Science Publishers B.V., 795-808.

Segui, J., Frauenfelder, U.H., & Mehler, J. (1981) Phoneme monitoring, syllable monitoring, and lexical access. *British Journal of Psychology*, **72**, 471-477.

Stemberger, J.P., Elman, J.L., & Haden, P. (1985). Interference between phonemes during phoneme monitoring: Evidence for an interactive activation model of speech perception. *Journal of Experimental Psychology: Human Performance and Perception*, **11**(4), 475-489.

Psychological Similarities of Spoken Words[1]

Paul A. Luce[2], Jan Charles-Luce[3] and David B. Pisoni

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

[1] This is a written version of a poster presented at the meeting of the Psychonomics Society, New Orleans, November, 1990. This research was supported by NIH Grant #DC-00879-01 to the State University of New York at Buffalo, and NIH Training Grant #DC-00012-12 and NIH Research Grant #DC-00111-14 to Indiana University.

[2] Now at Language Perception Laboratory, Department of Psychology, State University of New York at Buffalo, Buffalo, NY 14260.

[3] Now at Department of Communicative Disorders and Sciences and Department of Linguistics, State University of New York at Buffalo, Buffalo, NY 14260.

Abstract

Subjects judged the overall similarities of consonant-vowel-consonant (CVC) words, consonant-vowel (CV) syllables, and vowel-consonant (VC) syllables. In addition, subjects judged the similarity of the individual consonants and vowels in both the CV and VC syllables. Correlation and regression analyses were performed in order to examine the degree to which the overall similarity of the words and syllables could be accounted for by similarity judgments of the individual consonants and vowels. The results are consistent with recent theories of syllable structure that propose a hierarchical organization of the components of spoken syllables.

Psychological Similarities of Spoken Words

Some recent linguistic theories of syllable structure propose a hierarchical organization of the components (i.e., consonants and vowels) of spoken syllables (Clements & Keyser, 1985). In particular, these theories propose that the syllable may be composed of: (1) a *rime*, the vowel and any consonants following the vowel, and (2) an *onset*, any consonants preceding the vowel. We examined the degree to which subjects are sensitive to this hierarchical syllable structure by using a direct similarity scaling technique.

Method

Subjects

The subjects were 280 undergraduates at Indiana University. All were native English speakers and had no history of speech or hearing disorders.

Stimuli

The stimuli were CVC English words, and CV and VC syllables over headphones. The CVC words consisted of 22 stop-vowel-stop sequences containing the stops /p,t,k,b,d,g/ and the vowels /i,a,u/ (e.g., /pip/, /kad/, and /but/). The CV syllables were 13 stop-vowel syllables having the same CV sequences found in the CVC syllables (e.g., /pi/, /ka/, and /bu/). Finally, the VC syllables were 12 vowel-stop syllables having the same VC sequences found in CVC words (e.g., /ip/, /ad/, and /ut/).

Procedure

On each trial of the experiment, subjects heard two words or two syllables. The subjects indicated how similar sounding the stimuli were on a ten point scale ranging from 1-very dissimilar to 10-very similar. A given subject made one of two types of judgments.[4] For one type of judgment, subjects judged the similarity of CVC words, CV syllables, or VC syllables by ranking how similar they thought two words or syllables sounded *overall*. For the other type of judgment, subjects were asked to judge the *components* of the CV and VC syllables by ranking how similar they thought the consonants or vowels in the two syllables sounded.

The stimuli and similarity judgments are summarized in Table 1. The stimuli are listed in the left-hand column. The similarity judgments made on these stimuli are listed in the right-hand column. The segments underlined indicate the portion of the stimulus on which the subjects were asked to base their judgments. For example, CV stimuli were judged based on the entire CV ($\underline{\text{CV}}$), the consonant alone ($\underline{\text{C}}\text{V}$), and the vowel alone ($\text{C}\underline{\text{V}}$). Altogether, there were seven conditions with 40 subjects per condition.

Insert Table 1 about here

Results

Judgments of the components of the CV and VC syllables were correlated with overall judgments of the CVC words, CV syllables, and VC syllables. For example, the average judged similarity of the two CVC words /but/ and /kad/ was correlated with each of the following four judgments of the

[4] All judgment conditions were between-subjects.

Table 1

Seven stimulus conditions.

Stimulus	**Judgment**
CVC	$\underline{\mathrm{CVC}}$
CV	$\underline{\mathrm{CV}}$ $\underline{\mathrm{C}}\mathrm{V}$ $\mathrm{C}\underline{\mathrm{V}}$
VC	$\underline{\mathrm{VC}}$ $\underline{\mathrm{V}}\mathrm{C}$ $\mathrm{V}\underline{\mathrm{C}}$

components: (1) the initial consonants /b/ and /k/ in CV syllables, (2) the vowels /u/ and /a/ in CV syllables, (3) the vowels /u/ and /a/ in VC syllables, and (4) the final consonants /t/ and /d/ in VC syllables.

In addition to the correlation analysis, judgments of the components were regressed against the overall judgments using hierarchical multiple regression analyses. The independent variables (component judgments of consonants and vowels) were entered in two orders: those with the highest correlation to those with the lowest correlation, and vice versa.

Correlation of C and V Judgments with CVC Judgments

Figure 1 shows the correlations between similarity judgments of consonants and vowels in the CV and VC syllables with overall similarity judgments of the CVC words. The correlations are shown separately for consonants and vowels occurring in CV and VC syllables. Consonant judgments in CV syllables correlated more highly [r^2=.56] with overall CVC judgments than consonant judgments in VC syllables [r^2=.44]. However, the opposite pattern of results was found for the vowel judgments; judgments of vowels in VC syllables correlated more highly [r^2=.81] than judgments of vowels in CV syllables [r^2=.61].

Insert Figure 1 about here

Regression of C and V Judgements Against CVC Judgments

Figure 2 shows the results of a hierarchical regression analysis in which judgments of the individual components were regressed against the overall CVC judgments. The order of the steps in the regression analysis was determined by entering the independent variable with the highest correlation, followed by the variable with the next highest correlation, followed by the variable with the lowest correlation. Only the vowel judgment with the highest correlation (i.e., from the VC syllable) was used in the regression analysis. The proportion of variance accounted for is plotted on the y-axis. Successive steps in the analysis are plotted on the x-axis. From left to right, each bar represents a step in the regression analysis. Variables entered at each step are shown below each bar. V refers to the vowel, C1 refers to the consonant in CV syllables, and C2 refers to C2 in VC syllables. At the first step, the vowel accounts for 0.6592 of the variance. C1 accounts for 0.1279 additional variance, and C2 accounts for 0.0545 of the variance. Overall, 0.8416 of the total variance in the overall CVC similarity judgments is accounted for by these component judgments.

Insert Figure 2 about here

Figure 3 shows the results of hierarchical regression analysis in which the order of the variables was determined by first entering the variable with the lowest correlation, followed by the variable with the intermediate correlation, followed by the variable with the highest correlation. Again, only the vowel judgment with the highest correlation (i.e., from the VC syllable) was used in the regression analysis. Aside from the ordering of the variables, the format of Figure 3 is the same as that for Figure 2. At the first step, C2 accounts for 0.1942 of the variance. C1 accounts for 0.3111 additional variance, and V accounts for 0.3362 of the variance.

Insert Figure 3 about here

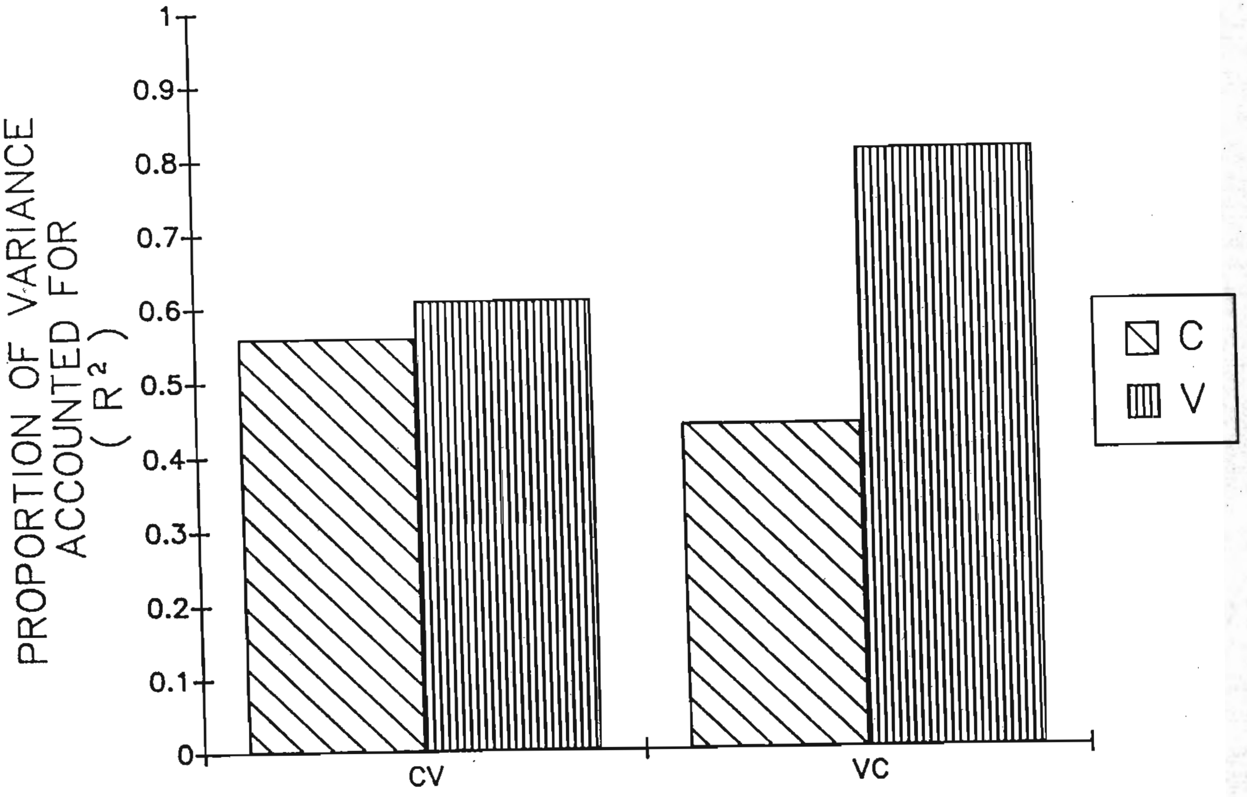


Figure 1. Correlation of C and V judgements with CVC judgements. Proportion of variance accounted for [r^2] is plotted on the y-axis. The type of syllable (CV or VC) on which the judgments were made is plotted on the x-axis. For CV and VC syllables, each bar shows the proportion of variance accounted for by the correlation between either consonant (diagonally striped) or vowel (vertically striped) similarity judgement and the overall CVC judgement.

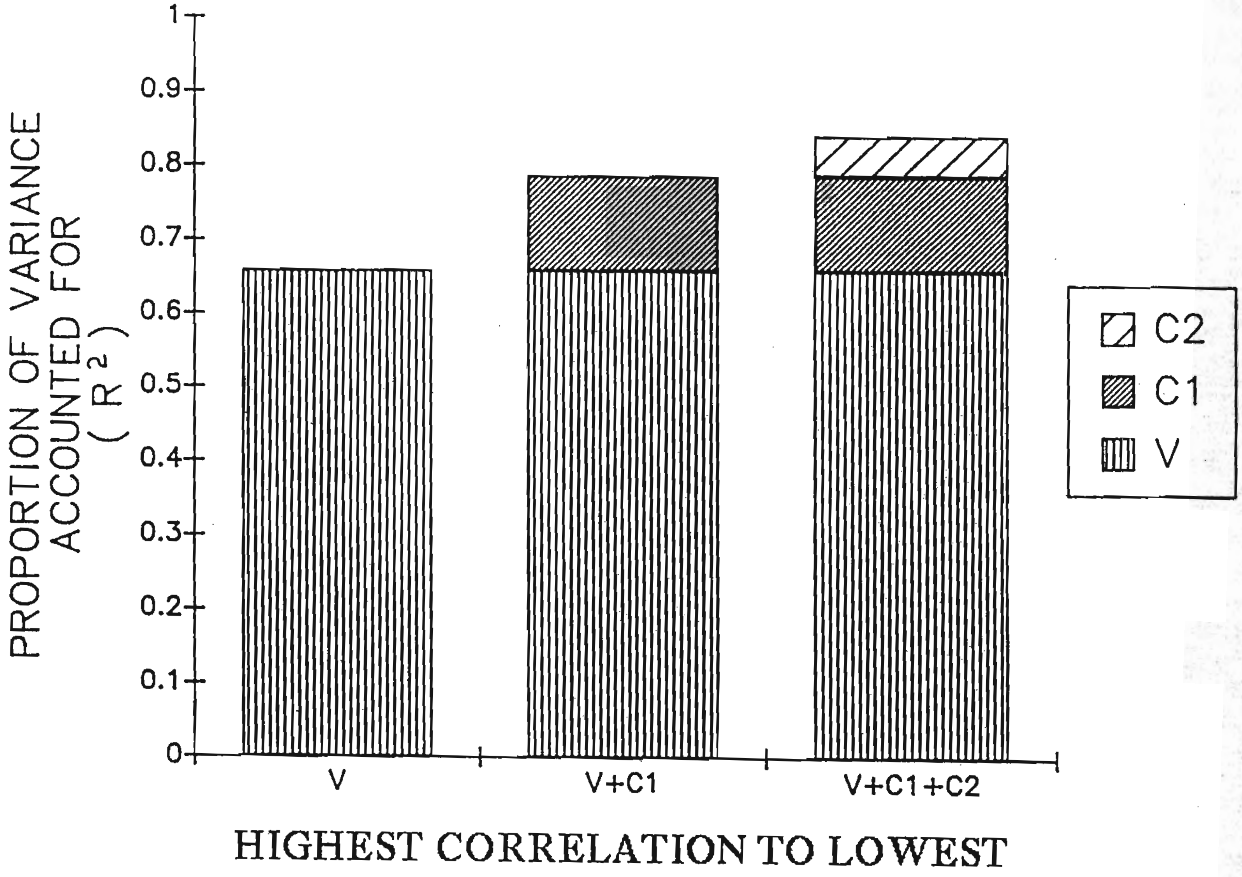


Figure 2. Regression of C and V judgements against CVC judgements. The independent variable with the highest correlation was entered into the regression first.

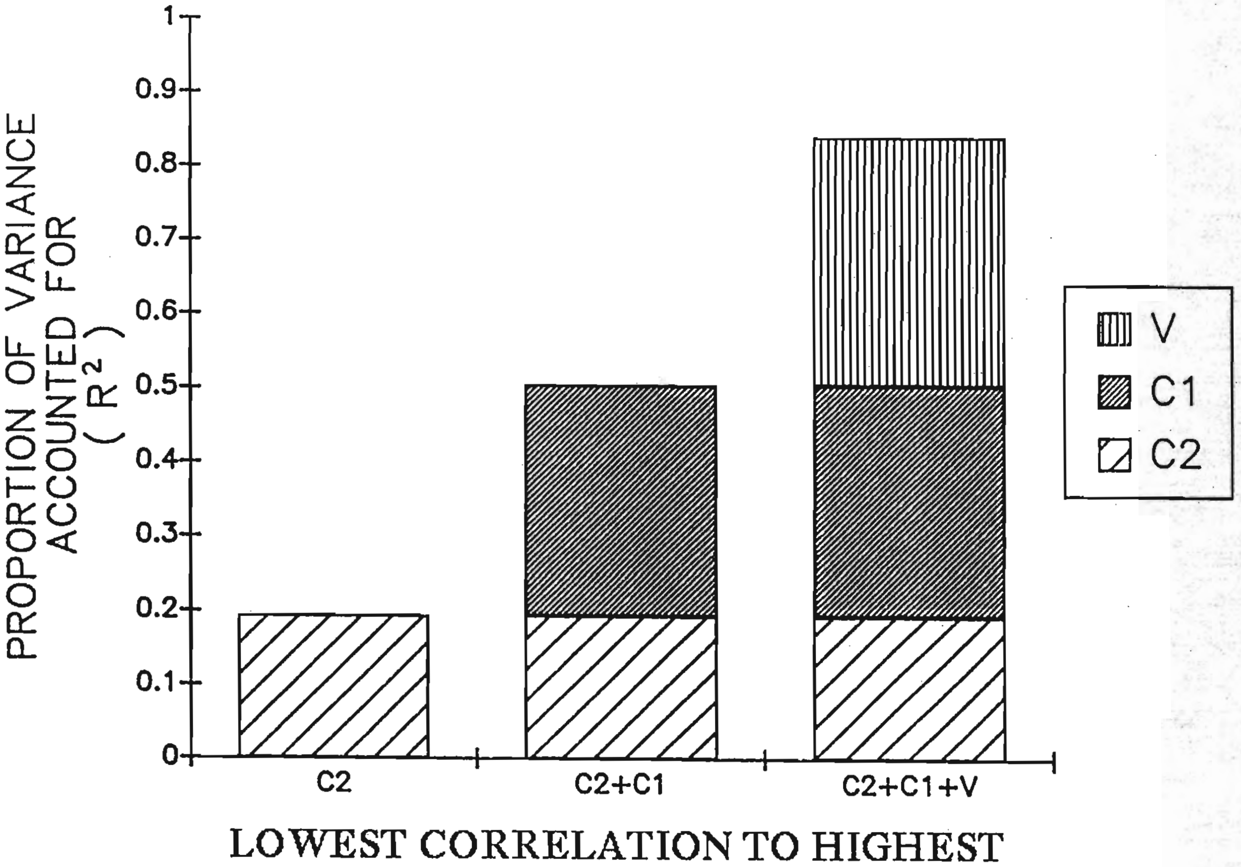


Figure 3. Regression of C and V judgements against CVC judgements. The independent variable with the lowest correlation was entered into the regression first.

In both regression analyses, the vowel accounts for most of the variance of the overall similarity judgments for the CVC words. C2, however, accounts for the least amount of variance. In both analyses, C1 proved to be a stronger variable in accounting for overall judgments than C2.

Correlation of C and V Judgments with CV Judgments

Figure 4 shows the correlations between similarity judgments of individual consonants and vowels in CV syllables with the overall similarity judgments of CV syllables themselves. In CV syllables vowel judgments correlated more highly [r^2=.95] than consonant judgements [r^2=.65] with overall CV judgments. However, consonants still appear, in this analysis, contribute to overall judged similarity. Regression analyses were again performed to evaluate the relative contributions of the independent variables to overall similarity judgments.

Insert Figure 4 about here

Regression of C and V Judgments Against CV Judgments

Figures 5 and 6 show the results of hierarchical regression analysis in which judgments of the individual components were regressed against the overall CV judgments. In Figure 5 the order of steps in the regression analysis was from highest to lowest correlation. At the first step the vowel accounts for 0.8933 of the variance. C1 accounts for 0.0892 additional variance. Overall, 0.9825 of the total variance in the overall CV similarity judgments is accounted for by these component judgments. In Figure 6 the order of steps in the regression analysis was from lowest to highest. C1 accounts for 0.4136 variance when entered first, and the vowel accounts for an additional 0.5689 of the variance. Therefore, in both regression analyses, the vowel is superior to C1, although C1 still contributes, especially when entered prior to the vowel.

Insert Figures 5 and 6 about here

Correlation of C and V Judgments with VC Judgments

Figure 7 shows the correlations between similarity judgments of component vowels and consonants in VC syllables with the overall similarity judgments of the VC syllables themselves. Vowel judgments in VC syllables correlated more highly [r^2=.96] than consonant judgments [r^2=.35] with overall VC judgments. The results further demonstrate that consonant judgments in VC syllables contribute much less in terms of variance accounted for [r^2=.35], than consonant judgments in CV syllables [r^2=.65] (compare Figures 4 and 7).

Insert Figure 7 about here

Regression of C and V Judgments Against VC Judgments

Figures 8 and 9 show the results of hierarchical regression analysis in which judgments of the individual components were regressed against the overall VC judgments. In Figure 8 the order of the steps in the regression analysis is from highest to lowest correlation. At the first step, the vowel accounts for 0.9326 of the variance and C1 accounts for only 0.0506 additional variance. For VC similarity jugdements, 0.9832 of the total variance is accounted for by these component judgments. In Figure 9 the order of steps in the regression analysis is from lowest to highest. C1 accounts for 0.1258 variance when entered first and the vowel accounts for an additional 0.8484 of the variance. Therefore, in both

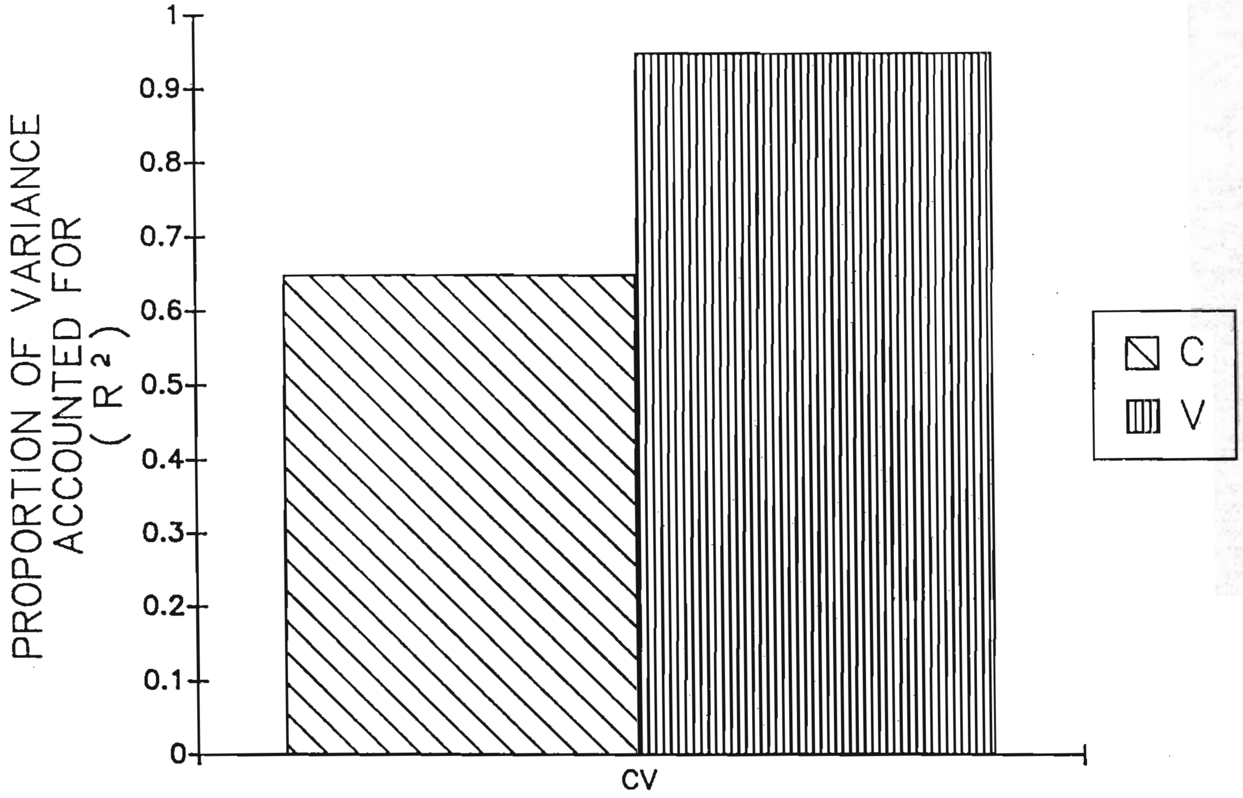


Figure 4. Correlation of C and V judgements with CV judgements. Proportion of variance accounted for [r^2] is plotted on the y-axis. Each bar shows the proportion of variance accounted for by the correlation between either consonant (diagonally striped) or vowel (vertically striped) similarity judgement and the overall CV judgement.

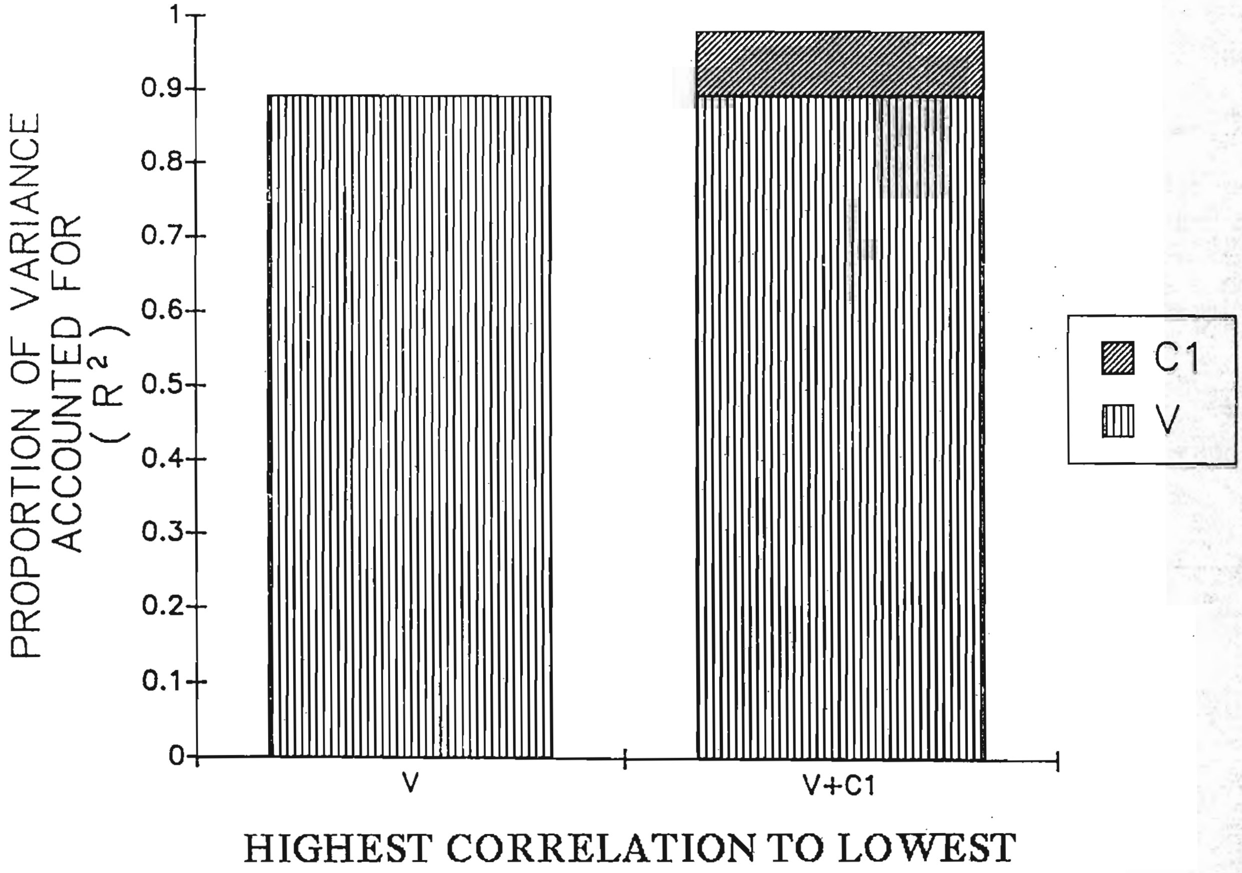


Figure 5. Regression of C and V judgements against CV judgements. The independent variable with the highest correlation was entered into the regression first.

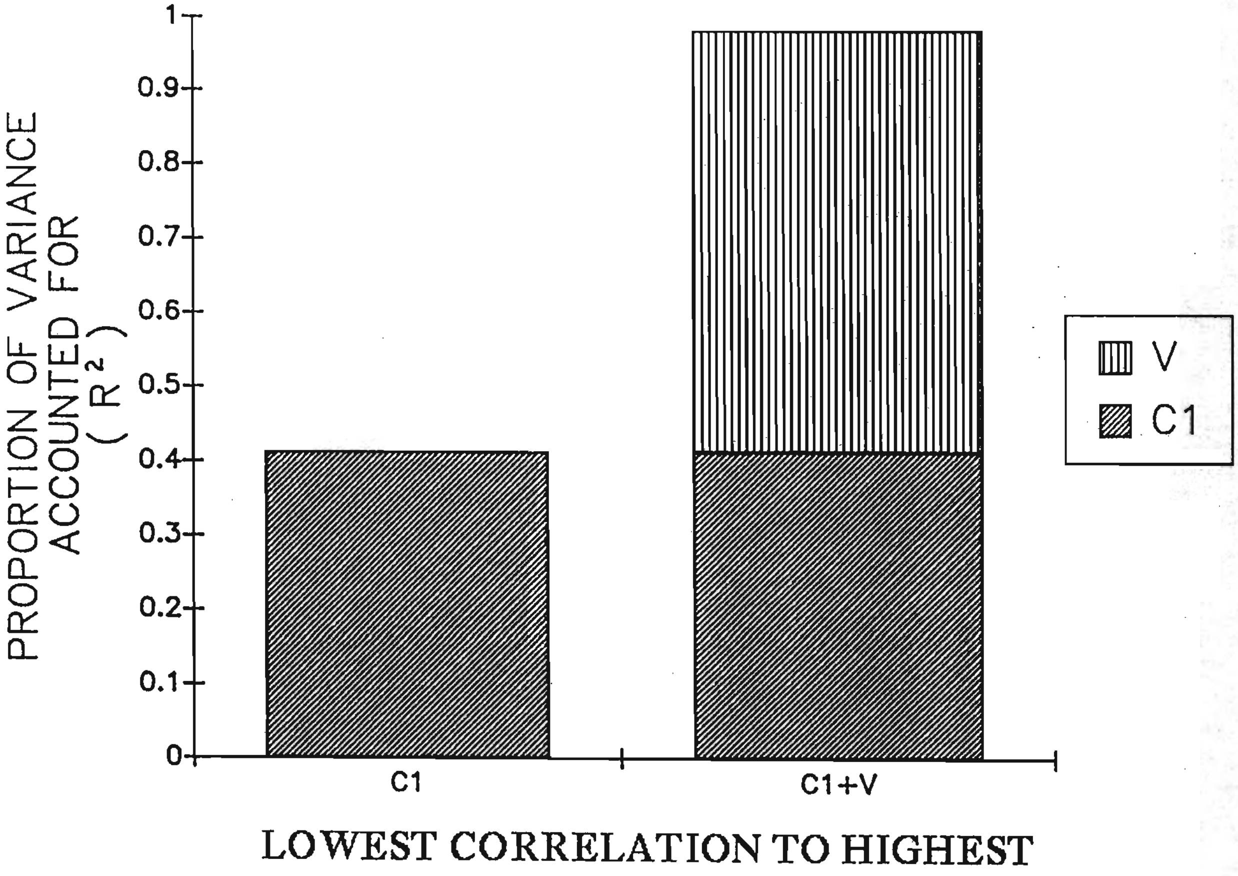


Figure 6. Regression of C and V judgements against CV judgements. The independent variable with the lowest correlation was entered into the regression first.

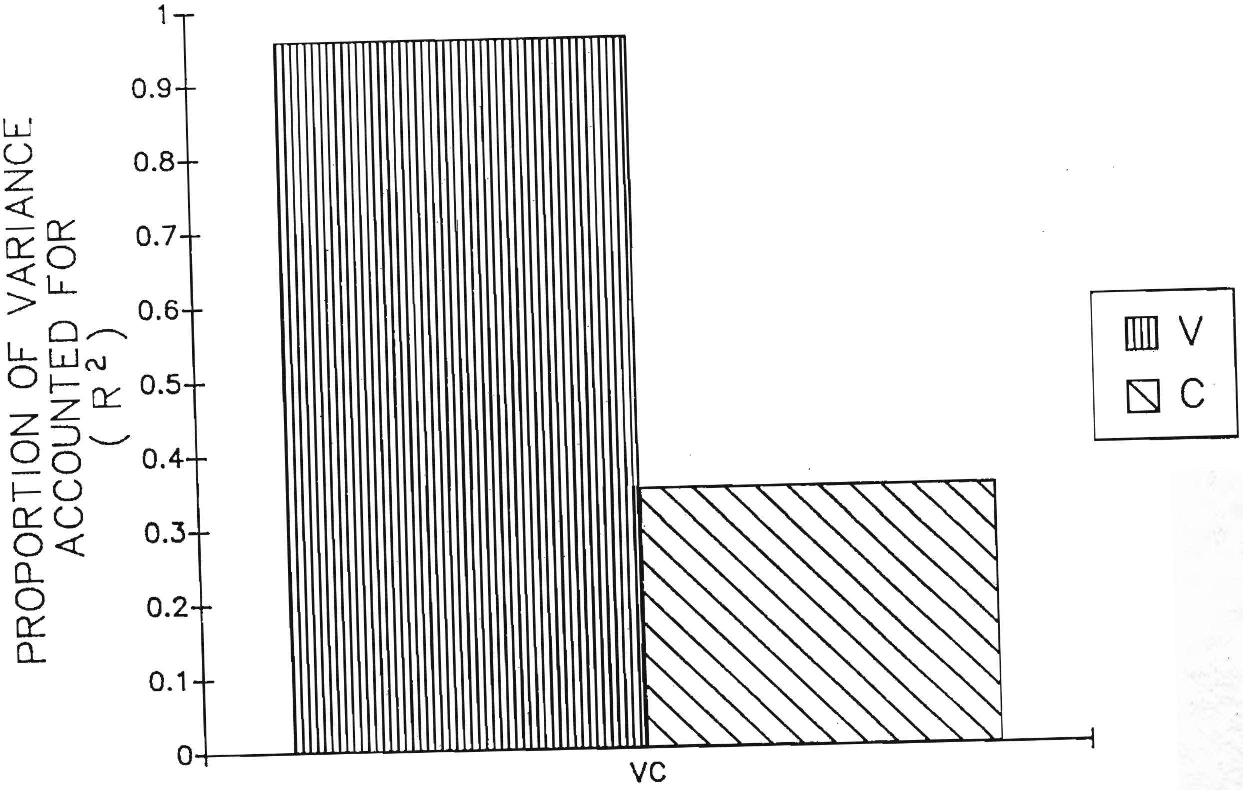


Figure 7. Correlation of C and V judgements with VC judgements. Proportion of variance accounted for [r^2] is plotted on the y-axis. Each bar shows the proportion of variance accounted for by the correlation between either consonant (diagonally striped) or vowel (vertically striped) similarity judgement and the overall VC judgement.

regression analyses the vowel is superior to C2. Furthermore, this analysis demonstrates that C2 contributes much less to overall VC judgment than C1 contributes to overall CV judgments.

Insert Figures 8 and 9 about here

Summary and Conclusions

Similarity judgments of vowels in both CV and VC syllables best account for overall judgments of CVC words. Similarity judgments of vowels in VC syllables can better account for overall CVC judgements than vowel judgments in CV syllables. Finally, similarity judgments of initial consonants in CV syllables are superior to judgments of final consonants in VC syllables in accounting for overall judgments of CVC words. In addition, consonants in CV syllables are superior to consonants in VC syllables in accounting for overall judgments of CV and VC syllables.

These data are consistent with a hierarchical theory of syllable structure (see Clements & Keyser, 1985) that states that syllables are composed of onsets and rimes. We found that vowels are the primary units upon which similarity judgments are based. More interesting, however, our results demonstrate that vowels and consonants in VC syllables appear to form a more cohesive unit than consonants and vowels in CV syllables. Consonants in CV syllables independently contribute more to overall similarity judgments than consonants in VC syllables, as might be predicted by a hierarchical theory of syllable structure.

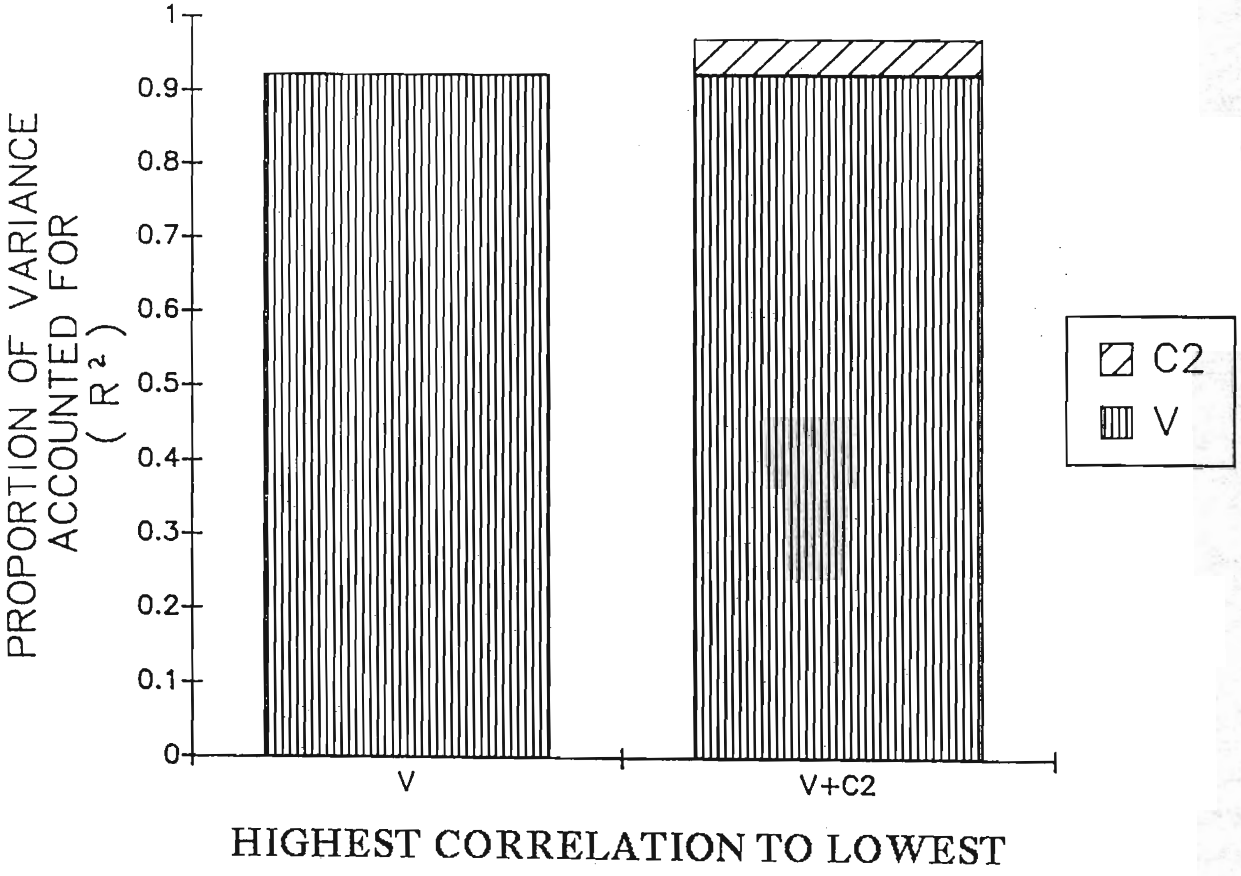


Figure 8. Regression of C and V judgements against VC judgements. The independent variable with the highest correlation was entered into the regression first.

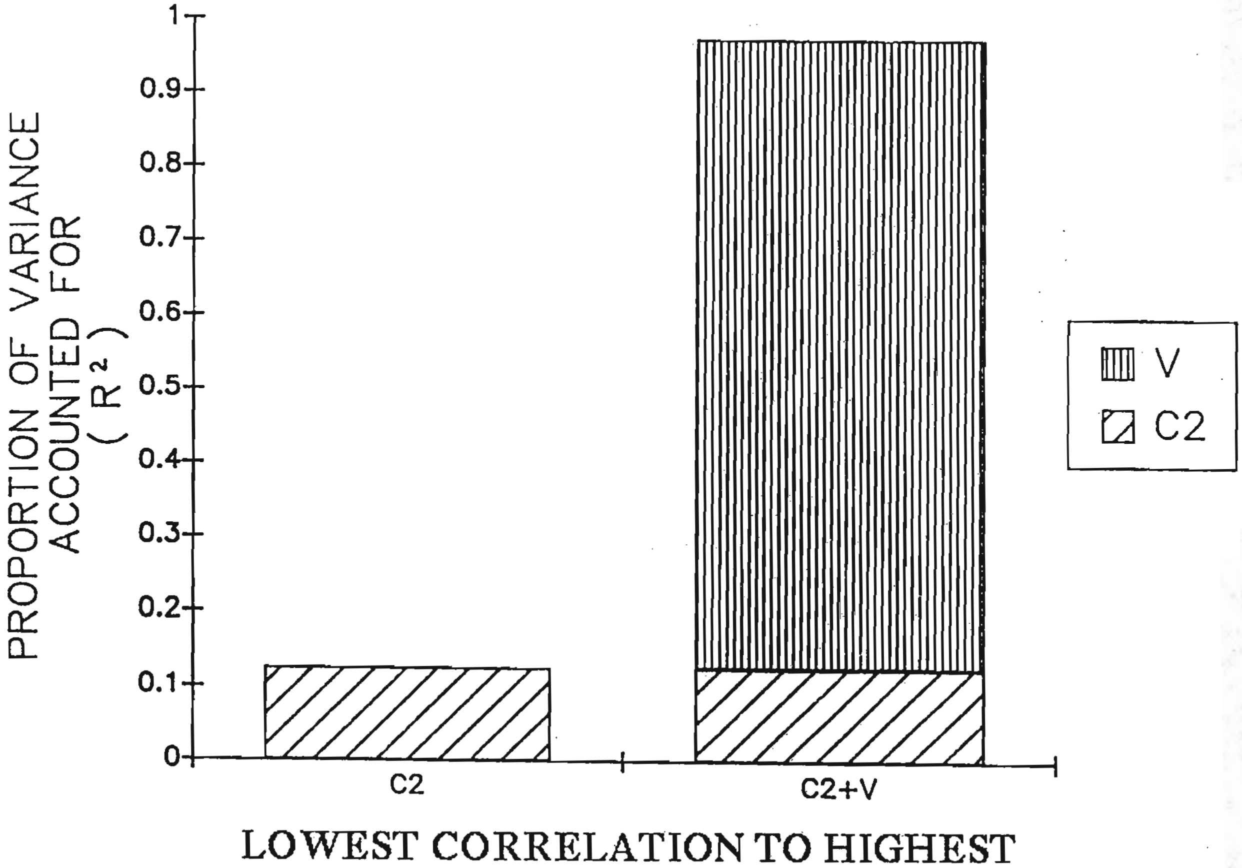


Figure 9. Regression of C and V judgements against VC judgements. The independent variable with the lowest correlation was entered into the regression first.

References

Clements, G.N., & Keyser, S.J. (1985). *CV Phonology: A Generative Theory of the Syllable.* Cambridge, MA: MIT Press.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Talker Normalization and Word Recognition in Preschool Children

Brigette R. Oliver[1]

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

[1] The author would like to thank David Pisoni, Linda B. Smith, and Andrea Hussong for their help in the planning and completion of this study. The research reported here was supported by NIH Research Grant #DC-0111-14 and NIH Training Grant #DC-00012-12 to Indiana University.

Abstract

The effects of talker variability on spoken word recognition were studied developmentally in three, four, and five-year old children. Subjects listened to lists of words presented in noise (S/N ratio=0) and identified each word by pointing to a picture in a six-alternative visual display. The words and pictures were taken from the Word Intelligibility by Picture Identification test (WIPI). Two talker conditions were examined: single talker and multiple talker. The design was within subjects, with each child hearing both a single-talker list and a multiple-talker list, counterbalanced for talker condition, list, and order. Results showed main effects of age and talker condition and a marginally significant interaction between these two variables. As expected, we found an increase in overall accuracy with age. Subjects also performed better on the single talker list than the multiple talker list at all ages. The marginal interaction indicates that, not only is there a cost for talker normalization for pre-school children, but that it appears it has a higher cost the younger children.

Talker Normalization and Word Recognition in Preschool Children

Every day we are likely to communicate through spoken language with numerous other individuals. Rarely, however, do we notice the sometimes dramatic acoustic differences between individual voices. Even more rarely does this inherent variability cause us perceptual difficulty in understanding what is being said. Most theorists believe that in order to perceive speech, listeners must "normalize" the acoustic cues that vary from speaker to speaker (Mullennix, Pisoni, & Martin, 1989). Until recently, however, very little was known about talker normalization, except for the obvious fact that we are quite adept at coping with different voices. In the past few years, a number of investigations into talker normalization have been completed in our laboratory. As a result, we now know when talker variability is likely to produce effects, and what these effects will look like. What we do not yet fully understand is how normalization is accomplished. In other words, what is the mechanism for talker normalization? In this study, we hope to shed light on the mechanism of talker normalization by investigating its operation developmentally. A critical question is whether or not the processes of normalization are in place from the beginning of language development, or whether the processes develop with experience.

One of the first reported studies on the effects of talker normalization was conducted by Creelman (1957) who found that performance in a speech intelligibility task was poorer when the lists were spoken by more than one talker. Specifically, Creelman found that lists spoken by two or more talkers resulted in less accurate performance than lists produced by a single talker. Because there were methodological problems with Creelman's study, Mullennix, Pisoni, and Martin (1989), attempted to replicate the perceptual identification experiment in a more controlled manner. They also extended on the earlier work by including a naming task, in addition to the identification paradigm. Mullennix et al. (1989), like Creelman, found that performance was reduced in the identification task. They also found that naming latencies were longer in the multiple talker condition than the single talker condition. Perhaps more importantly, these authors found that the differences between talker conditions were more robust and less task dependant than other commonly reported differences such as those due to word frequency and lexical density. Goldinger, Pisoni, and Logan (1991) have suggested that talker variability effects are more robust because talker information is physically concrete, unlike the more abstract nature of word frequency and neighborhood confusability. From these studies, we can conclude that a cost-incurring process of talker normalization does indeed play a part in the perception of speech.

Mullennix, Pisoni, and Martin (1989) concluded that talker variability has its effect in very early acoustic-phonetic encoding processes, and that normalization does not interact with higher level structural variables (e.g., word frequency). A second study was conducted in the same laboratory (Martin, Mullennix, Pisoni, & Summers, 1990) in order to examine the possibility that talker normalization might have effects on other cognitive processes, such as memory. More specifically, Martin et al. examined recall of spoken word lists produced by either a single talker or by multiple talkers. In a series of experiments, Martin et al. (1990) varied the memory load and the retention interval, as well as the talker conditions. Their findings indicate that the primacy effect (i.e., better recall of words presented early in a list) was decreased in the multiple-talker conditions as compared to the single-talker condition. The researchers explained their results by saying that multiple-talker lists require more processing resources which interferes with the rehearsal processes that are generally thought to produce the primacy effect in serial recall.

The 'interference' above could be caused by two alternative factors, between which Martin et al. (1990) were unable to distinguish in their study. The first possibility is that talker-normalization affects only early perceptual encoding, and that it is this extra cost that indirectly effects rehearsal capacity. The second hypothesis is that talker-normalization affects both encoding and rehearsal directly. In order to evaluate these possibilities, Goldinger, Pisoni, and Logan (1991) recently conducted a study in which rate of presentation was manipulated, among other variables. The rate of presentation of list items is believed to directly effect rehearsal processes (Goldinger et al., 1991). These authors found that recall of the multiple-talker lists was affected more by the rate of presentation manipulation than recall of the single-talker lists. They therefore concluded that talker normalization increases processing demands for both initial encoding and later rehearsal processes.

Based on the literature reviewed so far, we can draw several conclusions. First, there are costs for talker normalization in a number of different tasks. Second, talker normalization is costly for more than one processing stage in a recall paradigm, those being initial perceptual encoding and later rehearsal. What else do we know about the mechanisms and processes involved? One interesting finding is that of Mullennix and Pisoni (1990). They found, using a Garner (1973) speeded classification task, that subjects could not selectively ignore information about voice, even when it was beneficial to do so. This contrasts with the subjects ability to ignore the phonetic content. The authors concluded that voice information is processed in a mandatory fashion and stored as an integral component of the representation in memory. Their findings corroborates other findings by Geiselman and Bellezza (1976) who report that information about a speaker's voice is retained even when subjects are given no instructions to attend to such information.

While talker normalization produced negative effects in the studies reported above, the fact that voice information is stored as part of the memory representation leads us to question whether or not this information could be used as a cue to facilitate memory. A study by Lightfoot (1990) investigated this question by training subjects to recognize a set of voices who were then used in subsequent memory experiments. She found that familiar voices could indeed be used as a cue to memory. Specifically, Lightfoot found that familiar voices facilitated memory in the primacy portion of the list when presentation rates are slow. This is the same portion of the list that was shown to be negatively affected by multiple unfamiliar talkers (Martin, Mullennix, Pisoni, & Summers, 1990).

In addition to the studies done with adults, recent research has investigated the perceptual normalization abilities of infants. One of the first such studies was done by Kuhl (1979, 1983). In her studies, she found that infants only six-months of age could generalize a learned vowel contrast to different voices. This is not due to an inability to distinguish voices, however, since a newborn shows a preference for his mother's voice. Like adults, this ability to normalize across voices comes with some cost to processing. As demonstrated by Jusczyk, Pisoni, and Mullennix (1989), two-month old infants tested in a high amplitude sucking (HAS) paradigm who habituated to one set of voices repeating a syllable did not dishabituate when switched to another set of voices repeating a different syllable. This contrasts with their dishabituation in all single talker conditions and in the multiple talker condition where only the syllable changed. Although the infants did dishabituate in this latter multiple-talker condition, they took longer to habituate than in the single-talker conditions. Another study by Jusczyk et al. (1989) examined the effect that perceptual normalization might have on the infants memory for speech. Without going into great detail, the results suggest that different talkers disrupt infants' ability to remember the phonetic identity of syllables, but this effect was found only when there was a delay period between the habituation and dishabituation phases.

While the infant work is important and suggestive, it is difficult to interpret with regard to the mechanisms involved in talker-normalization. Researchers generally assume that infants this young are operating at a more general acoustic level, and only later come to process speech as a special stimulus (Jusczyk, 1989). In addition, the types of stimuli used in the infant studies are typically CV syllables, not words as are used with adults. It is possible that infants may be categorizing these stimuli based on perceptual similarity and not truly 'normalizing' in the same manner as adults. If we wish to investigate the development of talker normalization processes, it is important to choose subjects whom we can be sure are normalizing speech in the same manner as adults. The prime candidates for such study are preschoolers, whose language facility is improving at an amazing pace. For instance, they learn an average of nine new words a day (Carey, 1978). However, we can assume that preschoolers' speech perception skills are not yet fully developed (Walley, 1988). By studying these children we can begin to bridge the gap between the infant and adult studies while also asking very important questions about the mechanisms involved in talker normalization. That is, between the ages of three and five, do talker normalization abilities develop? How much of a processing cost is there for talker normalization in children this young? Given that stimuli must usually be degraded for adults to show differences due to talker variability, are children equally adept or is the process costly even under normal stimulus situations? Questions like these have never been posed. Answers are potentially important, both for understanding the development of speech perception, and also for shedding light on the mechanisms involved in talker-normalization in general. Our purpose is to begin to address these questions. Through this work we hope to describe in detail the talker-normalization skills possessed by preschool children, to examine any developmental trends, and in the process to possibly constrain the mechanisms that theorists propose to account for talker-normalization.

In the following study, we investigated the effect that talker variability has on word recognition in three, four, and five-year old children. We hypothesized that, when words are presented in noise, talker variability will result in poorer word identification performance than when words are spoken by a single talker at all ages. We also predict that if talker-normalization processes change with development, performance in the multiple-talker condition will be harder than in the single-talker condition the younger the child.

Method

Subjects

Thirty-six children, 12 each at ages 3, 4, and 5, were recruited to participate in this experiment from the surrounding community by an ad in the local paper. The average age for each group was 3.47, 4.62, and 5.68. Each subject was run separately in a single session lasting approximately half an hour. Subjects were paid for their participation.

Stimulus Materials

Two word lists of 25 words each from the Word Intelligibility by Picture Identification (WIPI) test were used as stimuli for this experiment (Ross & Lerman, 1970). The WIPI is designed to assess the speech discrimination abilities of young children. All words are monosyllabic and have an average familiarity of 6.957 (Nusbaum, Pisoni, & Davis, 1984) and an average frequency of 99.45 (Kucera & Francis, 1967). In its regular clinical usage, the words are read aloud in a live voice by the person administering the test. The child is shown a six-alternative visual display for each word and are instructed to identify the word by pointing to the correct picture. For our purposes, the lists of stimulus words were prerecorded on audio tape and were played back to the children over headphones. All other procedures remained the same.

As reported in a previous pilot study (Oliver, 1989), the stimulus materials were originally produced by seven males and seven females. The words were presented randomly via a CRT screen to the talker who was seated in a sound-attenuated booth (IAC model 401A). Their utterances were recorded using an Electro-Voice model D054 microphone and an Ampex AG-500 tape recorder. All talkers were told to read the words aloud in a normal voice at a constant speaking rate. The words were then converted to digital form using a 12-bit analog-to-digital converter running at a 10-kHz sampling rate. The RMS amplitude levels of the words were digitally equated and the test words were edited using a digitally controlled waveform editor (Luce & Carrell, 1981).

The resulting stimulus tokens spoken by the fourteen talkers were presented to adult subjects to obtain identification scores. Six subjects participated in two one-hour sessions. In one session the subjects rated the tokens spoken by males; in the other session, they rated the tokens spoken by the females. All stimuli were presented via headphones and subjects recorded the word they heard by typing responses into a computer. Results were tallied, with percentage of correct responses taken as a measure of intelligibility. The male talker with the highest overall identification score was chosen for use in the single-talker condition. All tokens produced by this speaker had at least an 86% correct identification score. Audio tapes were made using this voice for each of the two lists used in this experiment.[2] The tokens spoken by the five male and five female voices with the next highest identification scores were used to construct the multiple talker lists. The tokens produced by the one remaining male speaker and two remaining female talkers with the lowest identification scores were eliminated and were not used to construct stimulus tapes. Tokens were chosen at random from among the ten talkers with the requirement that the identification scores be at least 86% correct or higher. Each talker was used approximately the same number of times on each list, and lists were balanced for gender.

Design and Procedure

Each child was tested individually in a single session lasting less than 30 minutes. Prior to participation in the experiment, all subjects were given a pure-tone screening test at frequencies of 500Hz, 1000Hz, 2000Hz, and 4000Hz. No children were rejected due to hearing problems as indicated by this test. The child sat across from the experimenter, either next to or on the parent's lap. Parents were asked not to assist the child in any way during the experiment. The procedure was explained to the children as a game where they could win stickers. They were instructed to listen to the words presented through the earphones and point to the pictures of what they heard.

Before beginning the actual experiment, a practice trial was completed to insure that the child understood the instructions. The practice trial was similar to the experiment except that the words were said aloud in a live voice by the experimenter. The experimenter asked "What would you point to if you hear the word 'x'?" with 'x' being one of the six pictures on the sample page (e.g., cat). This procedure was repeated one or two more times to ensure the child understood his or her task. None of the children had difficulty understanding the procedure. The experiment then began with the first list of stimulus words presented in noise (S/N ratio=0) to the subjects through TDH-39 headphones using a Uher 4000 Report-L tape recorder.

The experimenter conducted each trial by saying "show me this", or a similar prompt, then playing a test word and pausing the recorder until the subject produced a response. Responses were recorded by the experimenter on a response sheet. The experiment continued until the 25 words from the first list were completed. Periodically throughout the procedure the experimenter would remind the child of the instructions. After the first list was completed, the child got to choose a 'prize' sticker, regardless of actual performance. After a short break of 1-3 minutes the child was asked if she or he

[2] A third list from the WIPI test was not used in this experiment.

would like a chance to earn another sticker by completing another list of words. None of the children refused and the second list was completed in the same manner as the first. The two lists (List 1 and List 2) and two talker conditions (single or multiple) were completely counterbalanced by list order (first or second). Three children at each age participated in each of the resulting four conditions (S1M2, S2M1, M1S2, and M2S1).

Results and Discussion

For analyses, the number of correct responses was converted to a percent correct score. Because there was no indication of a gender difference in the initial analysis, the data were collapsed across that variable. Performance was moderately high with an overall mean of 19.91, or 80% of the words chosen correctly. Figure 1 shows the data plotted as a function of age and talker condition.

Insert Figure 1 about here

Data were analyzed using a within-subjects ANOVA with the factors age and talker condition. There were three age levels and two levels of the talker condition: single-talker and multiple-talker. The ANOVA revealed a main effect of age [$F(2,33)=17.14$, $p<.001$], and a main effect of talker condition [$F(2,33)=56.27$, $p<.001$]. As expected, overall accuracy increased with age. Performance was better in the single talker condition than in the multiple talker condition at all ages (see Figure 1). This finding is not extremely informative. Given that talker variability effects are quite robust in adults, we expected the same or more from these children and were not disappointed in the results.

Insert Figure 2 about here

To evaluate the possibility of an interaction between age and talker condition, we calculated a ratio score for each subject by dividing the number of words correctly identified in the multiple talker condition by the number of words correctly identified in the single talker condition (see Figure 2). A ratio score of one indicates that performance was equal in both talker conditions. Values lower than one indicate that the multiple talker condition was more difficult than the single talker condition. These scores were analyzed in an ANOVA using the three-level factor of age. The result was marginally significant [$F(2,33)=3.21$, $p=.0533$]. Post-hoc Tukeys HSD analyses conducted on the ratio scores reveal that the 5-year olds differed significantly from the 3-year olds, but that neither were significantly different from the 4-year olds. We take this significant difference as an indication that the multiple talker list was harder for the three-year olds than the five-year old children, relative to the single talker list. In essence, this is an indication that talker normalization becomes easier or less costly with development. Presumably, the mechanism and/or processes involved in talker normalization do become more efficient, either through maturational processes or some effect of experience. Perhaps, as the children become familiar with more voices, they become more adept at 'tuning' the mechanism responsible for talker-normalization. Because initial adjustment to different voices would become easier, more processing capacities would be available for other operations. While this is mere speculation, it appears that the mechanisms proposed for talker-normalization must be ones capable of developing.

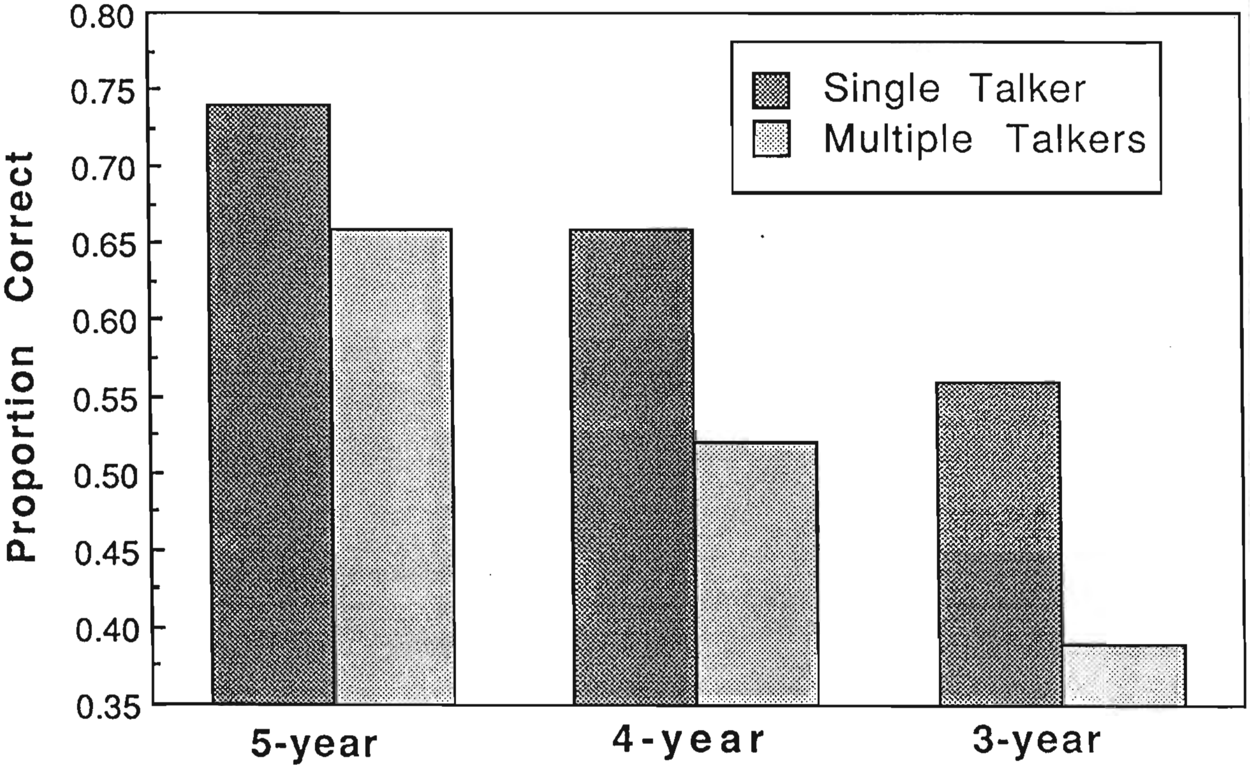


Figure 1. Proportion of words correctly identified, plotted by age and talker (S/N ratio=0).

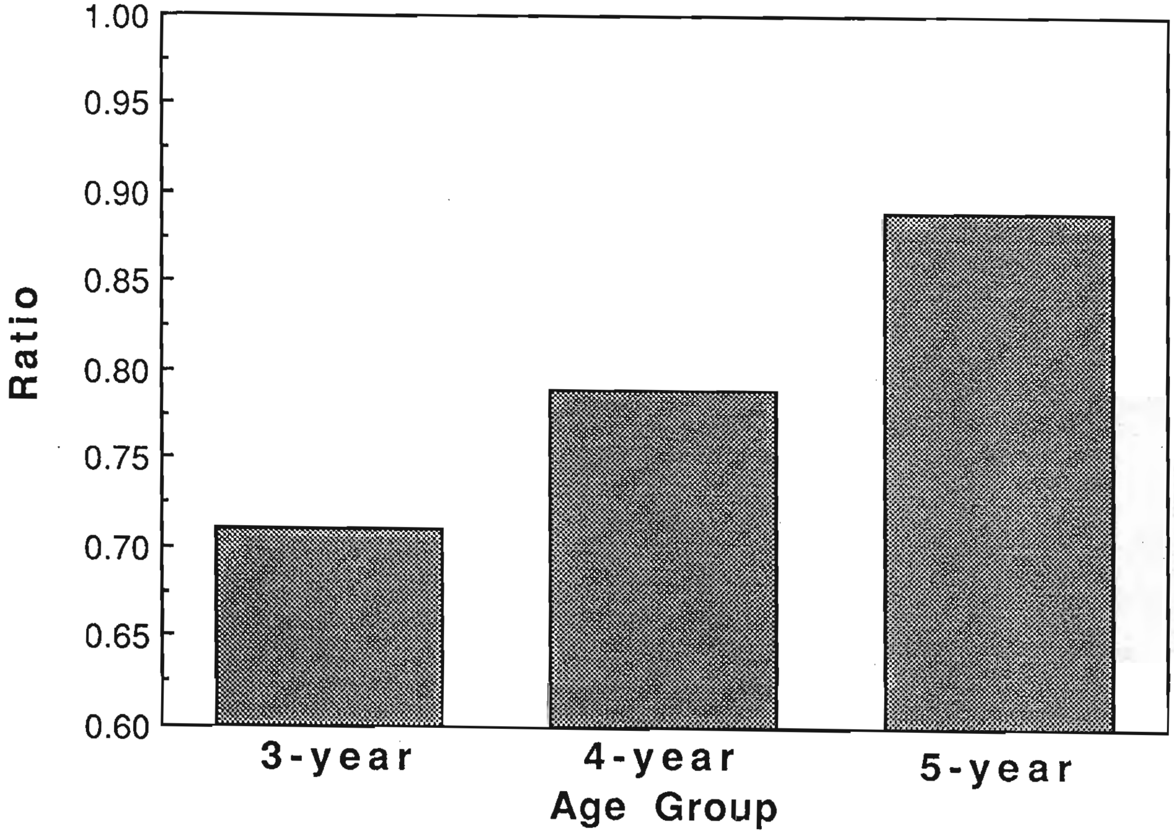


Figure 2. Ratio of single talker score to multiple talker score by age group ($p < .0533$).

References

Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., & Miller, G. (Eds.), *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.

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

Garner, W.R. (1973). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

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

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

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. W. (1989). Some effects of talker variability on speech perception in 2-month-old children. *Research on Speech Perception Progress Report No. 15.* Bloomington IN: Speech Research Laboratory, Indiana University.

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

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, **66**, 1668-1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, **6**, 263-285.

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

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

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1987). Effects of Talker Variability on recall of spoken word lists. *Research on Speech Perception Progress Report No. 13.* Bloomington IN: Speech Research Laboratory, Indiana University.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, **47**(4), 379-390.

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

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.

Oliver, B.R. (1989). Talker variability and word recognition: A developmental study. *Research on Speech Perception Progress Report No.15.* Bloomington, IN: Speech Research Laboratory, Indiana University.

Ross, M. & Lerman, J. (1970). A picture identification test for hearing impaired children. *Journal of Speech and Hearing Research*, **13**, 44-53.

Walley, A.C. (1988). Spoken word recognition by young children and adults. *Cognitive Development*, **3**, 137-165.

Episodic Encoding of Voice and Recognition Memory for Spoken Words[1]

Thomas J. Palmeri, Stephen D. Goldinger and David B. Pisoni

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

[1] We wish to thank Scott Lively, Lynne Nygaard, and especially Richard Shiffrin for their valuable comments. This research was supported by NIH Research Grant #DC-00111-14 to Indiana University.

Abstract

The effects of talker variability on recognition memory for spoken words were investigated using a continuous recognition memory task (Shepard & Teghtsoonian, 1961). The number of intervening items, or lag, between the initial presentation and the repetition of a spoken word, and the number of talkers who produced the words were varied. Half of the words in all lists were presented and repeated in the same voice and half were repeated in a different voice. Recognition judgments were based on word identity alone. Different groups of subjects heard words spoken from a set of either 1, 2, 6, 12, or 20 talkers. The results replicated and extended findings reported by Craik and Kirsner (1974) that same-voice repetitions are responded to more quickly and accurately than different-voice repetitions at all values of lag. No effect of increasing talker variability was found with either accuracy or response time measures. The present results suggest that information about a talker's voice is retained in long-term episodic memory representations. These findings are inconsistent with current views of talker normalization in theories of speech perception.

Episodic Encoding of Voice and Recognition Memory for Spoken Words

The speech signal varies across individual talkers due to differences in the shape and length of the vocal tract (Carrell, 1984; Fant, 1973; Summerfield & Haggard, 1973), glottal source functions (Carrell, 1984), positioning and control of articulators (Ladefoged, 1980), and dialect. It has long been known that the same vowel produced by two different talkers may sound the same, yet have strikingly different formant patterns (Peterson & Barney, 1952). Most contemporary theories of speech perception have regarded acoustic differences between talkers as "noise" that must be filtered or transformed in order to reveal the symbolic information in the speech signal (Blandon, Henton, & Pickering, 1984; Disner, 1980; Gerstman, 1968; Summerfield & Haggard, 1973). These theories assume, either explicitly or implicitly, a "talker normalization" process that compensates for talker variability (Joos, 1948).[2] Usually, such arguments have been made from a logical standpoint, with little empirical evidence regarding effects of talker variability on spoken word recognition. This *talker-normalization hypothesis* is consistent with traditional views of speech perception wherein invariances are sought, "redundant" surface forms of stimuli are quickly forgotten, and semantic information is retained in long-term memory.

In fact, there is a great deal of empirical evidence that the surface forms of stimuli are retained over relatively long periods of time. If voice information is "normalized," in the traditional sense, voice-specific information of a spoken message should be lost from memory after a short period of time. However, using a continuous recognition memory task (Shepard & Teghtsoonian, 1961), Craik and Kirsner (1974) found that memory for a spoken word was better when the word was repeated in the same voice as its original presentation than when it was repeated in a different voice. In a second experiment, they found that subjects were able to explicitly identify whether or not the repetition of a spoken word was in the same voice as its original presentation. Both results were found even after a delay of several minutes. When visual stimuli were used in a similar task, Kirsner (1973) found that recognition memory was better when words were presented and repeated in the same physical typography. Subjects were also able to explicitly identify whether or not the repetition was in the same typography as its original presentation. Using a similar procedure, Kirsner and Smith (1974) showed that information about the modality of a word's first presentation, either visual or auditory, was retained in memory over fairly long periods of time. Moreover, Kolers and Ostry (1974) observed a "savings" when subjects reread passages of inverted text that were in the same inverted form as an earlier presentation. This savings in reading time was found even *one year* after the original presentation of the inverted text, even though recognition memory for the content of the passages was at chance (Kolers, 1976). Taken together, these studies suggest that physical forms are not always filtered from sensory input, but remain part of the long-term memory representation. In the realm of speech perception, these findings suggest a *voice-encoding hypothesis* consistent with a nonanalytic view of cognition, wherein episodic traces of spoken words are not necessarily stripped of their surface forms (e.g., Kolers & Smythe, 1984; Jacoby & Brooks, 1984).

Several recent experiments have examined effects of talker variability on perception and memory for spoken words. Martin, Mullennix, Pisoni, and Summers (1989) examined serial recall of words spoken by a single talker and by multiple talkers. They found that only the recall of items in the primacy portion of a spoken word list was reduced by variations in voice; words in early list positions of multiple-talker lists were not recalled as well as words in early positions of single-talker lists. This effect

[2] *Talker variability* refers to differences between talkers. All references to talker variability and voice differences will refer to such between-talker differences. Differences between words produced by the same talker are not considered in this report.

has since been replicated in a number of studies (Goldinger, Pisoni, & Logan, 1991; Logan & Pisoni, 1987). A similar impairment in recall of words in the primacy portion of a list was found for synthetic speech compared to natural speech (Luce, Feustel, & Pisoni, 1983).

Deficits in the recall of words from the initial portion of a list have typically been taken as an indication of a reduced amount or efficiency of short-term memory rehearsal, which has a direct effect on the success of transfer to long-term memory (Atkinson & Shiffrin, 1968; Waugh & Norman, 1965). This reduced efficiency of rehearsal may arise from either of two processes related to talker variability. First, a demand for attentional resources in the initial talker normalization process could result in fewer available resources for subsequent rehearsal (see Mullennix & Pisoni, 1990). Second, the novel voice-specific information encoded in memory with each word might require greater resources for rehearsal of words produced by multiple talkers. Because the central and recency portions of the serial recall curve were not affected by talker variability, there did not appear to be an increase in perceptual encoding errors due to changes in voice from one word to the next.

Goldinger et al. (1991) conducted another experiment that varied the inter-stimulus interval (ISI) between presentation of words in a list to assess how talker variability interacted with rehearsal time in a serial recall task. With short ISIs, the results of Martin et al. (1989) were replicated. However, with longer ISIs, the effect of talker variability was reversed; words in the primacy portions of lists produced by multiple talkers were actually recalled *better* than words in the primacy portion of lists produced by a single talker. The presence of the additional voice cues apparently provided a means for improved recall of items from long-term memory. With shorter ISIs there was apparently not enough time to encode both the word and the voice. However, longer ISIs allowed for increased rehearsal of words produced by multiple talkers, hence providing better encoding of both the word and the voice as recall cues.

The present experiment sought to examine longer term memory of words spoken by multiple talkers by using a continuous recognition memory task (Shepard & Teghtsoonian, 1961). In this task, the subject was presented with a continuous stream of spoken words. Each word was presented twice. Half of the words were presented and repeated in the same voice and half were repeated in a different voice. After each word, the subject responded "new" when hearing a word for the first time, or "old" when hearing a word that was repeated, regardless of whether the voice was the same. The stimulus lists were constructed so that the initial presentation and repetition of each word was separated by a lag of between one and sixty-four intervening words, thus allowing measurements of recognition memory over different lengths of time. This distribution of lags was greater than that used by Craik and Kirsner (1974), who used the same task, but had a maximum lag of thirty-two intervening items.

Given the previous literature, the results of this experiment should show decreased recognition with an increase in lag. For repetitions after short lags, recognition is assumed to be based on words in short-term memory or an acoustic store (e.g., Crowder & Morton, 1969). Recognition is fast and accurate because the words are easily accessed. For repetitions after longer lags, recognition depends on long-term memory encoding. A decrease in recognition is expected with increases in lag because the words will not have all been transferred completely to long-term memory and because the intervening words may interfere with recognition.

More importantly, increased recognition performance is expected for same-voice repetitions, compared to different-voice repetitions, thus replicating the results found by Craik and Kirsner (1974) using two talkers. If word and voice can be considered two dimensions of a single stimuli, then same-voice repetitions match on both the word and voice dimensions, whereas different-voice repetitions match only on the word dimension. A matching voice provides a redundant cue which aids in the recognition of the word.

Whereas Craik and Kirsner (1974) compared recognition performance for words produced by a single talker with words produced by two talkers (a male and a female), the present study examined the effects of talker variability with up to twenty different talkers (ten male and ten female). With only two talkers it is not clear whether the voice, per se, is retained, or if only the gender of the talker is encoded and retained. Geiselman and Bellezza (1976, 1977) proposed a *voice-connotation hypothesis* to account for the retention of voice characteristics. In a series of experiments, Geiselman and colleagues found that subjects were able to recognize sentences as being repeated in the same voice, and argued that the talker's gender modified the semantic interpretation or connotation of the message (Geiselman, 1979; Geiselman & Bellezza, 1976, 1977; Geiselman & Crawley, 1983). In most experiments, encoding of voice was automatic and obligatory.[3] However, by modifying the task to include male and female agents in the sentences, Geiselman (1979) eliminated the obligatory encoding of voice information. Geiselman argued that, when the gender of the agent in the passage conflicted with the gender of the speaker, the agent took precedence and was preferentially encoded. Consistent with a traditional, symbolic view of cognition, Geiselman and colleagues have argued that voice is encoded by modifying the semantic interpretation of a sentence, rather than as information in its own right.

Given Geiselman's claim, there is some question as to whether improved performance on same-voice trials in the two-talker experiment by Craik and Kirsner (1974) indicates that voice information is encoded in memory or that gender information is encoded as connotative information about a word. In fact, Craik and Kirsner (1974) could not definitively state whether literal aspects of the spoken words were retained or whether the gender of the talker was retained by influencing the semantic coding of the word. By increasing the number of talkers, the effects of gender and voice can be distinguished because the different-voice repetitions can be spoken by a speaker of either the same gender or a different gender than the speaker who produced the original word.

With an increase in talker variability, three possible patterns of results could be observed. First, consistent with the talker-normalization hypothesis, overall recognition performance should decrease as talker variability is increased. Voice changes from trial to trial might incur a low-level processing deficit, taking resources away from memory encoding. Increasing the amount of talker variability should increase the need for low-level recalibration or adjustment, thereby causing an overall decrease in recognition performance. Second, consistent with the voice-connotation hypothesis, same-voice repetitions and different-voice repetitions of the same gender should result in similar recognition performance, and both should be better than different-voice repetitions of a different gender. If voice provides connotative information about a word, then a repetition in any voice of the same gender should yield the same semantic interpretation of a given word, thus aiding the recognition of its repetition. A repetition of a word by a speaker of a different gender should yield a different semantic interpretation, thereby attenuating the recognition of its repetition. Third, consistent with the voice-encoding hypothesis, overall recognition performance should not change as talker variability is increased. If voice is encoded directly into long-term memory representations, then voice can serve as an additional cue during recognition, making same-voice repetitions seem more familiar than different-voice repetitions. To understand the prediction of no effect, recall that the subject tries to discriminate repetitions from distractors. Global familiarity with each stimulus, as measured by a summed similarity with items in memory, is used as the basis of this discrimination (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988); familiar stimuli are judged

[3] This was also found by Mullennix and Pisoni, 1990, in a Garner speeded-classification task.

"old" and unfamiliar stimuli are judged "new." If only two voices are used in the stimulus set and voice familiarity contributes to global familiarity, then all words in the set, targets *and* distractors, will seem familiar to subjects. If twenty voices are used, all words in the set, targets *and* distractors, will seem unfamiliar, or distinctive, to subjects. In either case, the difficulty of the old/new discrimination is the same; changing the number of voices in the set should modify the familiarity or distinctiveness of both targets and distractors equivalently.

Method

Subjects

Subjects were two hundred undergraduate students from introductory psychology courses at Indiana University. Forty subjects served in each condition. Each subject participated for one half-hour session and received partial credit for an introductory psychology course. All subjects were native speakers of English who reported no history of a speech or hearing disorder at the time of testing.

Stimuli

The stimuli were lists of words spoken by either a single talker or a set of multiple talkers. The stimuli were monosyllabic words selected from the vocabulary of the Modified Rhyme Test (MRT) (House, Williams, Hecker, & Kryter, 1965). Each word was recorded in isolation on audiotape and digitized via a 12-bit analog-to-digital converter using a PDP 11/34 computer. The overall root mean squared (RMS) amplitude levels for each word were digitally equated. A database containing all 300 MRT words spoken by 20 different talkers was used to generate the stimulus lists.

The word lists were constructed such that each word was presented and repeated once. The repetition of any given word occurred after a lag of 1, 2, 4, 8, 16, 32, or 64 intervening words; the repetition counted as one of the intervening words. Each lag value was used an equal number of times in each list. At every position, except at the very beginning of the list, the probability of a repeated word was .5. One hundred-forty old-new test pairs were presented in each list. Unknown to the subject, the first thirty words of each list were used to establish a memory load and were not considered in the data analyses; none of these words were repeated in the test portion of the list. Twenty "filler" words were also distributed throughout the test portion of each list to simplify the list generation process; these words were never repeated and were not considered in the data analyses. Each subject was also given an initial practice list of 15 words to become familiarized with the task; none of these words were repeated in the experiment. The 30 initial words, 140 test pairs, 20 "filler" words, and 15 practice words provided a total of 345 spoken words in each session.

Talker variability was manipulated by selecting a subset of stimuli from a database of twenty different talkers, ten male and ten female, each speaking the entire set of 300 MRT words. Single-talker lists for each session were generated by randomly selecting one of the twenty talkers as the source of the words. Multiple-talker lists of 2, 6, 12, and 20 talkers were produced by randomly selecting an equal number of males and females from the pool of twenty talkers. On the initial presentation of a word, one of the available talkers in this set was selected at random. On the repetition of a word, the probability that the same talker repeated the word was equal to the probability that a randomly chosen different talker repeated the word. For the different-voice repetitions, the number of male-male, female-female, female-male, and male-female pairings were balanced.

Procedure

Subjects were tested in groups of five or fewer in a room equipped with sound-attenuated booths used for speech perception experiments. Stimulus presentation and data collection were controlled on-line by a PDP-11/34 minicomputer. Each stimulus word was low-pass filtered at 4.8 kHz and played to listeners binaurally through a 12-bit digital-to-analog converter over matched and calibrated TDH-39 headphones at 80dB (SPL). After hearing each word, the subject was given a maximum of 5-s to respond by pressing a button labeled "new" if the word was judged new, or a button labeled "old" if the word was judged old. There was a 1-s delay between response and presentation of the next word. Subjects rested one finger from each hand on the two response buttons. Subjects were asked to respond as quickly and as accurately as they could. Button selections and response times were recorded.

The lag between the initial presentation and the repetition of a word, and the voice of the repetitions (same vs. different voice) were within-subjects variables. The total number of voices in the set was a between-subjects variable. Each group listened to a different list of stimulus words.

Results

In the following analyses, a hit was defined as responding "old" to an old word, a miss as responding "new" to an old word, a false alarm as responding "old" to a new word, and a correct rejection as responding "new" to a new word. Unless otherwise stated, all reported response time data are for hits.

Insert Figure 1 about here

Hit Rates

Hit rates were calculated for all subjects in all conditions. Figure 1 displays the hit rates from all of the multiple-talker conditions. The upper panel displays the hit rates for same- and different-voice repetitions as a function of talker variability, collapsed across values of lag; the lower panel displays the hit rates for same- and different-voice repetitions as a function of lag, collapsed across levels of talker variability. A 4 X 7 X 2 analysis of variance (ANOVA) was conducted on the hit rate data for the between-subjects variable of talker variability (Variability), and the within-subjects variables of lag (Lag) and same- and different-voice repetitions (Voice). No significant main effect of Variability was observed [$F(3,156)=0.53, p=.66$]. Increasing the amount of talker variability on the lists had no significant effect on hit rates or on false alarms, as described below. This null finding can be seen by the parallel lines in the upper panel of Figure 1. A significant main effect of Voice was observed [$F(1,156)=186.94$, $p<.0001$], as revealed by the separation of lines in both panels of Figure 1. Same-voice repetitions yielded higher hit rates than different-voice repetitions at all levels of talker variability and at all lag values. A significant main effect of Lag was observed [$F(6,936)=196.16, p<.0001$]. Increasing the lag caused a decrease in the hit rate, as shown by the negative slope of the lines in the lower panel of Figure 1. There was no significant two-way interaction of Voice X Lag in this analysis [$F(6,936)=1.37$, $p=.22$].

In addition to the main effects, two significant interactions were observed. The two-way Variability X Lag interaction was significant [$F(18,936)=1.86, p<.05$]. Tukey's HSD analyses revealed that this interaction was due to a higher hit rate at a lag of thirty-two with twelve or twenty voices than with two or six voices.[4] A significant three-way Variability X Lag X Voice interaction was also observed [$F(18,936)=2.14$, $p<.01$]. Although statistically significant effects, the magnitude of these interactions were small, compared to the main effects of Lag and Voice. They each accounted for less

[4] All post-hoc comparisons reported in this paper are $p<.05$ or beyond.

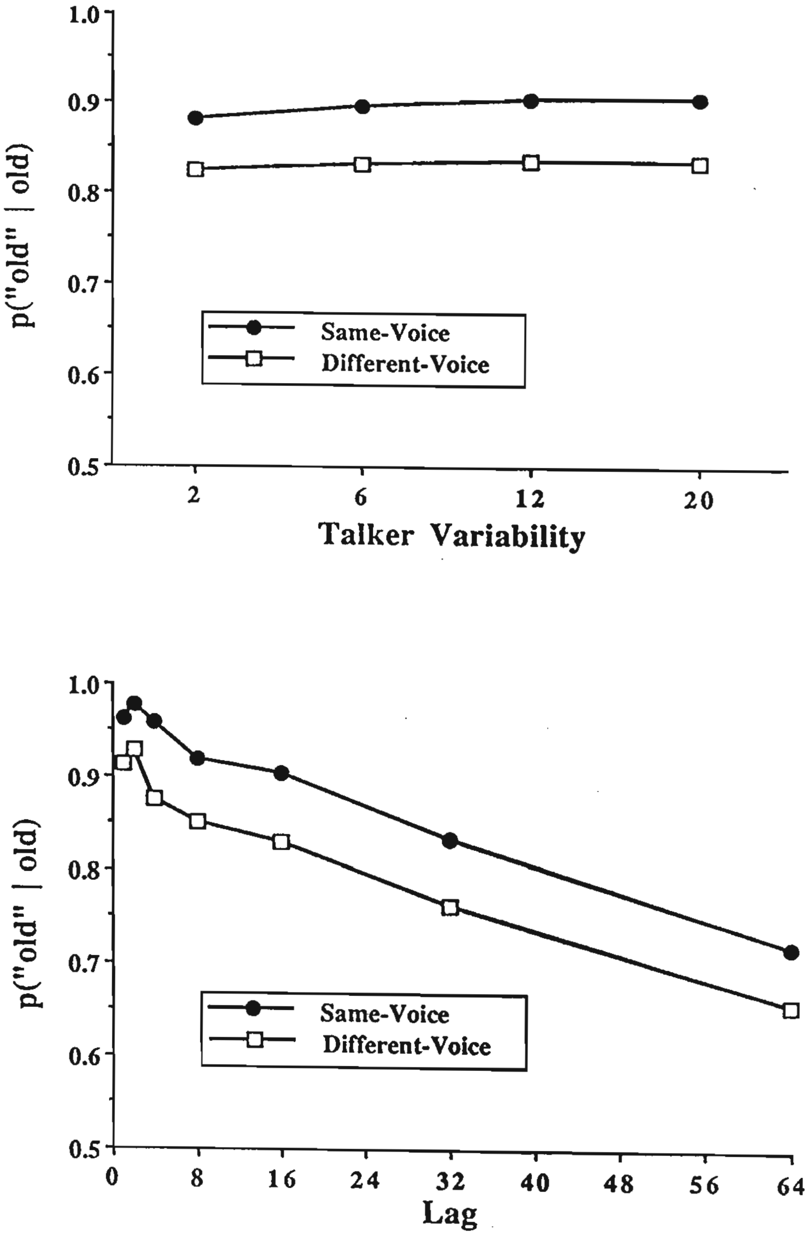


Figure 1. Hit rates [p("old" | old)] from all of the multiple-talker conditions. The upper panel displays the hit rates for same- and different-voice repetitions as a function of talker variability, collapsed across values of lag; the lower panel displays the hit rates for same- and different-voice repetitions as a function of lag, collapsed across levels of talker variability.

than 1% of the variance, compared to 55.8% and 7.8% of the variance accounted for by Lag and Voice, respectively.

Insert Figure 2 about here

Figure 2 displays the hit rates from the single-talker condition and the same-voice repetitions of the multiple-talker conditions. The upper panel displays the hit rates as a function of talker variability, collapsed across values of lag. The lower panel compares the hit rates for the single-talker condition to the average hit rates for the same-voice repetitions of the multiple-talker conditions as a function of lag, collapsed across levels of talker variability. A 5 X 7 (Variability X Lag) ANOVA was conducted on the hit rates. As in the analysis of the multiple-talker conditions alone, no significant main effect of Variability was observed [$F(4,195)=0.82, p=.51$]. The hit rates in the single-talker condition were not significantly different from the hit rates in the multiple-talker conditions. A significant main effect of Lag was obtained [$F(6,1770)=176.24, p<.0001$], as reflected by the negative slopes in the lower panel of Figure 2. The two-way Variability X Lag interaction was not significant [$F(24,1170)=1.24, p=.19$].

Insert Figure 3 about here

Figure 3 displays the hit rates from the six-, twelve-, and twenty-talker conditions. In both panels of Figure 3, the same-voice repetitions are compared with different-voice/same-gender and different-voice/different-gender repetitions. Only the six-, twelve-, and twenty- talker conditions were included because only these had both same and different gender for different-voice repetitions. The upper panel displays the hit rates as a function of talker variability, collapsed across values of lag. The lower panel displays the hit rates as a function of lag, collapsed across levels of talker variability. A 3 X 7 X 3 (Variability X Lag X Voice) ANOVA was conducted on the hit rate data. The three Voice conditions in the analysis were same-voice, different-voice/same-gender, and different-voice/different-gender. No significant main effect of Variability was observed [$F(2,117)=0.09, p=.91$]. As in the previous analyses, increasing the level of talker variability had no effect on hit rates. A significant main effect of Voice was obtained [$F(2,234)=53.64, p<.0001$]. Tukey's HSD analyses revealed that the hit rates for different-voice/same-gender and different-voice/different-gender repetitions did not significantly differ, although both were significantly lower than the hit rates for same-voice repetitions, as shown in both panels of Figure 3. A significant main effect of Lag was observed [$F(6,702)=117.20, p<.0001$], as shown by the negative slopes in the lower panel of Figure 3. There was, however, no significant two-way Voice X Lag interaction [$F(12,1404)=1.50, p=.12$].

Insert Figure 4 about here

Response Time

Response time to hits were calculated for all subjects in all conditions. Figure 4 displays the response times from all of the multiple-talker conditions. The upper panel displays the response times for same- and different-voice repetitions as a function of talker variability, collapsed across values of lag. The lower panel displays the response times for same- and different-voice repetitions as a function of lag, collapsed across levels of talker variability. A 4 X 7 X 2 (Variability X Lag X Voice) ANOVA was conducted on the response times. No significant main effect of Variability was observed [$F(3,156)=0.55, p=.65$]. Increasing the amount of talker variability of the lists had no significant effect on response times. This null finding is shown by the parallel lines in the upper panel of Figure 4.

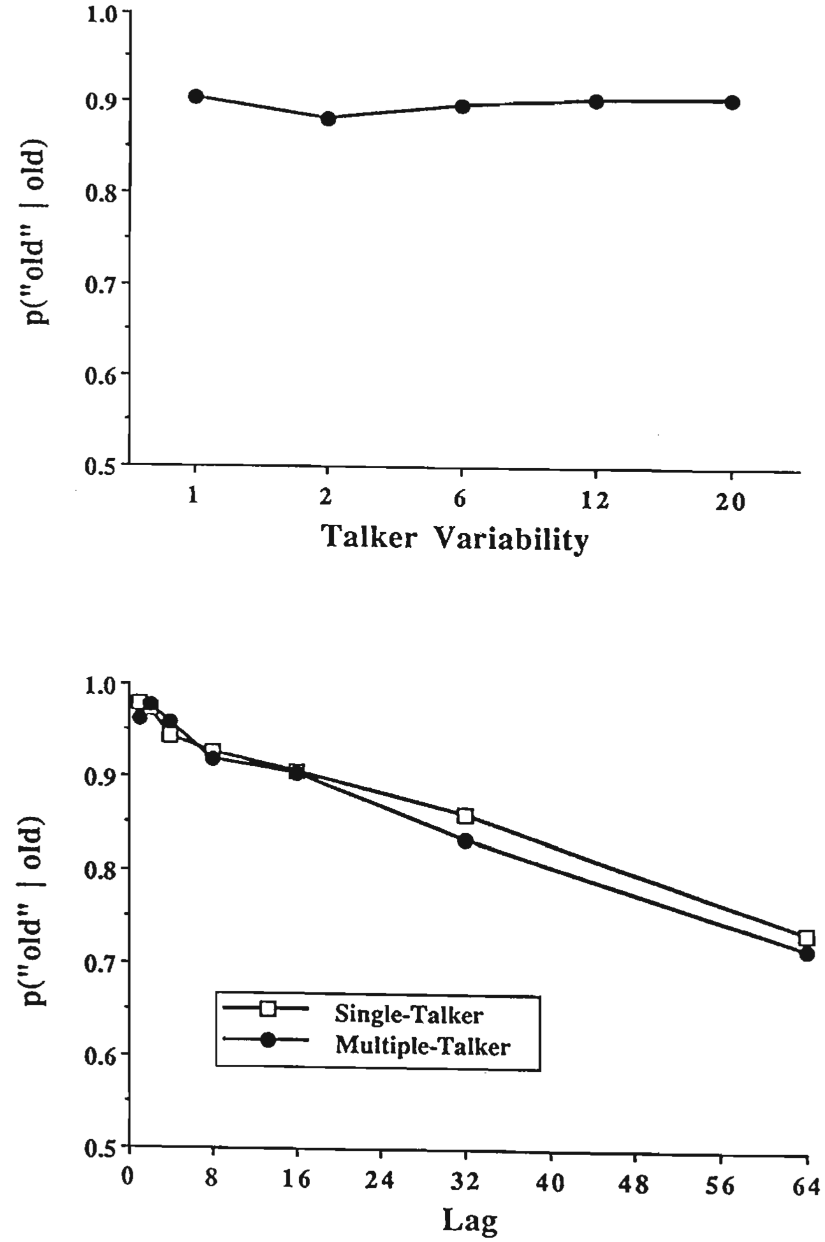


Figure 2. Hit rates [p("old" | old)] from the single-talker condition and the same-voice repetitions of the multiple-talker conditions. The upper panel displays the hit rates as a function of talker variability, collapsed across values of lag; the lower panel compares the hit rates for the single-talker condition to the average hit rates for the same-voice repetitions of the multiple-talker conditions as a function of lag.

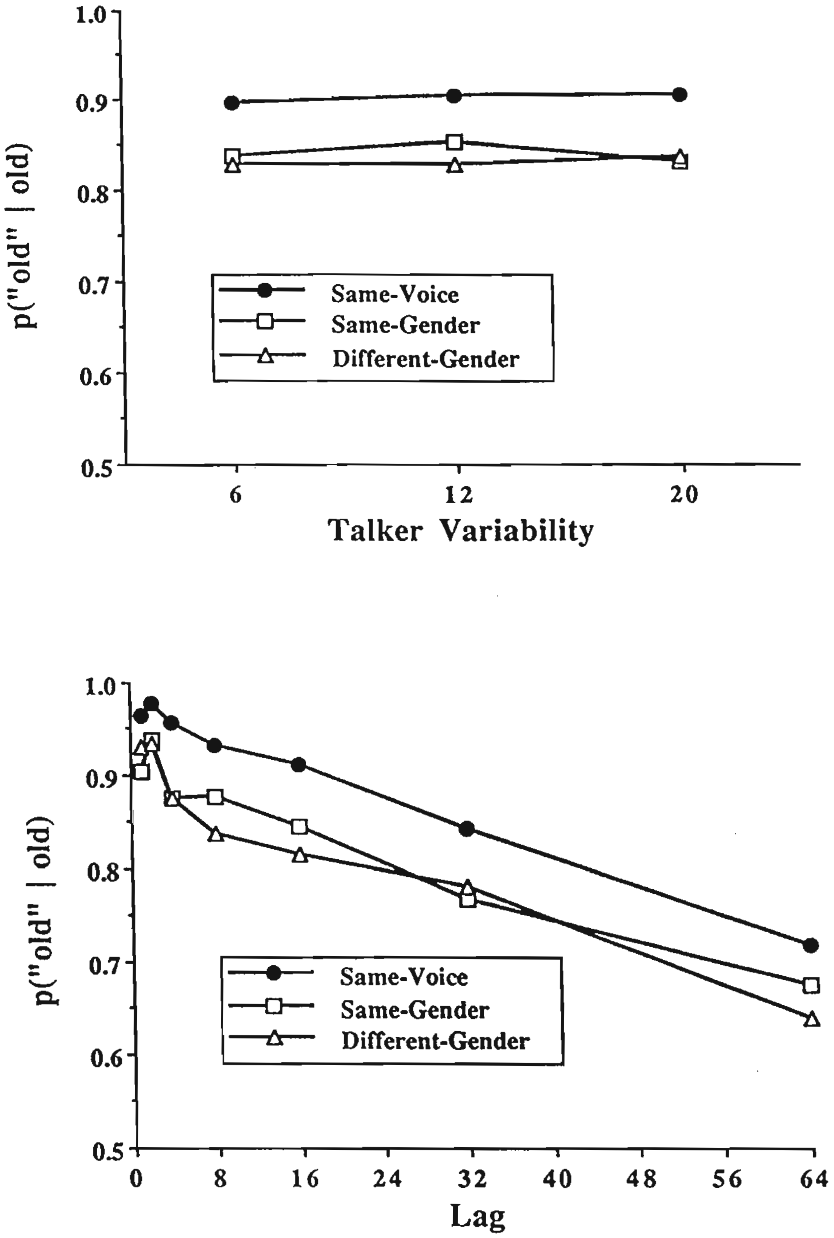


Figure 3. Hit rates [p("old" | old)] from the six-, twelve-, and twenty-talker conditions. In both panels, the same-voice repetitions are compared with different-voice/same-gender and different-voice/different-gender repetitions. The upper panel displays the hit rates as a function of talker variability, collapsed across values of lag; the lower panel displays the hit rates as a function of lag, collapsed across levels of talker variability.

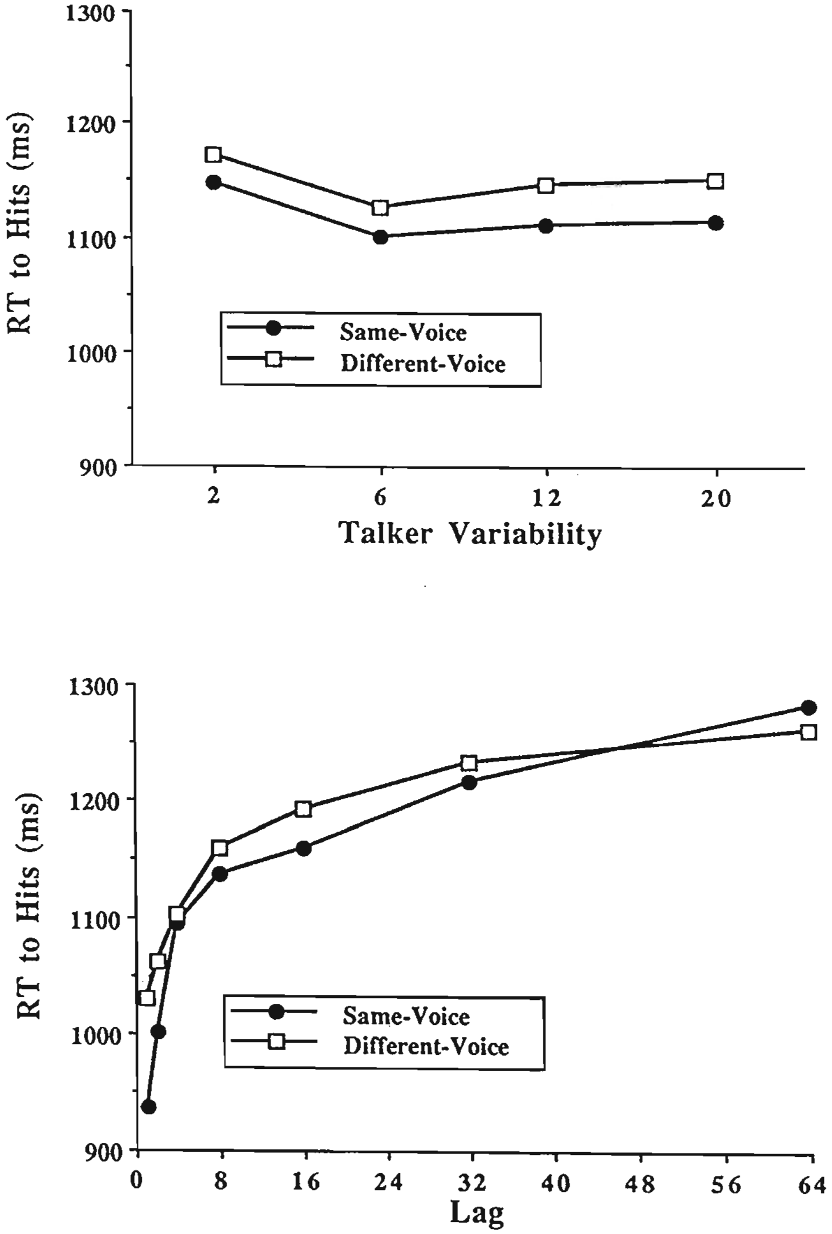


Figure 4. Response times (ms) to hits from all of the multiple-talker conditions. The upper panel displays the response times for same- and different-voice repetitions as a function of talker variability, collapsed across values of lag; the lower panel displays the response times for same- and different-voice repetitions as a function of lag, collapsed across levels of talker variability.

A significant main effect of Voice was observed [*F*(1,156)=30.00, *p*<.0001], as revealed by the separation of lines in both panels of Figure 4. Same-voice repetitions yielded faster response times than different-voice repetitions at all levels of talker variability. A significant main effect of Lag was observed [*F*(6,936)=161.20, *p*<.0001]. Increasing the lag yielded an increase in the response time, as shown by the monotonic increase of the curves in the lower panel of Figure 4. In addition, a significant Lag X Voice interaction was observed [*F*(6.936)=7.08, *p*<.0001]. Tukey's HSD analyses revealed that response times to hits were significantly faster for same-voice repetitions than different-voice repetitions only at lags of one or two intervening items. However, responses to same-voice repetitions tended to be faster than different-voice repetitions at all values of lag except sixty four, as is evident by the separation of the curves in the lower panel of Figure 4.

A significant three-way Variability X Lag X Voice interaction was also observed [*F*(18,936)=2.09, *p*<.01]. As with the comparable three-way interaction found in the hit rate data, this was a statistically significant effect, however, the magnitude was small compared to the main effects of Lag and Voice. The interaction accounted for less than 1% of the variance compared to 53.5% of the variance accounted for by Lag. Voice and the Voice X Lag interaction accounted for 1.3% and 1.5% of the variance, respectively.

Insert Figure 5 about here

Figure 5 displays the response times from the single-talker condition and the same-voice repetitions of the multiple-talker conditions. The upper panel displays the response times as a function of talker variability, collapsed across values of lag; the lower panel compares the response times for the single-talker condition to the average response times for the same-voice repetitions of the multiple-talker conditions as a function of lag. A 5 X 7 (Variability X Lag) ANOVA was conducted on the response times. The main effect of Variability only approached significance [*F*(4,195)=2.34, *p*=.06]. Tukey's HSD analyses revealed that the response times to hits were significantly faster for the single-talker condition than the two-talker condition, but were not significantly faster than the six-, twelve-, or twenty-talker condition. This effect is evident from the faster response times for the single-talker condition compared to the multiple-talker conditions in the upper panel of Figure 5. A significant main effect of Lag was obtained [*F*(6,1770)=181.40, *p*<.0001], as reflected by the monotonically increasing curves in Figure 5. A significant two-way Variability X Lag interaction was observed [*F*(24,1170)=1.63, *p*<.05].

Insert Figure 6 about here

Figure 6 displays the response times from the six-, twelve-, and twenty-talker conditions. In both panels of Figure 6, the same-voice repetitions are compared with different-voice/same-gender and different-voice/different-gender repetitions. Only the six-, twelve-, and twenty- talker conditions were included because only these had both same and different gender for different-voice repetitions. The upper panel displays the response times as a function of talker variability, collapsed across values of lag; the lower panel displays the response times as a function of lag, collapsed across levels of talker variability. A 3 X 7 X 3 (Variability X Lag X Voice) ANOVA was conducted on the response times. The three Voice conditions in the analysis were: same-voice, different-voice/same-gender, and different-voice/different-gender. No significant main effect of Variability was observed [*F*(2,117)=0.14, p=.87]. Increasing the level of talker variability had no effect on response times, as in the previous analyses. A significant main effect of Voice was observed [*F*(2,234)=4.07, *p*<.05]. Tukey's HSD analyses revealed that the response times for different-voice/same-gender and different-voice/different-gender repetitions did not significantly differ, although both were significantly

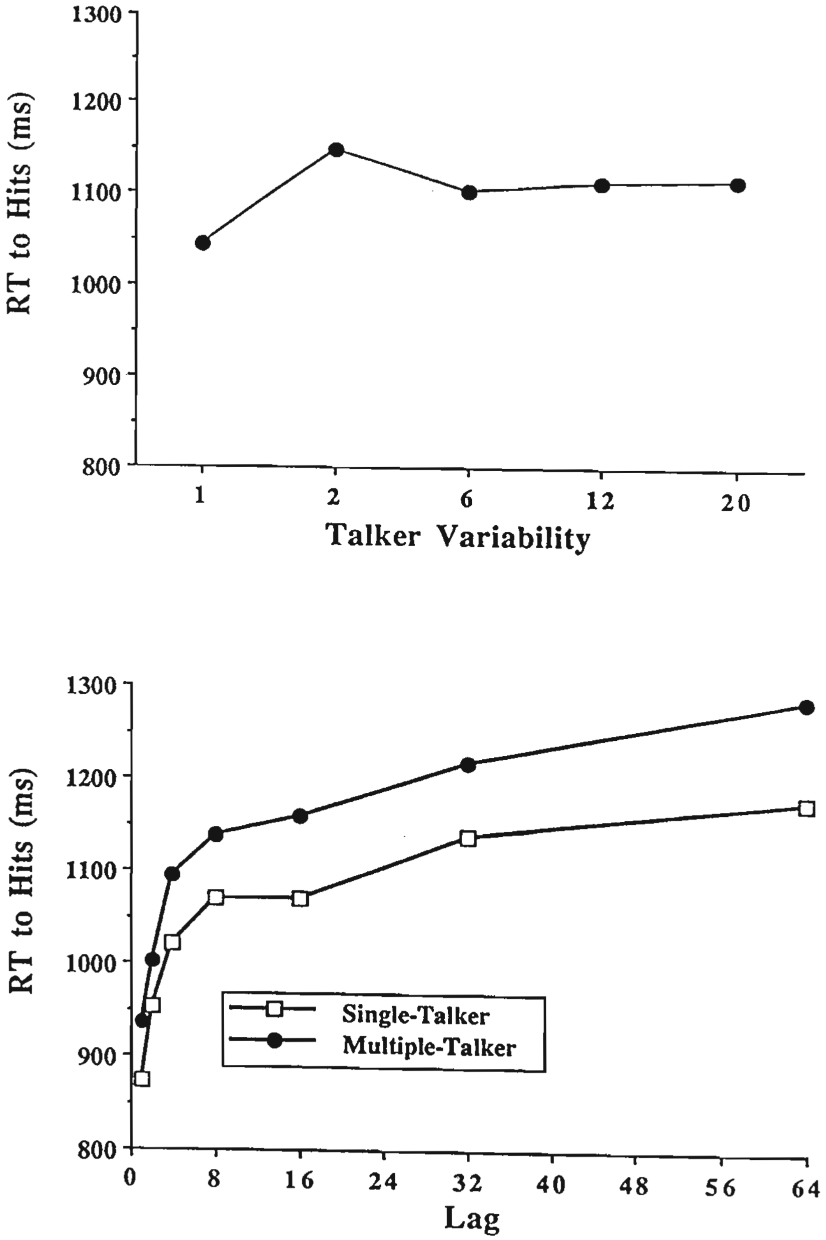


Figure 5. Response times (ms) to hits from the single-talker condition and the same-voice repetitions of the multiple-talker conditions. The upper panel displays the response times as a function of talker variability, collapsed across values of lag; the lower panel compares the response times for the single-talker condition to the average response times for the same-voice repetitions of the multiple-talker conditions as a function of lag.

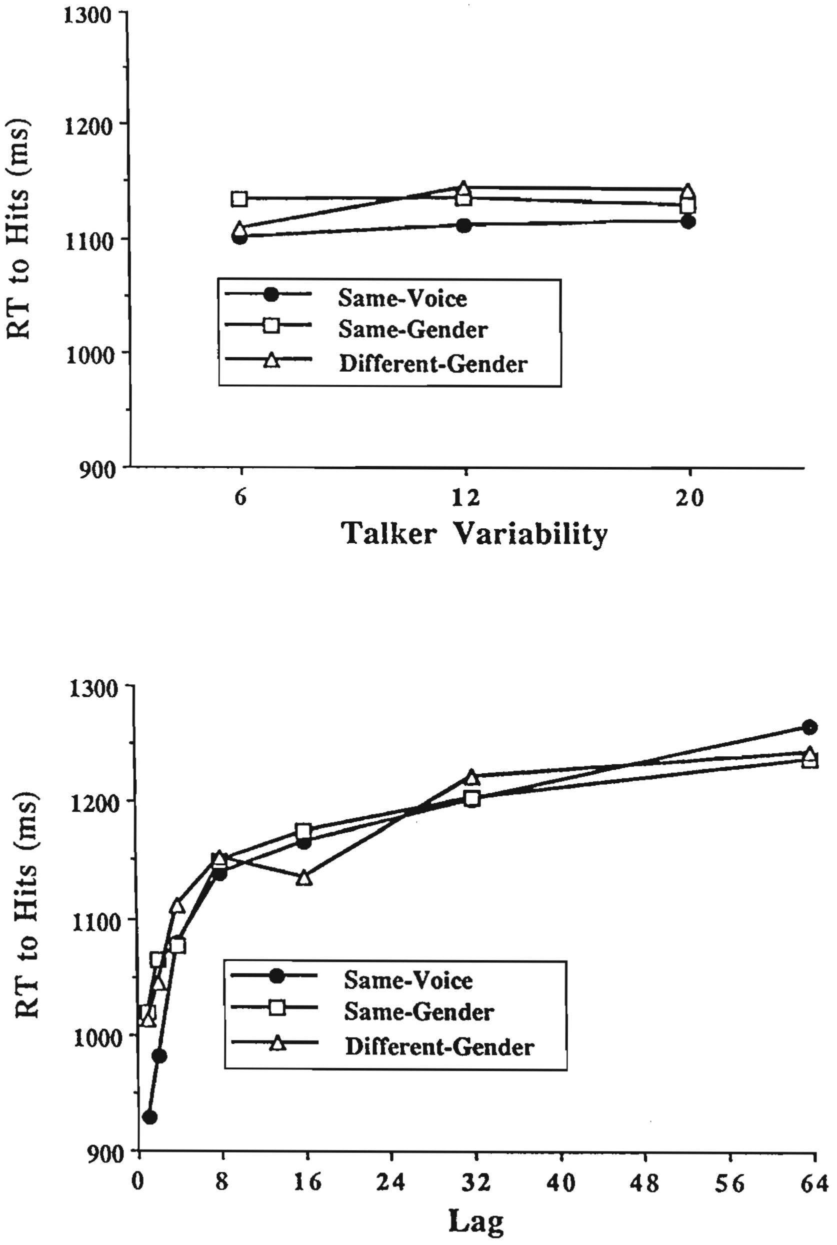


Figure 6. Response times (ms) to hits from the six-, twelve-, and twenty-talker conditions. In both panels, the same-voice repetitions are compared with different-voice/same-gender and different-voice/different-gender repetitions. The upper panel displays the response times as a function of talker variability, collapsed across values of lag; the lower panel displays the response times as a function of lag, collapsed across levels of talker variability.

slower than the response times for same-voice repetitions, as shown in both panels of Figure 6. A significant main effect of Lag was observed [$F(6,702)=80.02$, $p<.0001$], as shown by monotonically increasing curves in the lower panel of Figure 6. There was also a significant two-way Voice X Lag interaction [$F(12,1404)=2.39$, $p<.01$].

Insert Figure 7 about here

Comparison of Hits and False Alarms

The upper panel of Figure 7 displays the hits and false alarms as a function of talker variability. The lower panel displays the response times for hits and false alarms as a function of talker variability. The hit rates are collapsed across Lag and Voice in both panels. Comparison of hits and false alarms provides an assessment of overall recognition performance. Because false alarms are "old" responses to new words, they cannot be analyzed in terms of lag, but only in terms of talker variability. The hit rates were analyzed previously. A one-way ANOVA was conducted on the false alarm rates over levels of talker variability. No significant main effect of Variability was observed [$F(3,156)=2.18$, $p=.09$]. An ANOVA including the single-talker condition was also conducted. Again, no significant main effect of Variability was observed [$F(4,195)=1.87$, $p=.12$]. A one-way ANOVA was conducted on the response time to false alarms over talker variability. No effect of Variability was observed [$F(3,156)=0.21$, $p=.89$]. An ANOVA including the single talker condition was also conducted. No effect of Variability was observed [$F(4,195)=1.26$, $p=.29$].

Discussion

The present experiment was designed to study the effects of voice on recognition memory for spoken words. Several major findings stand out. First, we replicated the earlier findings of Craik and Kirsner (1974). Increasing the lag between the initial presentation of a spoken word and its repetition decreased accuracy and increased response time, regardless of the voice of the repetition. This basic finding has been reported by others using the continuous recognition memory task (Craik & Kirsner, 1974; Hockley 1982; Kirsner, 1973; Kirsner & Smith, 1974; Shepard & Teghtsoonian, 1961). The explanation for the decrease in performance over time depends on the model of memory one chooses to employ. Such results have been attributed to decay of activation of nodes in memory over time (Anderson, 1983), increased noise due to shared memory vectors which get convolved with later items (Eich, 1985; Murdock, 1982), or changes in context over the course of the experiment (Gillund & Shiffrin, 1984). All three of the hypotheses discussed in the introduction are compatible with this finding.

Second, we replicated Craik and Kirsner's (1974) finding that same-voice repetitions resulted in higher accuracy and faster responses than different-voice repetitions at all values of lag, and at all levels of talker variability. For very short lags, one could argue that the advantage for same-voice repetitions is due to some form of short-term acoustic storage (Crowder & Morton, 1969). However, the advantage for same-voice repetitions at longer lags suggests that the talker's voice is encoded as an integral part of the long-term memory representation of a spoken word. This is a problem for any talker-normalization hypothesis which assumes that voice information is "stripped away" from the listener's representation of the speech signal and only lexical information is retained in memory. According to most models of memory (e.g., Gillund & Shiffrin, 1984), the presence of additional cues improves recognition performance. If voice serves as an additional cue, same-voice repetitions can be retrieved using the word cue as well as the voice cue, whereas different-voice repetitions can only be retrieved with the word cue. This prediction is consistent with the voice-encoding hypothesis and was reflected in our data. This result, however, can also be explained by the voice-connotation hypothesis, since the gender of the talker

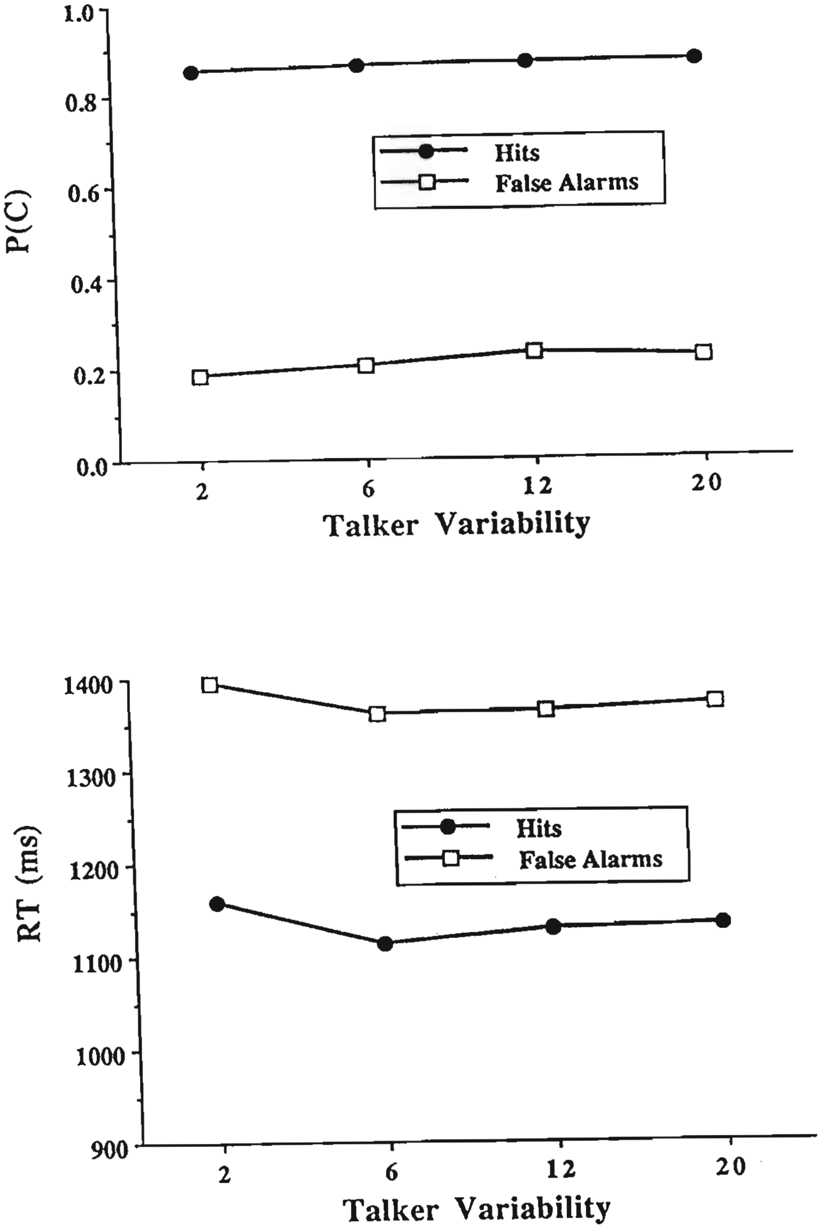


Figure 7. The upper panel displays the hits and false alarms as a function of talker variability; the lower panel displays the response times for hits and false alarms as a function of talker variability. The hit rates are collapsed across Lag and Voice in both panels.

can provide a similar connotation for a word repeated by the same talker or for a word repeated by a different talker of the same gender.

The present findings demonstrate that some aspect of the talker's voice is retained in long-term memory and is used as a cue during recognition. The question is, what aspects of the talker's voice are encoded? Geiselman and Belezza (1976, 1977) would argue that the gender of the talker influences the connotation of a given word. If this were the case, then recognition of different-voice repetitions should have been greater for words spoken by talkers of the same gender as the talker of the original word, because a similar connotation would be given. Analysis of different-voice repetitions revealed no significant differences between words repeated in a different voice of the same gender or different gender; the voice-connotation hypothesis was not supported. It appears that something more detailed than a connotative influence of gender is retained in memory. Perhaps either some physical representation of the spoken word or some record of the operations used to decode the stimuli into a lexical item (e.g., Kolers, 1976) would better characterize episodic traces of spoken words.

Third, no difference was found in recognition performance with an increase in talker variability. This result may be difficult to explain with the talker-normalization hypothesis. A talker-normalization hypothesis would predict that increasing the number of talkers would decrease recognition performance because the normalization process must be constantly applied. However, it may be that the uncertainty introduced by *any* of the multiple-talker conditions forces the listener to allocate resources towards normalization, even if two consecutive words are spoken by the same talker. Thus, the finding of no differences between the two talker and the twenty talker conditions may not reject the talker normalization hypothesis. The results from the single-talker condition provide no clear evidence on this matter. The accuracy of recognition in the single-talker condition was the same as the same-voice trials of the multiple-talker conditions, although the response times were faster in the single-talker condition. All of the multiple-talker conditions were equivalent and all were slower than the single-talker condition. Thus, it appears that any amount of talker variability slows down responses but does not affect accuracy.

The absence of any effects due to talker variability is consistent with a voice-encoding hypothesis. If voice information is encoded into long-term memory representations, then the voice can be used as an additional retrieval cue during recognition. A familiarity-based model of recognition is assumed (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988). Words spoken in a set with low variability are similar to many previous items because of numerous voice matches, thus increasing the overall level of familiarity. Words spoken in a set with high variability are similar to few previous items because of few voice matches, thus decreasing the overall level of familiarity. Since any increase or decrease in familiarity is equivalent for targets *and* distractors, there is no net change in recognition performance. Subjects responding in an optimal way should adjust their criterion so that both hit rates and false alarm rates are the same, regardless of the level of talker variability. The results show that both hit rates and false alarm rates are constant over increases in talker variability, suggesting that voices are retrieval cues just as words are retrieval cues.

Most experiments which have examined the effects of talker variability on spoken word recognition have compared a single-talker condition with one multiple-talker condition (Craik & Kirsner, 1974; Goldinger et al., 1991; Logan & Pisoni, 1987; Martin et al., 1989; Mullennix, Pisoni, & Martin, 1988). In these experiments, different degrees of talker variability, either in number of voices or perceptual similarity of voices, are not employed. With the exception of Creelman (1957), who found a small decrease in perceptual identification accuracy with an increase in talker variability, and Mullennix and Pisoni (1990), who found that increasing talker variability slowed down classification in a Garner task, there have been no experiments that systematically varied the level of talker variability. Additional experiments are currently underway to determine how the degree of talker variability affects various aspects of spoken word recognition and memory.

The result that same-voice repetitions are recognized better and more accurately than different-voice repetitions, at all values of lag, suggests that a talker normalization process that removes talker-specific information from representations is not plausible. Indeed, the evidence from this experiment, as well as others (e.g., Goldinger et al., 1991), suggests that detailed information about a talker's voice is encoded along with words into long-term memory, and that this information persists for at least several minutes (Allard & Henderson, 1976; Cole, Coltheart, & Allard, 1974). In addition, the null effect of increasing talker variability is consistent with the hypothesis that voice attributes can be used as retrieval cues in much the same way as words are. There is some disagreement as to whether a talker normalization mechanism is even necessary in a theory of speech perception (e.g., Strange, Verbrugge, Shankweiler, & Edman, 1976; Verbrugge, Strange, Shankweiler, & Edman, 1976). If, however, there is a low-level talker normalization mechanism, as suggested by many researchers (Disner, 1980; Gerstman, 1968; Summerfield & Haggard, 1973) and if it is correctly implied by recent perceptual experiments (e.g., Mullennix & Pisoni, 1990), then it must be a weak form of normalization in which linguistic information *extracted* from the speech signal is encoded along with the physical form of the utterance. It cannot be a strong form of normalization, which the term "normalization" often implies, in which talker variability is removed from the speech signal or the speech signal is rescaled, and only linguistic information remains.

References

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In Spence, K.W., & Spence, J.T. (Eds.), *The Psychology of Learning and Motivation, Volume 2*. New York: Academic Press, 89-105.

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, IN: Speech Research Laboratory, Indiana University.

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

Crowder, R.G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception and Psychophysics*, **5**, 365-373.

Disner, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, **67**, 253-261.

Eich, J.M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review*, **92**, 1-38.

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

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

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

Geiselman, R.E., & Bellezza, F.S. (1977). Incidental retention of speaker's voice. *Memory and 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.

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

Gillund, G., & Shiffrin, R.M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.

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

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

Hockley, W.E. (1982). Retrieval processes in continuous recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **8**, 497-512.

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.

Jacoby, L.L, & Brooks, L.R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In Bower, G.H. (Ed.), *The Psychology of Learning and Motivation, Volume 18*. New York: Academic Press, 1-47.

Joos, M.A. (1948). Acoustic phonetics. *Language*, Supplement 24, 1-136.
Kirsner, K. (1973). An analysis of the visual component in recognition memory for verbal stimuli. *Memory and Cognition*, **1**, 449-453.

Kirsner, K., & Smith, M.C. (1974). Modality effects in word identification. *Memory and Cognition*, **2**, 637-640.

Kolers, P.A. (1976). Reading a year later. *Journal of Experimental Psychology: Human Learning and Memory*, **2**, 554-565.

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.

Kolers, P.A., & Smythe, W.E. (1984). Symbol manipulation: Alternatives to the computational view of mind. *Journal of Verbal Learning and Verbal Behavior*, **23**, 289-314.

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

Logan, J.S., & Pisoni, D.B. (1987). Talker variability and the recall of spoken word lists: A replication and extension. *Research on Speech Perception Progress Report No. 13*. Bloomington, IN: Speech Research Laboratory, Indiana University.

Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25**, 17-32.

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 and Psychophysics*, **47**, 379-390.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1988). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Murdock, B.B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, **89**, 609-626.

Peterson, G.E., & Barney, H.L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

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

Strange, W., Verbrugge, R.R., Shankweiler, D.P., & Edman, T.R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, **60**, 213-224.

Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report of Speech Research in Progress, Volume 2*. Belfast, North Ireland: Queen's University, 1-12.

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.

Waugh, N.C., & Norman, D.A. (1965). Primary memory. *Psychological Review*, **72**, 89-104.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Comments on Talker Normalization in Speech Perception[1]

David B. Pisoni

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

[1] To appear in Yoh'ichi Tohkura (Ed.), *Speech Perception, Production and Linguistic Structure*. Tokyo: OHM Publishing Co. Ltd. This research has been supported by NIH Research Grant #DC-00111-14 to Indiana University.

Comments on Talker Normalization in Speech Perception

For many years, researchers working in the field of speech perception have assumed that the units derived from perceptual analysis of the speech signal are isomorphic with the abstract idealized units postulated in linguistic theory. In his well-known review chapter, Studdert-Kennedy (1974) makes this point quite explicitly:

> The signal is a more or less continuously varying acoustic wave,.... The message is a string of lexical and grammatical items that may be transcribed as an appropriately marked sequence of discrete phonemic symbols....

Central to this approach has always been the assumption of some kind of abstraction process which takes the time-varying physical signal and converts it into a symbolic representation that is equivalent to a linear sequence of phonemes. Until recently, few researchers working on human speech perception had any reason to question this approach. Most theoretical accounts of speech perception have postulated some type of perceptual normalization process to compensate for the many sources of variability in the speech signal. Unfortunately, it is precisely these sources of variability in speech that have prevented researchers from making much progress in solving what is often called the "primary recognition problem," that is, the problem of mapping invariant attributes of the physical signal onto abstract linguistic units.

According to traditional accounts of speech perception, the process of perceptual normalization involves a substantial reduction in information and transformation of the signal into a discrete symbolic representation. Thus, physically different tokens of the same word are made equivalent by removing irrelevant variability or "noise" from the speech signal. In the case of talker normalization, the topic of this session at the ATR workshop, it has been commonly assumed that the speech signal is "stripped" of its source characteristics. Detailed information about the talker is excluded from the symbolic representation. Only the linguistically distinctive features are preserved in memory in an idealized form.

Research from our laboratory over the last few years on the role of talker variability in speech perception demonstrates that this particular source of stimulus variability should not be thought of as just noise in the speech signal. Several of our experiments demonstrate that source characteristics -- detailed information about the talker -- become an integral part of the perceptual record and are encoded into long-term memory along with the more abstract symbolic representation derived from phonetic analysis. From these results, it appears that the representation of speech in memory is actually much richer and more detailed than is necessary for the linguist's description of the speech signal as an idealized sequence of meaningful units. It is possible that these more elaborate representations may provide important new clues to the pattern analyzing operations that human listeners use to map speech signals onto lexical representations in long-term memory. Thus, rather than thinking of stimulus variability as something that needs to be filtered out of the speech signal to extract the idealized linguistic message, stimulus variability may actually provide very useful information to the listener about aspects of the speech that are used for its perceptual analysis and subsequent encoding into memory.

In this commentary on talker normalization, I will briefly review two early studies that examined the role of talker variability in speech perception. Then, I will report on several recent findings from my research group at Indiana. More details about these experiments are provided in my ICSLP-90 plenary lecture (Pisoni, 1990).

Perceptual Experiments on Talker Variability

One of the earliest studies on the effects of talker variability in speech perception was reported by Peters in 1955. He studied the relative intelligibility of single-voice and multiple-voice messages in noise. The hypothesis under test was that continuity of voice during a transmission will improve listener reception. The results supported his hypothesis. Single-voice transmissions were consistently more intelligible than multiple-voice transmissions. Moreover, multiple-voice transmissions also tended to be more adversely affected by increasing levels of noise than single-voice transmissions. Peters speculated that continuity of voice during a message transmission contributes to listener efficiency through the focusing of selective attention and continuing adaptation to the same speaker. Thus, there appears to be some "fine tuning" and long-term calibration to the voice of a single talker that is retained in a listener's memory for some period of time and subsequently used in perceptual analysis. We do not know very much about the nature of this fine-tuning or adjustment process, but it is clearly a robust phenomenon in speech perception.

Another study was reported by Creelman in 1957. Using PB words, he collected speech intelligibility data from a group of listeners who heard two types of audio tapes. In one condition, all the test words were spoken by a single talker; in the other condition, the test words were spoken by either two, four, eight or sixteen talkers. The test items were mixed with noise and presented at three speech-to-noise ratios. Like Peters, Creelman found a tendency for identification performance to decrease as the number of talkers increased. The average score for individual talkers under all conditions was about 7 percent greater than the average score when two or more talkers were used. Both Peters and Creelman touched on the interesting problem of talker variability in speech perception in the 1950's but, as far as I know, neither of them ever continued this line of research any further.

Several years ago we became interested in the effect of talker variability on perception of isolated words. Almost all of the theoretical work in the field of speech perception has been based on experiments that used only a single talker to generate the stimuli. We wondered what would happen to a listeners perceptual response if we introduced variability into the stimulus ensemble. At the time this work was carried out, we were unaware of the earlier findings reported by Peters and Creelman.

In our first experiment, talker variability was manipulated by having listeners identify familiar English words under two conditions. In the first condition, all the stimulus words were produced by a single talker; in the second condition, the same words were produced by 15 different talkers (Mullennix, Pisoni, & Martin, 1989). Subjects identified the words at three different speech-to-noise ratios, +10 dB, 0 dB, and -10 dB. Across all three conditions, we found that identification performance was more accurate for single-talker lists than multiple-talker lists. This finding replicated the earlier results of Peters and Creelman (see Mullennix, Pisoni, & Martin, 1989). In a second experiment, we measured naming latencies for the same two conditions. The results showed that subjects were not only slower, but were also less accurate in naming words from multiple-talker lists than words from single-talker lists.

Apparently, listeners are quite sensitive to the trial-by-trial variability and context of the test items, specifically, the uncertainty associated with changes in the talker's voice from one stimulus to the next in the experiment. This was true despite the fact that all the test items were highly intelligible when presented in isolation without any masking noise.

One could argue, as we have done in several papers, that these findings support the view that talker normalization requires attention and processing resources and does not occur automatically without some cost to the listener (See Mullennix & Pisoni, 1990). It is as if the human listener were a very fast time-varying adaptive filter that changes its characteristics as the stimulus input changes, even from trial to trial in an experiment, in order to "optimize" recognition performance. The exact properties of this filter need to be studied in greater detail.

Another issue that we have investigated recently concerned the relationship between talker variability and phonetic coding. Do the perceptual processes used to encode voice information function independently of processes that are used to encode phonetic information? One way to determine whether perceptual processes are related to one another is to assess whether the stimulus dimensions are perceived independently or whether there is some dependency relation between them.

In experiments of this type, known as a Garner speeded classification task, subjects are required to attend selectively to one stimulus dimension while simultaneously ignoring another stimulus dimension (Garner, 1973). The two stimulus dimensions are combined in various ways. In the "control set," the unattended dimension is held constant while the attended dimension varies randomly from trial to trial. The control set for each dimension provides a baseline measure for classifying each dimension and permits one to assess whether both dimensions are, a priori, equally discriminable. In the "orthogonal set," both the attended and the unattended dimensions vary randomly. The degree to which response latencies increase from the control set to the orthogonal set for each dimension indicates the extent to which the stimulus dimensions are processed separably or in an integral fashion. If the same stimulus dimensions are classified as quickly in the orthogonal conditions as they are in the control conditions, then the stimulus dimensions are said to be processed separably. That is, decisions about the relevant dimension are unaffected by the variation on the irrelevant dimension. However, if there is a significant increase in response latencies from the control conditions to the orthogonal conditions, then the stimulus dimensions are said to be processed in an integral manner. In this case, subjects cannot ignore or "filter out" variation on the irrelevant dimension. This result, which is termed orthogonal interference, indicates that a failure of selective attention to the attended dimension has occurred.

Insert Figure 1 about here

Figure 1. shows the amount of orthogonal interference (in msec.) for the voice and word dimensions for each of four stimulus conditions from an experiment carried out by Mullennix and Pisoni (1990). Across all conditions, significant increases in orthogonal interference were obtained when subjects were required to attend to either the word or the voice dimension. The pattern of results shows clearly that the processing of each dimension affects classification of the other dimension. Moreover, this effect increased as stimulus variability increased. Thus, each dimension affects decisions on the other dimension and the magnitude of the effect increases as stimulus variability increases.

Memory Experiments on Talker Variability

We have also completed a series of memory studies to learn more about the perceptual representation of spoken words in long-term memory (see Martin, Mullennix, Pisoni, & Sommers, 1989). In one recent study, carried out by Lightfoot (1989), listeners were trained to explicitly identify voices in a perceptual learning task. After ten days of training, subjects learned arbitrary names for an unfamiliar set of male and female voices; then they were given several memory tests with a novel set of words produced by the same talkers in order to assess the effects of familiarity on recall. Performance on serial recall tests with the novel words was improved relative to the control group who heard the same test words but did not have any prior experience of familiarity identifying the talkers who produced the test items. Lightfoot's results suggest that as a listener becomes familiar with a particular voice, some aspect of the rehearsal process becomes more efficient so that additional talker specific cues can be used to aid retrieval of these items from long-term memory (see also Goldinger, Pisoni, & Logan, 1991).

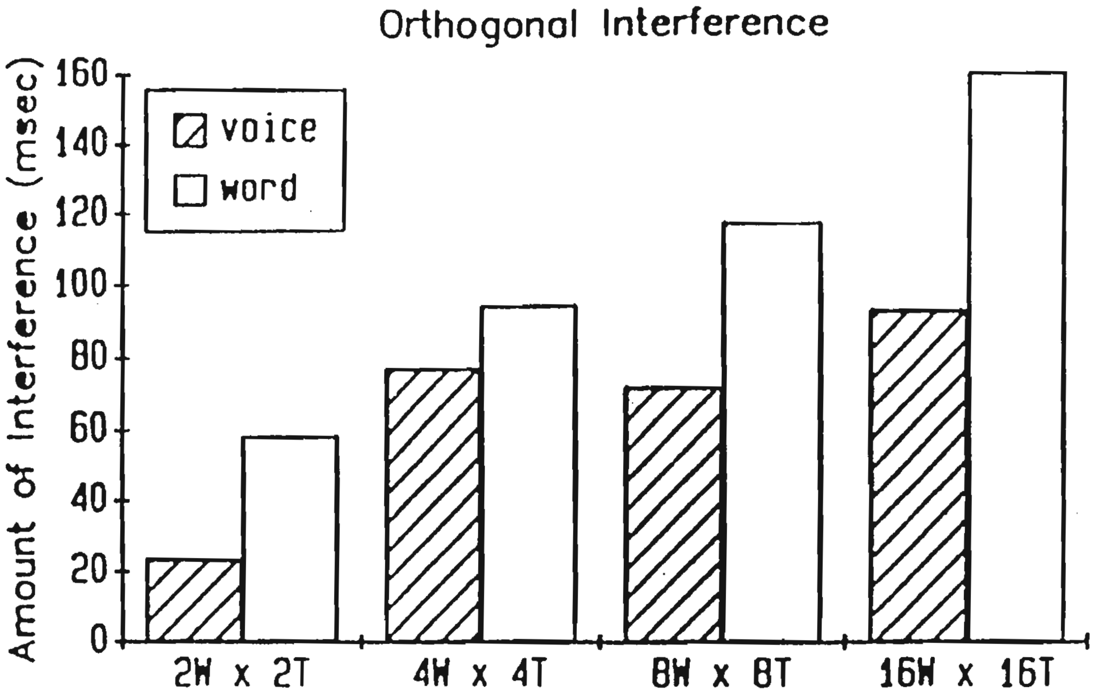


Figure 1. Amount of orthogonal interference expressed in milliseconds for word and voice dimensions at each of four levels of stimulus variability (From Mullennix & Pisoni, 1990).

In another recent study using a continuous recognition memory task, Palmeri, Goldinger and Pisoni (1990) found that detailed information about a talker's voice is preserved in memory even when there is a great deal of competition from other voices in the set of stimulus items. In this experimental procedure, subjects received a very long list of words and were required to determine if each word had occurred previously during the test. Every word was repeated once, either in the same voice or a different voice, at lags of from 1 to 64 items.

Insert Figures 2 and 3 about here

The major results for "old" responses on the recognition test are shown in Figure 2 as a function of lag. Across all lag conditions, hit rates for same-voice repetitions were more accurate than hit rates for different-voice repetitions. This finding replicates earlier results of Craik and Kirsner (1974) who found that same-voice repetitions were responded to more accurately than different-voice repetitions when two voices were used. However, the most interesting findings from our study are shown in Figure 3. When performance on both "old" and "new" trials is examined as a function of the amount of talker variability in the test sequence, we found no change in performance for both same-voice or different-voice repetitions. Performance is expressed here by d' scores. The same pattern was observed across all four conditions of talker variability ranging from the two-talker condition to the twenty-talker condition. These results provide additional support for the proposal that detailed information about a talker's voice is preserved in some form after perceptual analysis and is subsequently encoded into long-term memory.

Taken together, these new finding raise several important theoretical questions, not only about the kinds of processing operations and representations that have been assumed in the past, but also about the close dependency that most current models of speech perception have traditionally had on linguistic theory, particularly in terms of their reliance on abstract units such as phonemes. Given our recent findings, I believe there are some good reasons to be a little more skeptical of the psychological reality of phonemes as perceptual units for human listeners.

The present findings also bear a close similarity to a long series of studies in the field of cognitive psychology on implicit memory phenomena. Many of these studies, carried out in the growing tradition of what is called "nonanalytic cognition", have suggested that memory, particularly recognition, depends on reference to specific instances or records of the stimulus pattern (Jacoby & Brooks, 1984). Thus, according to this view, each stimulus event creates a permanent record in long-term memory. The ensemble of these events functions as a set of exemplars that can be used to classify novel stimuli into familiar categories. Within this framework, the nonanalytic instance-based views of memory have been able to deal with stimulus variability more directly than the more popular analytic approaches to cognition which have emphasized abstraction and symbolic coding of the stimulus input with substantial loss of the details of the perceptual input.

Concluding Remarks

This is an exciting time to be working in the field of speech perception. While many of the old problems still remain and occupy the attention of a very diverse group of researchers, new findings such as those summarized here on talker variability have begun to raise important questions about the old dogma and theoretical assumptions that have been so pervasive over the last 45 years. Our findings suggest several general conclusions that are basically incompatible with traditional views of speech perception. First, detailed information about the source characteristics of the talker is retained in memory for some period of time. Second, most current accounts of speech perception have assumed that talker normalization processes eliminate variability in the speech signal and produce an idealized symbolic representation that is isomorphic with the linguist's description of speech as a sequence of phonemes;

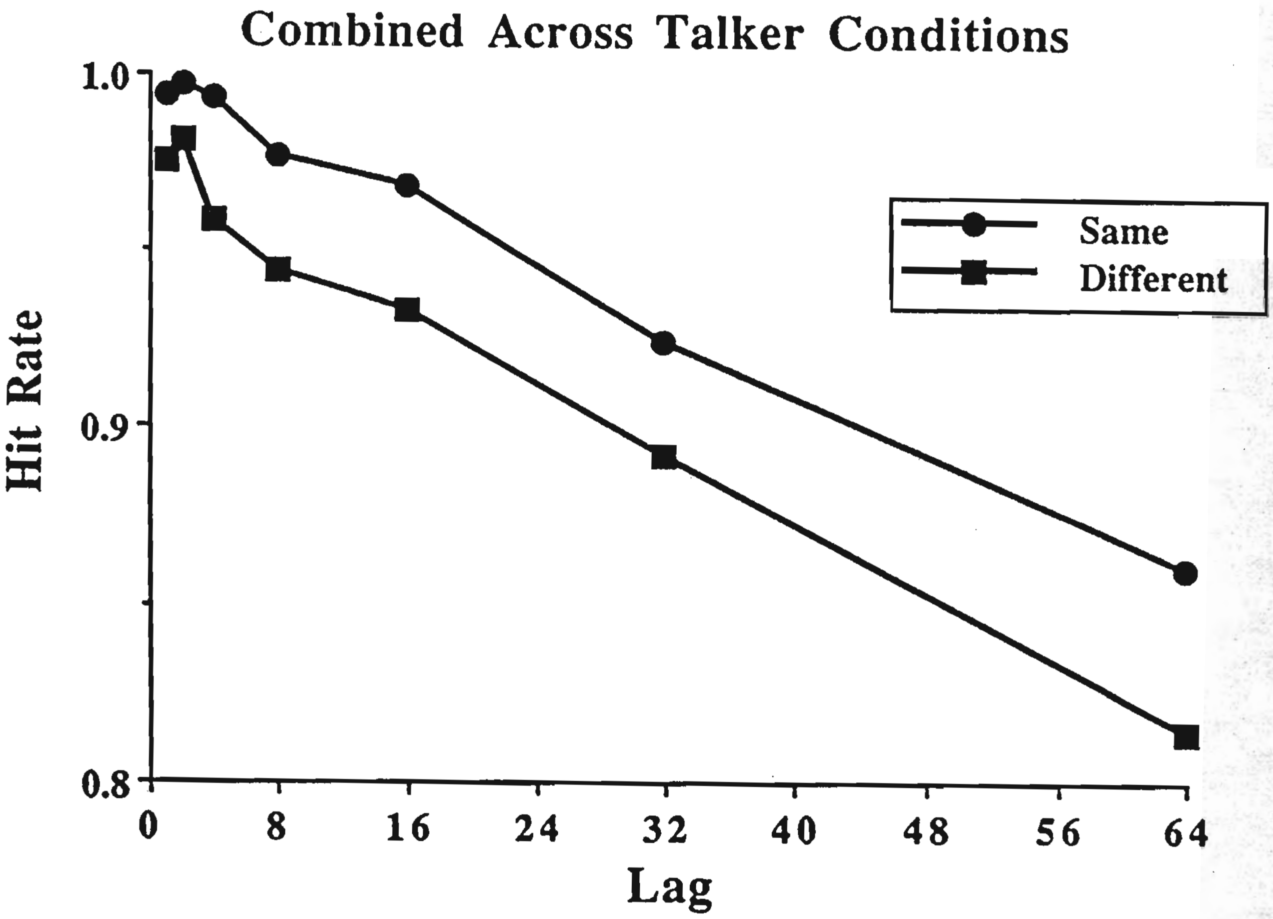


Figure 2. Hit rates for "old" responses across talker variability groups as a function of lag for same-voice and different-voice repetitions (From Palmeri et al., 1990).

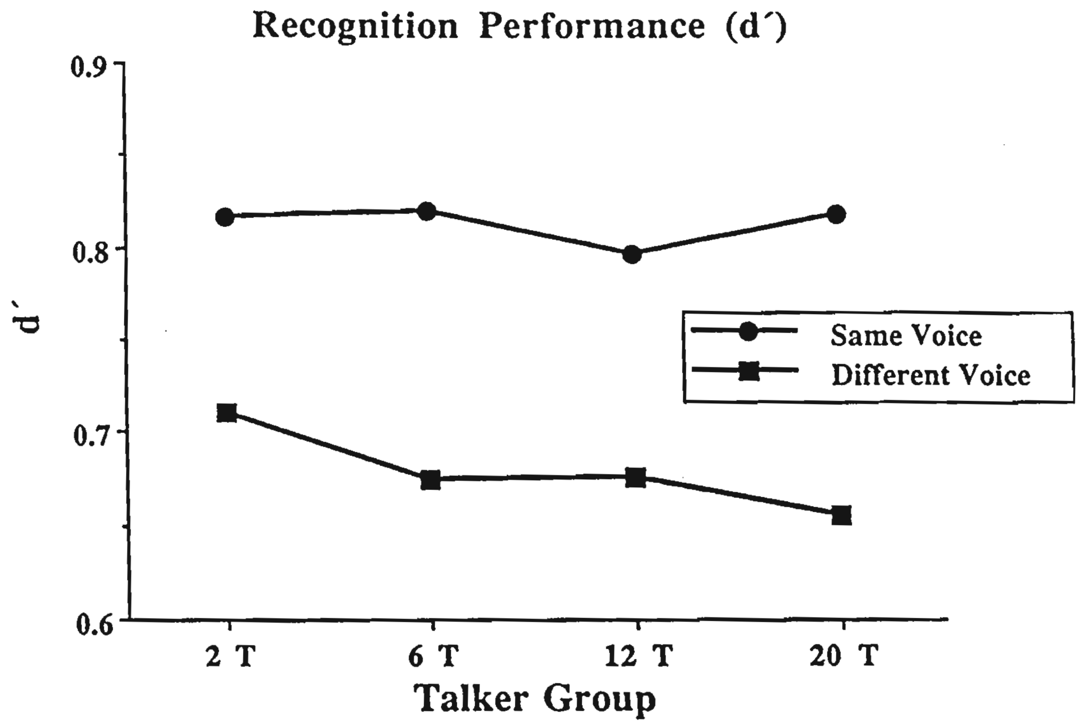


Figure 3. Disçriminability scores (d') collapsed across lag for same-voice and different-voice repetitions as a function of talker variability (From Palmeri et al., 1990).

the present findings suggest that this view is incorrect. Third, talker variability not only affects speech perception but our findings show that this source of variability also influences performance in auditory memory tasks by modifying the long-term memory representation of the stimulus item. Listeners apparently encode quite detailed talker-specific attributes and use these additional distinctive properties to encode and retrieve information in long-term lexical memory.

These conclusions are consistent with proposals made by Dennis Klatt (1979) in motivating his LAFS model of speech perception. He argued that traditional phonetic representations discard detailed acoustic information that would be useful for lexical access. According to Klatt, the loss of this detailed acoustic information may produce errors in lexical interpretation that would be difficult to correct if only an idealized symbolic representation were available in memory. Klatt strongly believed that canonical phonetic representations were sub-optimal in human speech perception and machine speech recognition because they violated the principle of "delayed commitment." This principle states that information in the speech signal should not be discarded until it is no longer of any potential use in perception or recognition. We believe that Klatt was basically correct in his emphasis on preserving detailed acoustic-phonetic information in the speech signal. In the years to come, we hope to learn much more about the perceptual representation of speech and the manner in which this information becomes a part of the long-term memory representation for spoken words in the mental lexicon.

References

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.

Garner, W.R. (1973). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

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

Jacoby, L.L., & Brooks, L.R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In Bower, G.H. (Ed.), *The Psychology of Learning and Motivation, Volume 18*. New York: Academic Press, 1-47.

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

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

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 and Psychophysics*, **47**, 379-390.

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1990). Episodic encoding of Voice and Recognition Memory for Spoken Words. *Research on Speech Perception Progress Report No. 16*. Bloomington, IN: Speech Research Laboratory, Indiana University.

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

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

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

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Lexical Memory in Auditory and Visual Modalities: The Case for a Common Mental Lexicon[1]

David B. Pisoni and Ellen E. Garber

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

[1] This research was supported by NIH Training Grant #DC00012-12 and NIH Research Grant #DC00111-14 to Indiana University. We thank Paul Luce for his help and advice.

Abstract

This paper reports the results of a study designed to measure differences in familiarity for spoken and written words. Two sets of 450 English words were randomly selected from a computerized version of Webster's pocket dictionary. Four groups of subjects were presented with both lists of words for familiarity judgements. The first group (VV), saw all the words on each list presented visually; the second group (AA) heard all the words; the third and fourth groups (AV, VA) received one list visually and another list auditorily. Subjects rated the familiarity of each word using a seven-point scale. Correlations of the familiarity scores across both lists and modalities were very high. The mean ratings were not significantly different for visual and auditory groups. The absence of modality differences suggests that familiarity effects occur late in the processing system where information from the input modality converges on a common lexical store in long-term memory.

Lexical Memory in Auditory and Visual Modalities: The Case for a Common Mental Lexicon

For many years, the field of speech perception has been conspicuously isolated from the mainstream of research on spoken language processing that has addressed problems of word recognition, lexical access, sentence comprehension and discourse processing. This situation came about, in part, because the primary concern in speech perception research had been on the physical correlates of speech, that is, acoustic-phonetics (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Very little interest or attention was devoted to the interface between speech and more abstract levels of language processing. Fortunately, this situation has changed recently as more researchers have turned their attention to somewhat broader issues that encompass not only the search for the acoustic cues to phoneme perception but the mapping problem between speech and the lexical representations of words in the listener's long-term memory. Many investigators realized that our understanding of how human listeners perceive words and understand fluent speech was very impoverished compared to the extensive literature on the perception of phonemes in isolated nonsense syllables. As researchers began to pursue these issues, new research strategies and techniques began to evolve to study speech perception and spoken word recognition (Pisoni, 1985a).

For the last five years, we have been working with several computational techniques to learn more about the structural organization of spoken words in the mental lexicon, and the role this organization may play in perception and memory (Pisoni, 1985b). To study these issues, we acquired several large lexical databases and carried out a number of analyses of the structural patterns of words and their organization. We computed similarity spaces using several quantitative metrics that were designed to index the phonetic distance among sound patterns in English. And, we studied how these similarity spaces affect perception in numerous behavioral experiments. One of these databases, Webster's Pocket Dictionary, contains 20,000 words. Each entry contains the standard English orthography for a word, a phonetic transcription, and special codes indicating the syntactic functions of the word. Frequency information was added to each entry when available by merging information from the Kucera and Francis word count. We also developed several ways of computing similarity neighborhoods for words in this database so that quantitative measures of "lexical density" could be included with each entry. Numerous experiments over the last five years have demonstrated that the number and nature of words in a similarity neighborhood strongly affects word recognition performance and memory in a variety of experimental tasks. In particular, neighborhood structure was shown to be an important determinant of the speed and accuracy of word recognition (Luce, 1986). The number of words in a similarity neighborhood and their relative frequencies have been shown to affect spoken word recognition performance in perceptual identification, lexical decision, and naming tasks (Luce, 1986). These same structural factors also affect encoding and retrieval of words from long-term memory (Goldinger, Pisoni, & Logan, 1991).

In the course of carrying out this program of research, we also collected familiarity ratings from human observers for each of the 20,000 words in the Webster's database in order to have both objective frequency counts and subjective familiarity judgements for each entry (Nusbaum, Pisoni, & Davis, 1986). Although many researchers have routinely assumed that objective frequency counts provide valid estimates of experienced frequency, other theorists have argued that this practice may be inappropriate because frequency counts substantially underestimate exposure to many words that are actually present in a listener's mental lexicon but were simply not included in the corpora because of sampling problems. A number of years ago, Gernsbacher (1984) reported that familiarity ratings from subjects were more accurate predictors of performance on several psycholinguistic tasks than published estimates of word frequency derived from objective word counts.

The familiarity data that we collected several years ago and routinely use in our experimental work were all obtained with visually presented materials. The question naturally arises as to whether these familiarity ratings can be generalized across modalities and whether they are appropriate for spoken as well as written words. Do these familiarity ratings reflect information that is, in some sense, modality-specific, or do they index information about words that is modality-independent, information that presumably resides in some common mental lexicon that is shared across modalities? If familiarity judgements for the same set of words using visual and auditory presentation are comparable, or nearly so, one could argue that these results reflect access to or retrieval of information contained in a modality-independent lexicon rather than separate modality-specific lexicons that would be functionally autonomous for auditory and visual inputs. Results of this kind would not only be relevant to issues about the existence of a single modality-independent lexicon in language processing but they would also be extremely useful in resolving several long-standing problems concerning the locus of frequency effects in word recognition studies (Balota & Chumbley, 1984).

A number of theorists have argued that frequency effects occur very early in perceptual processing prior to lexical access and retrieval of the meaning of a word from long-term memory (Marslen-Wilson, 1984). If familiarity effects are modality-independent and reflect access to information that is common to both auditory and visual inputs, the locus of these effects may therefore occur relatively late in perceptual processing presumably sometime after the two input modalities have converged on a common representation in lexical memory. To study this problem, we presented lists of words both visually and auditorily to separate groups of subjects to determine if the subjective familiarity ratings assigned to words would be dependent on the modality of input.

Method

Subjects

Ninety-six undergraduate students from introductory psychology courses served as subjects. Each subject attended two fifty-minute sessions and received partial course credit for their participation. All subjects were native speakers of English who reported no history of a speech or hearing disorder at the time of testing.

Materials

Two lists of 450 words were constructed using a computerized lexical database derived from Webster's Pocket Dictionary. This database included frequency counts for 11,750 words from the Kucera and Francis (1967) word count and familiarity ratings from the earlier study by Nusbaum, Pisoni and Davis (1986). The two sets of words were randomly selected from the database with the restriction that each list contain 150 words with high familiarity ratings (5.1-7.0), 150 words with medium familiarity ratings (3.1-5.0), and 150 words with low familiarity ratings (1.0-3.0). For the auditory presentation, the words on each list were spoken in isolation by a male talker. For the visual presentation, the words were presented on a CRT using a video character generator controlled by a computer.

Procedure

Subjects were tested in groups of six or fewer in a sound treated room. Each subject was seated in an individual testing booth which was equipped with a GBC CRT monitor and a pair of TDH-39 headphones. Subjects were instructed to rate the subjective familiarity of each word using a seven point scale, ranging from 7 (very familiar word) to 1 (word was unknown). Subjects recorded their familiarity judgements by pressing the appropriate button on a response box that was interfaced to a computer.

Each subject received two lists of words, one list in each of two sessions. The lists were presented either visually (V) or auditorily (A) in each session. Twenty-four subjects participated in each of the four conditions (VV, AA, AV and VA). Within each condition, the order of presentation of the two lists was counterbalanced so that 48 subjects received List 1 first and another 48 subjects received List 2 first. Presentation of the stimuli was randomized for each group of subjects.

Presentation of the stimuli and collection of the data were controlled on-line by a PDP 11/34 minicomputer. In the auditory condition subjects listened to the words over headphones. The onset of a light on the response box indicated the beginning of a new trial. In the visual condition, the words were presented in uppercase letters on a CRT monitor. A line of asterisks on the screen indicated the onset of a new trial. The word remained on the screen until the subject responded. In both the visual and auditory conditions, response times were measured from the onset of each stimulus until the subject initiated a response. Subjects had eight seconds to respond. At the end of the time period, the computer proceeded to the next trial.

Results

Mean familiarity ratings were computed for each word in each modality of presentation. A summary of these results is shown in Figure 1. All four conditions displayed the same pattern of ratings across the three levels of familiarity. Analysis of variance confirmed the trends shown in the figure. Modality was not significant. Thus, the familiarity ratings were unaffected by mode of presentation. Not surprisingly, familiarity produced a highly significant effect [*min* $F'(2,181)=196.62$, $p<.0001$]. No significant interactions were obtained.

Insert Figure 1 about here

The response latency data shown in Figure 2 display a similar trend across the three levels of word familiarity. Analysis of variance revealed significant main effects for modality [*min* $F'(2,242)=3.9$, $p<.01$] and familiarity [*min* $F'(2,343)=10.29$, $p<.0001$]. Overall, highly familiar words were responded to more rapidly than unfamiliar words and auditory presentation was slower than visual presentation. None of the interactions were significant.

Insert Figure 2 about here

Figure 3 shows a scatter plot of the familiarity ratings for auditory and visual presentations of the two word lists. The correlations between the auditory and visual ratings were high and statistically reliable (Auditory List 1 vs. Visual List 2, $r=+.93$; Auditory List 2 vs. Visual List 1, $r=+.92$).

Insert Figure 3 about here

Discussion

The results of the present experiment clearly demonstrate that familiarity judgements are independent of input modality. The sensory modality used to present the word lists apparently did not affect subject's familiarity judgements in this task. Moreover, the correlations between the two word lists

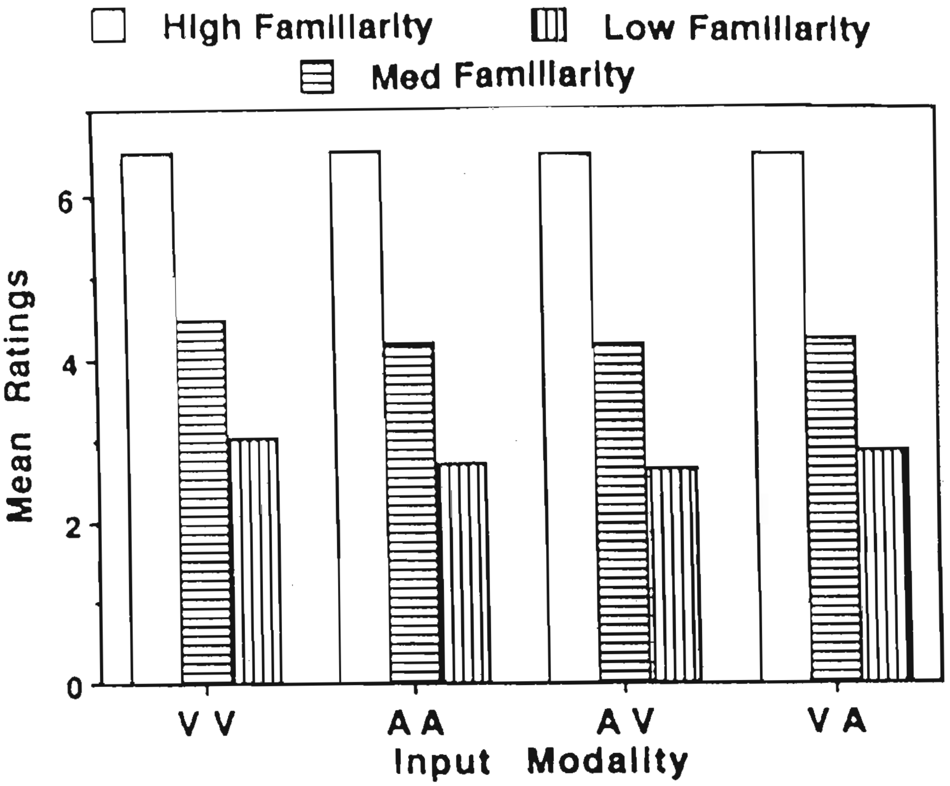


Figure 1. Mean familiarity ratings for high, medium and low words as a function of input modality.

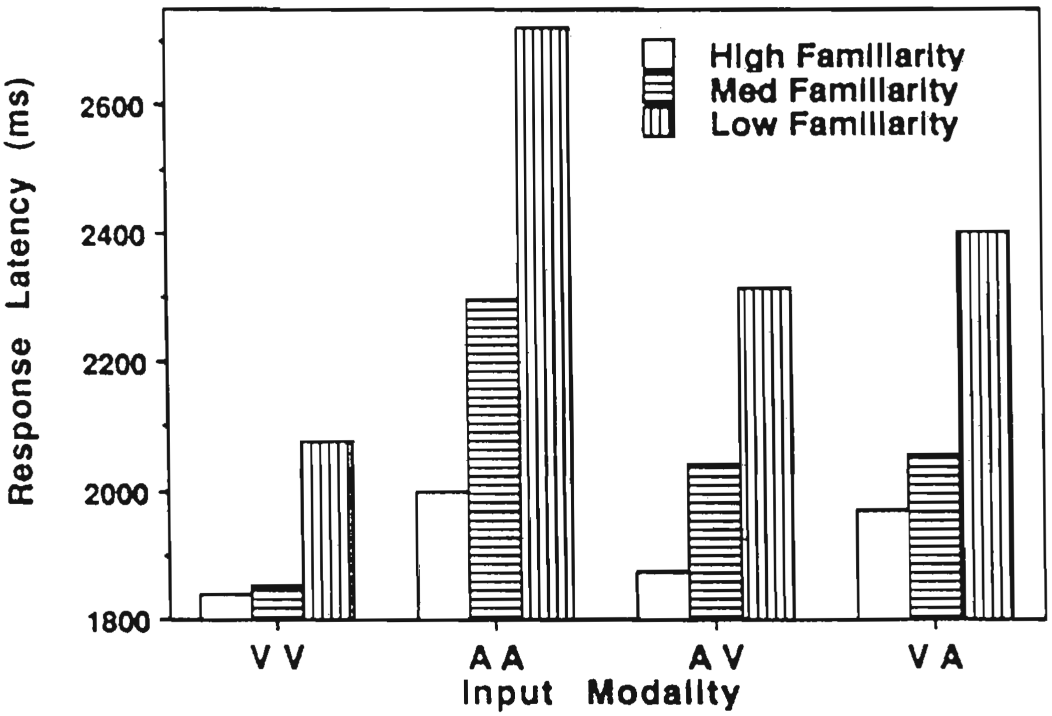


Figure 2. Response latencies for high, medium and low words as a function of input modality.

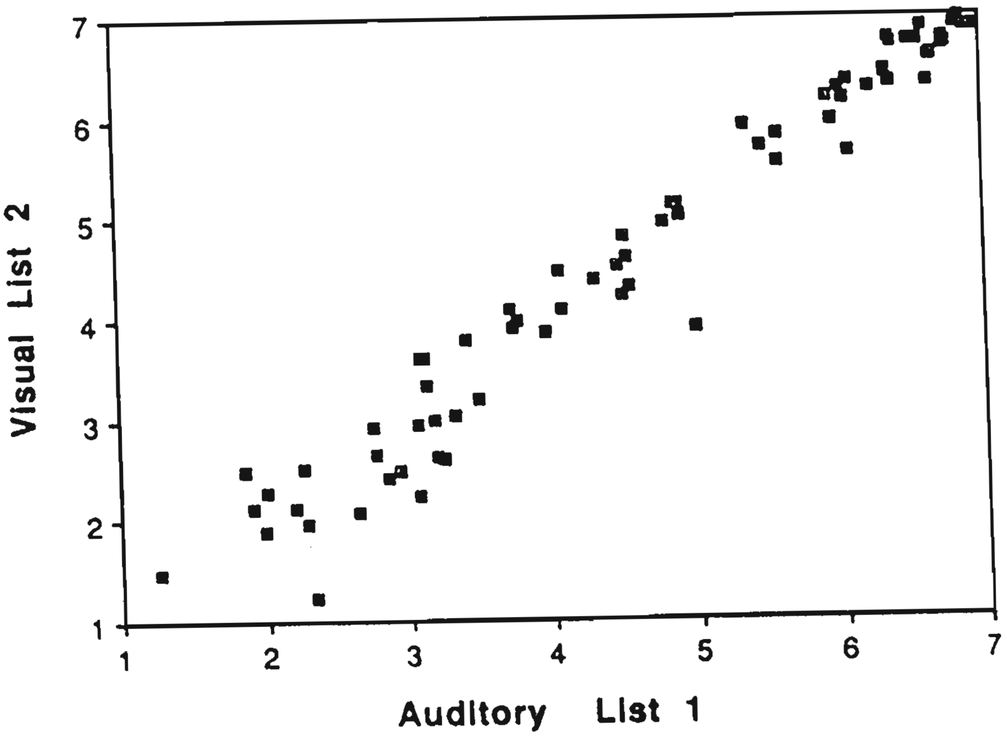


Figure 3. Scatterplot of familiarity ratings for auditory and visual presentation.

both within and between modalities were all positive and extremely high, suggesting a common basis for the familiarity judgements that is independent of modality of input.

The results of this study have a number of implications for current research on word recognition and models of lexical access. First, in terms of our original goals, the present findings demonstrate that the familiarity ratings obtained in our earlier investigation, which used only visual presentation, can be generalized to spoken words with some degree of confidence. This is useful information to have when constructing materials for perceptual experiments and making generalizations across modalities.

Second, the absence of modality effects in this task raises several questions about the claims made by Forster (1976) concerning the locus of frequency effects in word recognition experiments. Forster states that the frequency of a printed word may differ widely from the frequency of the spoken form of the same word. This assertion, among others, was used to motivate the proposal of three separate peripheral access files or bins in his search model. The entries in each bin are listed according to the frequency of occurrence of the word in the language. According to Forster's search model, a word is recognized by a two-step process. First, a description of the stimulus features of a word, which Forster calls the "access code" is located by searching for an entry in one of the peripheral access files. When a match is found, the search process terminates and a pointer specifies the location of that word in the master file. The master file contains all the information that a listener has about a word whereas the peripheral access files contain only access codes and pointers to words in the master file. Thus, the peripheral access files are organized by both modality and frequency to permit efficient search. Within each modality-specific bin, the entries are arranged by frequency so that high-frequency words can be located earlier than low-frequency words. If we assume that word frequency somehow underlies or controls familiarity judgements through exposure to words in the language, Forster's claim about differential frequency effects would appear to be incorrect because neither frequency nor familiarity effects are influenced by input modality. A similar set of criticisms concerning frequency effects can be raised about Morton's (1979) proposal of two input logogens to deal with modality-specific facilitation effects. In Morton's model, the logogens are not only frequency-sensitive but they are modality-specific as well. Again, the present results question the need for proposing modality specific logogens to account for frequency effects.

Third, another closely related issue deals with the question of the precise locus of frequency effects in current models of word recognition. The well-known models proposed by Morton (1979), Forster (1976) and Marslen-Wilson (1984) all place the locus of frequency effects relatively early in perceptual processing which precedes lexical access. The absence of any modality differences and the high correlations obtained in the present study within and across modalities suggests that frequency and familiarity effects occur fairly late in processing, after input has converged on a common representation rather than early on during search or activation as suggested by these investigators.

Finally, the present results provide additional evidence for a dissociation between frequency and modality effects. Studies by Lee, Ovid, Tzeng, and Hung (1978) and Kirsner, Milech, and Standen (1983) have found that frequency effects are largely independent of input modality. Their findings, taken together with the present results, suggest either a common, modality-independent representation for words or the use of differential retrieval processes for modality-specific and frequency-specific information in long-term memory. Whatever the final explanation turns out to be, it is clear from the results of the present study that judgements of subjective familiarity for words are made on the basis of information that is not coded or indexed by input modality. Thus, there appear to be both common and modality-specific components to the mental lexicon. The extent to which these two sets of attributes can be accessed in various psycholinguistic tasks will depend largely on the specific requirements of the information processing task and the strategies adopted by the subject.

References

Balota, D.A. & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Forster, K.I. (1976). Accessing the mental lexicon. In Wales, R.J. & Walker, E.C.T. (Eds.), *New Approaches to Language Mechanisms*. Amsterdam: North-Holland, 257-287.

Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysomy. *Journal of Experimental Psychology: General*, **113**, 256-281.

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.

Kirsner, K., Milech, D., & Standen, P. (1983). Common and modality-specific processes in the mental lexicon. *Memory and Cognition*, **11**, 621-630.

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

Lee, A.T., Ovid, J.L., Tzeng, L.C., & Hung, D.L. (1978). Sensory modality and the word-frequency effect. *Memory and Cognition*, **6**, 306-311.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

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, 1-91.

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

Morton, J. (1979). Facilitation in word recognition: Experiments causing a change in the logogen model. In Kolers, P.A., Wrolstad, M.E., & Bouma, H. (Eds.), *Processing Visible Language*. New York: Plenum Press, 259-268.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1986). 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, 357-376.

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

Pisoni, D.B. (1985b). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, **4**, 75-95.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 16 (1990)
Indiana University

Some New Directions in Research on Comprehension of Synthetic Speech[1]

David B. Pisoni, James V. Ralston and Scott E. Lively

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

[1] This research is supported in part by NSF Research Grant #IRI-86-17847 to Indiana University.

Some New Directions in Research on Comprehension of Synthetic Speech

In thinking about new directions for future research, it is often a good idea to sit back and review what has been done in the past. This exercise not only helps to identify the major accomplishments of the past but it frequently highlights some of the gaps in our fundamental knowledge that will need attention in the future. A review of the previous research on the perception of synthetic speech over the last ten years reveals that the overwhelming bulk of the research efforts have been devoted to questions dealing with the measurement of segmental intelligibility. For the most part, research on word recognition, sentence processing and comprehension of long passages of fluent speech has received very little attention.

This should not come as a surprise to anyone who has worked in the field of speech processing. Intelligibility tests are not only very easy to carry out, but they have a long and distinguished history that goes back to before World War II. Moreover, it is fairly easy to interpret the results of intelligibility tests and make relevant comparisons between various speech communication systems. Most engineers and speech scientists are familiar with many of the traditional speech intelligibility tests that are widely used at the present time. They know what the test is and what the results mean in terms of assessing gross aspects of overall system performance. For example, a voice output system with a low MRT or DRT score is a system that users will have some difficulty with; a system with higher scores on these tests will "sound" better and users will report fewer problems. If these are the only behavioral criteria that engineers and speech scientists are interested in then speech intelligibility tests such as these will have served their purpose. But there are several other important issues and questions that cannot be addressed by simple standardized speech intelligibility tests like the MRT or the DRT.

When we begin to ask questions about word recognition, sentence processing, or comprehension of linguistic messages, it is necessary to move away from simple tests that measure well-defined acoustic-phonetic properties of speech to more complicated experimental procedures that assess the cognitive and linguistic contributions of the listener. These contributions are enormous and are often underestimated by engineers and system designers who would like to have a single number to describe the performance of a particular speech communication system. Because human listeners are extremely robust observers and spoken language is a highly redundant symbol system, it has been difficult to measure the more cognitive aspects of spoken language processing such as comprehension. Moreover, there have been very few attempts to relate standardized measures of segmental intelligibility to measures of comprehension performance. These two areas of research have been pretty much independent of each other for many years. Researchers concerned with measuring speech intelligibility and developing standards for assessment of speech communication systems rarely, if ever, show any interest in the study of comprehension. One of the main reasons for the lack of interest is that comprehension is very complicated and relies on linguistic knowledge that is hard to quantify with a few simple numbers like those typically obtained in intelligibility tests.

Most cognitive psychologists agree that comprehension is a process that involves encoding of sensory information, the retrieval of previously stored knowledge from long-term memory, and the subsequent interpretation, integration and assimilation of various sources of knowledge. Comprehension of spoken language depends on a relatively large number of diverse factors, some of which are still only poorly understood by psycholinguists at the present time. Measuring comprehension is an even more difficult problem because of the interaction of many factors and the absence of any coherent model of comprehension that is broad enough to encompass the diverse nature of language understanding.

In addition to our research on segmental intelligibility over the last ten years, we have also been interested in studying comprehension. One of our goals has been to relate traditional measures of segmental intelligibility such as the MRT to several different measures of comprehension performance. Another somewhat broader goal has been to simply learn more about the comprehension process itself. At the present time, as far as we know, there are no accepted standards for measuring comprehension performance. We believe that before any standards are even considered, it is important to acquire more detailed information about how to measure comprehension. Although we have devoted some of our efforts to these problems over the past few years, we actually made some initial attempts to measure comprehension and relate those findings to measures of speech intelligibility back in 1979.

Two distinct lines of research on comprehension have been carried out in our laboratory at Indiana University. One set of studies has examined the comprehension of isolated sentences using a sentence verification task (SVT). In this procedure, we measure the time it takes a listener to answer simple questions. Another set of studies has been concerned with developing online measures of comprehension for much longer passages of fluent speech.

In the sentence verification paradigm, subjects are required to judge the truth value of short sentences in relation to some prior information, such as a picture or their general world knowledge (Gough, 1965; Larkey, & Danly, 1983). Because the verification task is quite easy, error rates are typically very low and response latency is the primary dependent variable of interest. One of our studies used the sentence verification technique in combination with a transcription task to study the comprehension of natural speech and synthetic speech produced by five different text-to-speech systems. The verification latency data yielded a rank ordering of the performance of the systems that was very similar to the ranking found in our study of the segmental intelligibility of these systems (Logan, Greene, & Pisoni, 1989). In addition, moderate to high correlations were observed between transcription error rates and verification measures. Other studies have used the SVT with synthetic speech, natural speech and coded speech and found similar results (Pisoni & Dedina, 1986; Pisoni, Manous, & Dedina, 1987).

The close relationship found between segmental intelligibility and sentence comprehension in the SVT suggests that comprehension of isolated sentences depends, to a large extent, on phoneme intelligibility and accurate word recognition. It seems unlikely that the results obtained with the SVT can be generalized to the comprehension of passages of fluent speech where listeners must integrate information across several sentences to derive the semantic content of an entire passage. How well global comprehension performance can be predicted from traditional measures of segmental intelligibility is obviously a topic of some interest given the extensive reliance on standardized measures of speech intelligibility.

In order to address this problem, we have recently completed a series of experiments that have employed two different online tasks to assess comprehension of long passages of fluent speech (Lively, Ralston, Pisoni, & Rivera, 1990; Ralston, Pisoni, Lively, Greene, & Mullennix, 1991). Measures of segmental intelligibility using the MRT were also obtained from each listener so that correlations could be computed between speech intelligibility and comprehension performance.

In one experiment, we used a word monitoring task in which subjects had to detect the presence of a set of target words while they simultaneously listened to the passage in order to understand it (Lively et al., 1990). After the passage was presented, subjects were required to respond to a series of questions that were designed to assess aspects of their memory for specific words and propositions that were contained in the passage. We found that word monitoring performance was slower and less accurate for passages of synthetic speech than passages of natural speech. Performance in answering questions on the memory test was also slower and less accurate for synthetic speech.

Insert Figures 1, 2, and 3 about here

In another study, we used a self-paced sentence-by-sentence listening task to measure differences in online processing between natural speech and synthetic speech (Lively et al., 1990). This procedure permitted subjects to control the presentation of successive sentences in a passage by pressing a response button. Subjects heard one sentence at a time and they could control the rate of presentation of successive sentences. Response latencies were used as the dependent variable to index the amount of listening time needed to understand each sentence. As in the previous experiment, the sentence-by-sentence listening times were slower for passages of synthetic speech than for passages of natural speech. Performance on the memory test was very similar to the findings obtained in the word monitoring experiment. Subjects were slower and less accurate in responding to the memory questions after listening to synthetic speech.

Insert Figures 4 and 5 about here

When the individual sentences in each passage were randomly presented thus preventing subjects from constructing a coherent textbase, the sentence-by-sentence listening times increased dramatically for both natural speech and synthetic speech (Ralston et al., 1991). However, the increase was even larger for the passages of synthetic speech. These results suggest that when subjects listen to long passages of synthetic speech, they may make much greater use of global information about the structure of the text and the organization of the component sentences into an abstract propositional macrostructure or schema than when they listen to natural speech.

Insert Figure 6 about here

We also computed correlations between the MRT intelligibility scores and several measures of comprehension performance. The results of these analyses revealed that word monitoring performance, sentence-by-sentence listening times, and memory were only moderately correlated with the speech intelligibility scores obtained from the MRT.

Insert Tables 1 and 2 about here

Several conclusions can be drawn from the results obtained with these new techniques. First, both measures show that comprehension of synthetic speech is poorer than natural speech. This is not surprising at all given the differences in segmental intelligibility.

Second, the online word monitoring task and the sentence-by-sentence listening procedure appear to be extremely sensitive measures of performance that can be successfully used to study the real-time comprehension of spoken language. The difficulties that listeners encounter in comprehending long passages of synthetic speech can be detected with two very different types of procedures. Moreover, at least in the case of the listening time task, we were able to detect changes in the comprehension process when very high level text variables were manipulated by randomizing the sentences in the passage. Independently of issues surrounding the perception of synthetic speech, this finding is of interest more generally to questions about the time course of comprehension and how the listener allocates attentional resources across very different types of processing activities.

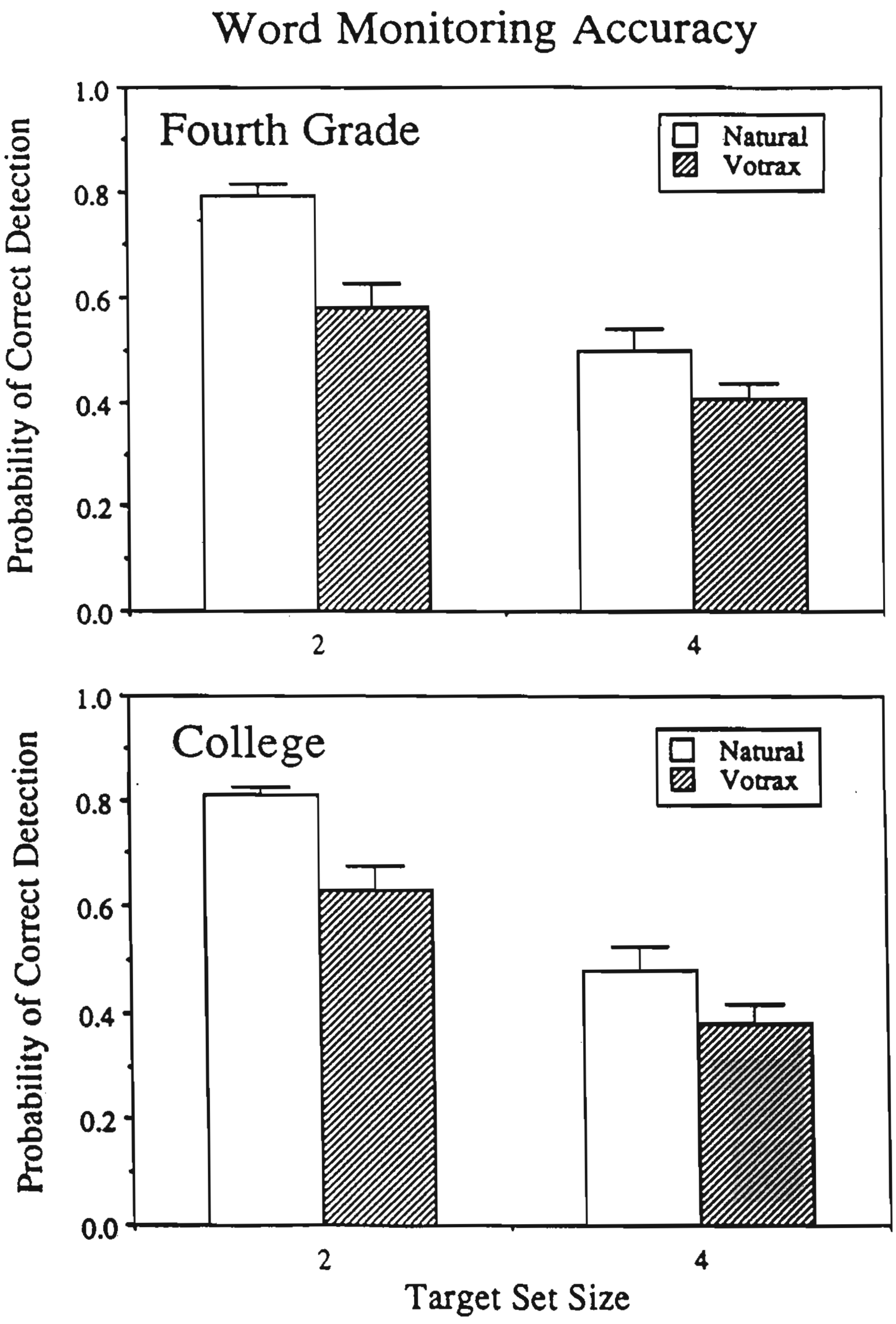


Figure 1. Word monitoring accuracy as a function of target set size. The upper panel shows the data for the fourth-grade passages; the lower panel shows the data for the college-level passages. Open bars are for natural speech; striped bars are for Votrax synthetic speech (From Lively et al., 1990).

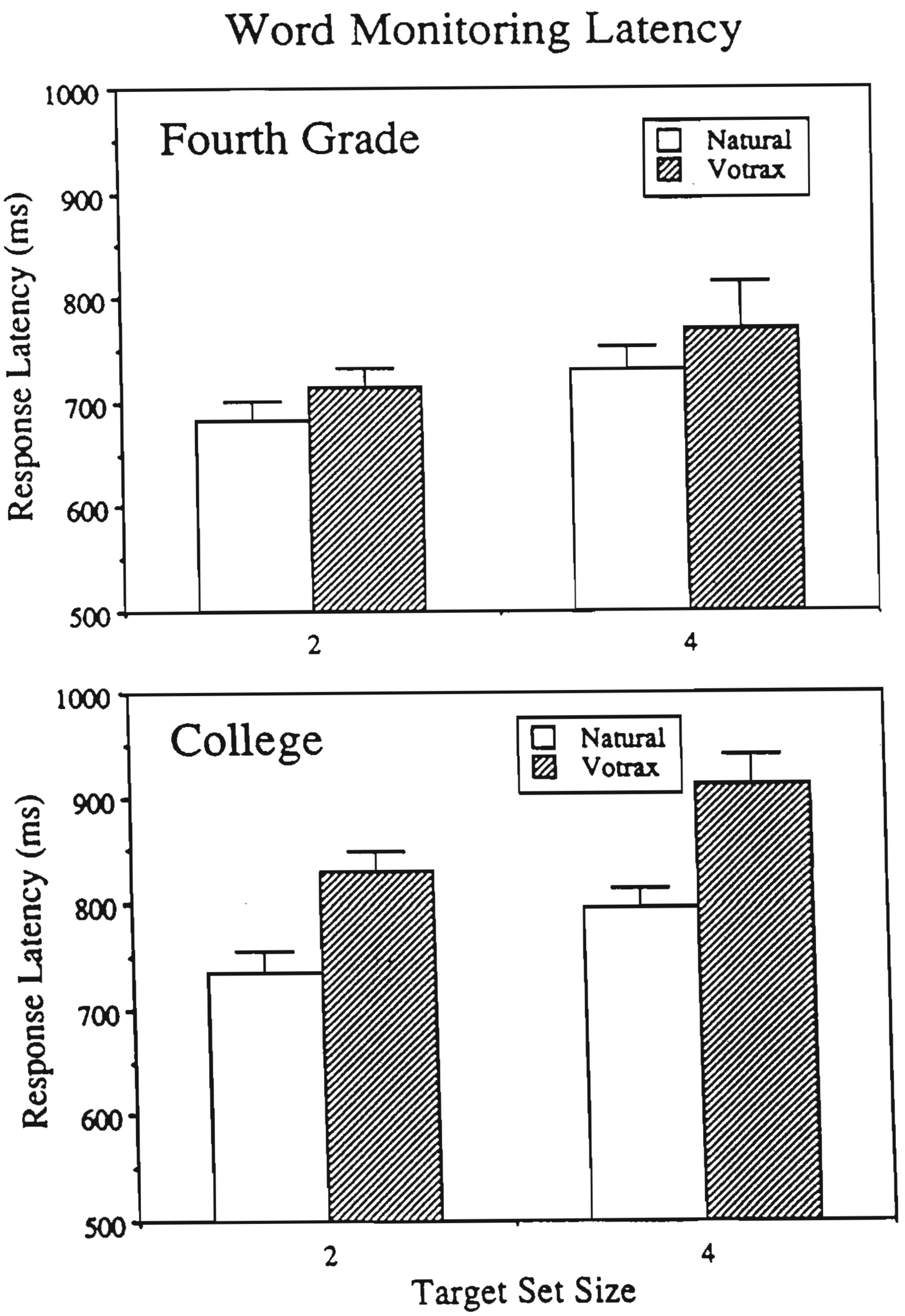


Figure 2. Word monitoring latencies (in ms) as a function of target set size for fourth grade (upper panel) and college-level (lower panel) passages. Open bars show the latencies for passages of natural speech; the striped bars show the latencies for passages of synthetic speech (From Lively et al., 1990).

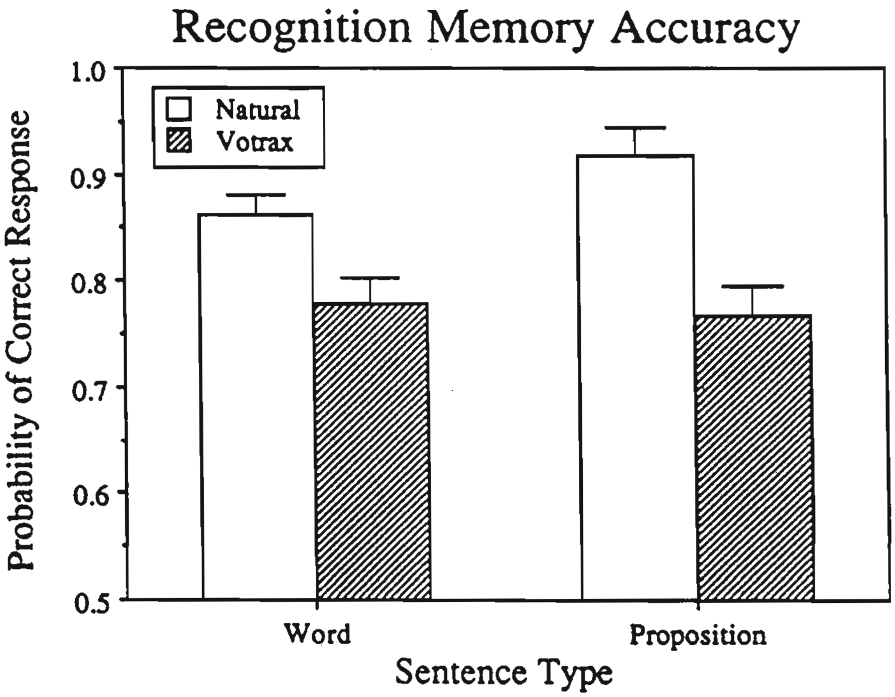


Figure 3. Recognition memory accuracy for words and propositions as a function of voice. Natural speech is shown in the open bars; Votrax synthetic speech is shown in the striped bars (From Lively et al., 1990).

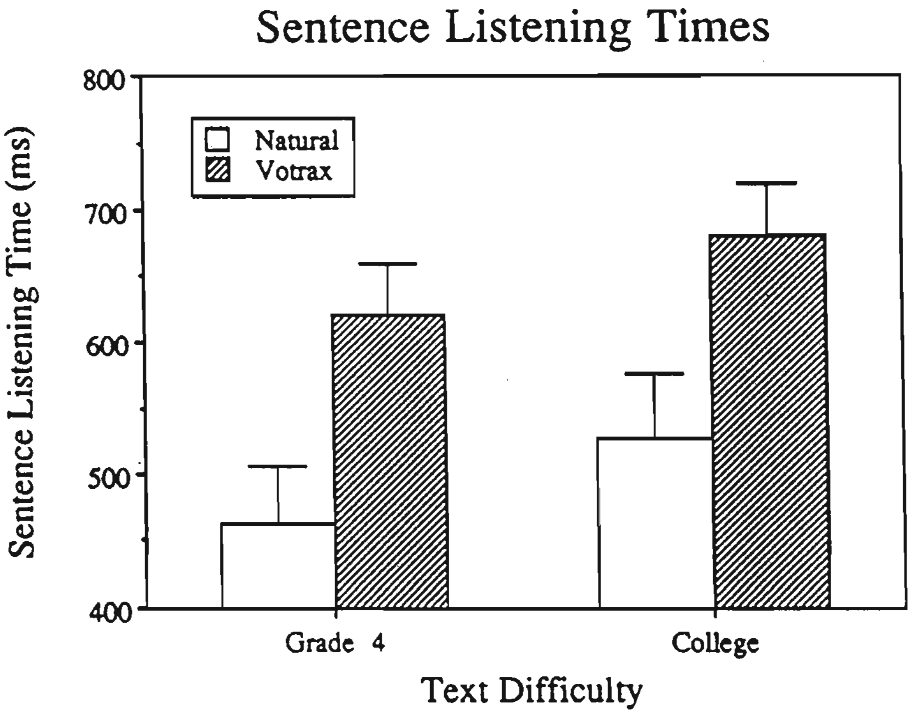


Figure 4. Sentence-by-sentence listening times as a function of voice. The bars on the left display the mean latencies for the fourth-grade passages; the bars on the right display the mean latencies for the college-level passages. Open bars show the data for natural speech; striped bars show the data for Votrax synthetic speech (From Lively et al., 1990).

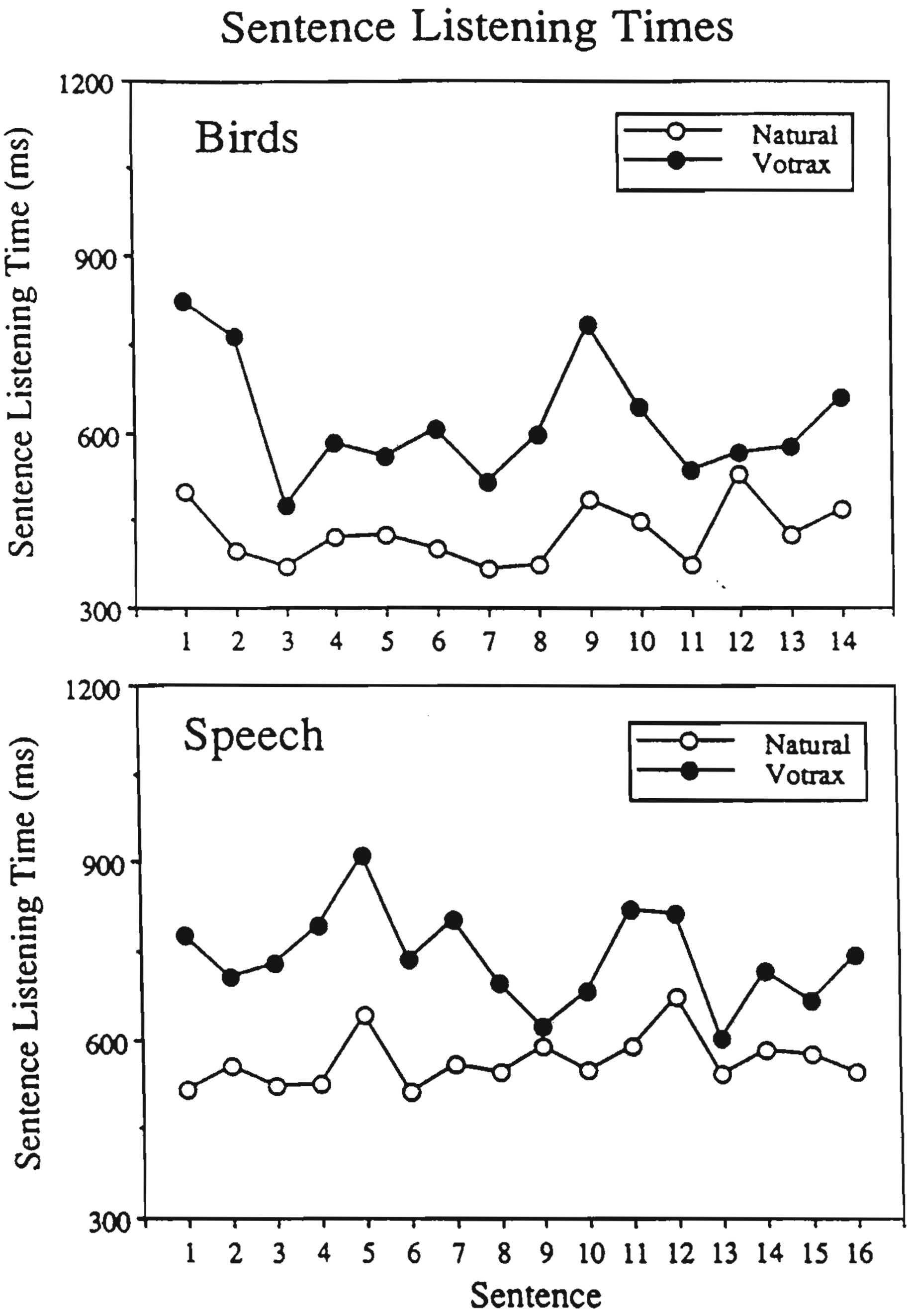


Figure 5. Sentence-by-sentence listening times for two representative passages for the individual sentences. The top panel shows data for a fourth-grade passage about birds; the bottom panel shows data for a college-level passage about speech. Natural speech is shown by the open circles; Votrax synthetic speech is shown by the filled circles (From Lively et al., 1990).

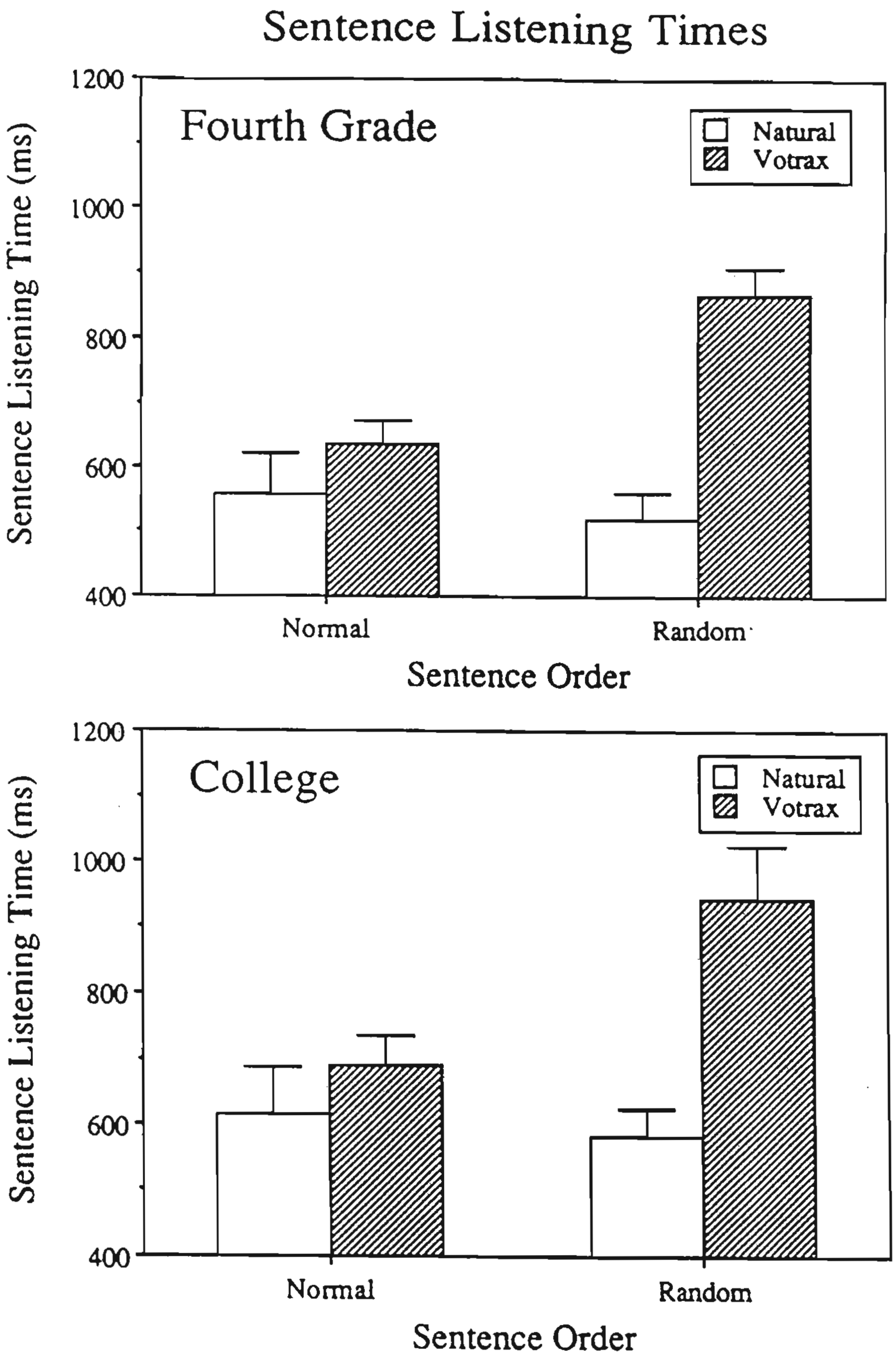


Figure 6. Sentence-by-sentence listening times for natural speech (open bars) and synthetic speech (striped bars) in normal and random sentences orders. The top panel shows data for fourth-grade passages; the bottom panel shows data for the college-level passages (From Ralston et al., 1991).

Table 1

Experiment 1 (Word Monitoring):
Correlations of MRT accuracy and comprehension measures.

Predicted Variable	Correlation	Probability
Word Monitoring		
Probability of Correct Detection	0.31	0.01
Response Latency	-0.34	<0.01
Recognition Memory		
Word Recognition	0.46	<0.01
Proposition Recognition	0.47	<0.01

Table 2

Experiment 2 (Listening Time):
Correlations of MRT accuracy and comprehension measures.

Predicted Variable	**Correlation**	**Probability**
Response Latency	-0.40	0.03
Recognition Memory		
Word Recognition	0.55	<0.01
Proposition Recognition	0.63	<0.01

Third, from the correlational analyses, we found that comprehension performance on passages of connected speech can be predicted only moderately well from standardized tests of segmental intelligibility using isolated monosyllabic words. This finding reinforces the belief that comprehension involves something more than just phoneme perception and word recognition. While segmental intelligibility contributes to the early sensory encoding of the speech waveform into a sequence of discrete phonemes, there appears to be a great deal more information used in comprehension that is abstract in nature and not part of the acoustic-phonetic properties of the signal.

In terms of directions for future research, our findings on comprehension suggest that more effort should be focused on the study of spoken language comprehension, particularly in terms of developing reliable experimental methods for measuring the online aspects of the process. The two methods we have used, the word-monitoring task and the self-paced sentence-by-sentence listening task, have provided us with important new information about the time course of comprehension that is difficult to obtain with either post-perceptual measures based on accuracy or more traditional measures of speech intelligibility using forced-choice tests such as the MRT or DRT. Spoken language comprehension makes use of several sources of knowledge that are closely dependent on each other. Although comprehension is complex and there are very few reliable assessment techniques at the present time, this does not mean that we cannot study the process experimentally. The online techniques that we have developed recently are the first steps in our program of research on the comprehension of connected speech. Given the extensive amount of work already done on segmental intelligibility, we feel that the problem of comprehension is a very fruitful area for future research.

References

Gough, P.B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, **4**, 107-111.

Larkey, L.S., & Danly, M. (1983). Fundamental frequency and sentence comprehension. *MIT Speech Group Working Papers, Vol. II.* Cambridge, MA: Research Laboratory of Electronics, Massachusetts Institute of Technology.

Lively, S.E., Ralston, J.V., Pisoni, D.B., & Rivera, S.M. (1990). Some effects of text structure on the comprehension of natural and synthetic speech. *Research on Speech Perception Progress Report No. 16.* Bloomington, IN: Speech Research Laboratory, Indiana University.

Logan, J.S., Greene, B.G., & Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566-581.

Pisoni, D.B., & Dedina, M.J. (1986). Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report. *Research on Speech Perception Progress Report No. 12.* Bloomington, IN: Speech Research Laboratory, Indiana University.

Pisoni, D.B., Manous, L.M., & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on sentence verification of sentences controlled for intelligibility. *Computer Speech and Language*, **2**, 303-320.

Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G., & Mullennix, J.W. (1991, in press). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*, **33**(4).

III. Publications

Papers Published:

Charles-Luce, J., Luce, P.A., & Cluff, M. (1990). Retroactive influence of syllable neighborhoods. In Altmann, G.T.M. (Ed.), *Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspectives*. Cambridge, MA: MIT Press, 173-183.

Connine, C.M., & Mullennix, J.W. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 1084-1096.

Davis, S., & Napoli, D.J. (1990). The destiny of second conjugation verbs in Romance. *Probus*, **2**, 125-166.

Davis, S. (1990). An argument for the underspecification of coronal in English. *Linguistic Inquiry*, **21**, 301-306.

Davis, S. (1990). Coronals and the phonotactics of nonadjacent consonants in English. In Paradis, C. & Prunet, J.-F. (Eds.), *The Special Status of Coronals*. New York, NY: Academic Press, 49-60.

Davis, S. (1990). Italian onset structure and the distribution of "il" and "lo". *Linguistics*, **28**, 43-55.

Dedina, M.J., & Nusbaum, H.C. (1991). PRONOUNCE: A program for pronunciation of new words by analogy. *Computer Speech and Language*, **5**, 55-64.

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**, (1), 152-162.

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

Johnson, K. (1990). Contrast and normalization in vowel perception. *Journal of Phonetics*, **18**, 229-254.

Johnson, K., Pisoni, D.B., & Bernacki, R.H. (1990). Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica*, **47**, 215-237.

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**(2), 874-886.

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

Makhoul, J. (Chairman), Crystal, T., Green, D., Hogan, D., McAulay, D., Pisoni, D., Sorkin, R., & Stockham, D. (1989). *Removal of Noise from Noise-Degraded Speech Signals*. Washington, D.C.: National Academy Press.

Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, **47**(4), 379-390.

Pisoni, D.B. (1989). Perceptual evaluation of synthetic speech: A tutorial review. *ESCA Tutorial on Speech Input/Output Assessment and Speech Databases*. Noordwijkerhout, The Netherlands: European Speech Communication Association, 1-13.

Pisoni, D.B., Greene, B.G., & Logan, J.S. (1989). An overview of ten years of research on the perception of synthetic speech. *ESCA Workshop on Speech Input/Output Assessment and Speech Databases*. Noordwijkerhout, The Netherlands: European Speech Communication Association, 1-4.

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*. Tokyo: The Acoustical Society of Japan, 1399-1407.

Pisoni, D.B., & Garber, E.E. (1991). Lexical memory in visual and auditory modalities: A case for a common lexicon. *Proceedings of the 1990 International Conference on Spoken Language Processing, Kobe, Japan*. Tokyo: The Acoustical Society of Japan, 401-404.

Pisoni, D.B., & Greene, B.G.(1990). The role of cognitive factors in the perception of synthetic speech. In Fujisaki, H. (Ed.), *International Symposium on International Coordination and Standardization of Speech Database and Assessment Techniques for Speech Input/Output*, Kobe, Japan, 3-25.

Pisoni, D.B., Ralston, J.V., & Lively, S.E. (1990). Some new directions in research on comprehension of synthetic speech. In Fujisaki, H. (Ed.), *International Symposium on International Coordination and Standardization of Speech Databse and Assessment Techniques for Speech Input/Output*, Kobe, Japan, 29-42.

Pisoni, D.B. (1991). Modes of processing speech and nonspeech signals. In Mattingly, I.G. and Studdert-Kennedy, M. (Eds.), *Modularity and the Motor Theory of Speech Perception*. Hillsdale, NJ: Erlbaum, 225-238.

Manuscripts Accepted for Publication (In Press):

Charles-Luce, J. The effects of semantic context on voicing neutralization. *Phonetica*.

Davis, S., & Summers, W.V. Vowel length and closure duration in word-medial VC sequences. *Journal of Phonetics*.

Goldinger, S.D., Pisoni, D.B., & Luce, P.A. Speech perception: Research and theory. In Lass, N.J. (Ed.), *Principles of Experimental Phonetics*. Toronto, Canada: B.C. Decker.

Humes, L.E., Nelson, K.J., & Pisoni, D.B. Recognition of synthetic speech by hearing-impaired elderly listeners. *Journal of Speech and Hearing Research*.

Johnson, K. Review of Speech Physiology, Speech Perception, and Acoustic Phonetics, by P. Lieberman and S.E. Blumstein. *Journal of Phonetics*.

Lively, S.E., Pisoni, D.B., & Logan, J.S. Some effects of training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura (Ed.), *Speech Perception, Production and Linguistic Structure*. Tokyo: Ohmsha Publishing.

Mullennix, J.W., Goldinger, S.D., & Pisoni, D.B. Some characteristics of talker normalization. In Charles-Luce, J., Luce, P.A., & Sawusch, J.R. (Eds.), *Theories in Spoken Language: Perception, Production and Development*. Norwood, NJ: Ablex.

Pisoni, D.B., Logan, J.S. & Lively, S.E. Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In Nusbaum, H. & Goodman, J. (Eds.), *Development of Speech Perception: The Transition from Recognizing Speech Sounds to Spoken Words*. Cambridge, MA: MIT Press.

Pisoni, D.B. Some Comments on Talker Normalization in Speech Perception. In Tohkura, Y., Vatikiotis-Bateson, E. & Sagisaka, Y. (Eds.), *Speech Perception, Production and Linguistic Structure*. Tokyo: Ohmsha Publishing.

Pisoni, D.B., Johnson, K., & Bernacki, R.H. Effects of alcohol on speech. *Proceedings of the Human Factors Society*. Santa Monica, CA: Human Factors Society.

Ralston, J.V., Pisoni, D.B., & Mullennix, J.W. Comprehension of synthetic speech produced by rule. In Bennett, R., Syrdal, A., & Greenspan, S. (Eds.), *Behavioral Aspects of Speech Technology: Theory and Applications*. New York, NY: Elsevier.

Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G., & Mullennix, J.W. Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*.

Tanford, J.A., Pisoni, D.B., & Johnson, K. Novel scientific evidence of intoxication: Acoustic analysis of voice recordings from the Exxon-Valdez. *Journal of Criminal Law and Criminology*.