RESEARCH ON SPOKEN LANGUAGE PROCESSING

Progress Report No. 18
(1992)

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-00111-16

and

National Institutes of Health
Training Grant No. DC-00012-14

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No.18 (1992)

Table of Contents

INTRODUCTION

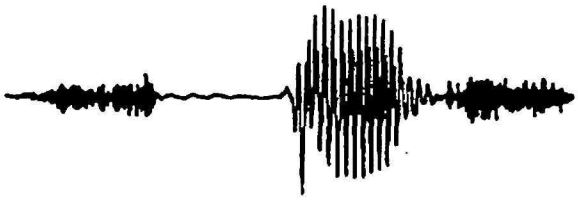
This is the eighteenth annual report summarizing research activities on speech perception and spoken language processing carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize various accomplishments over the past year and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief summaries of "on-going" research projects in the laboratory. From time to time, we also have included new information on instrumentation and software developments when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating 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 spoken language processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception and spoken language processing and we would be grateful if you 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
Phone: (812) 855-1155, 855-1768
FAX: (812)855-4691
E-mail: (BITNET) "PISONI@INDIANA"
E-mail: (INTERNET) "PISONI@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 for further information.

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 SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Long-term Memory in Speech Perception: Some New Findings on Talker Variability, Speaking Rate and Perceptual Learning[1]

David B. Pisoni

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

[1]This research was supported by NIDCD Research Grant DC-00111-17 to Indiana University in Bloomington. I thank Steve Goldinger, Scott Lively, Lynne Nygaard, Mitchell Sommers and Thomas Palmeri for their help and collaboration in various phases of this research program. This chapter to appear in a Festschrift in honor of Hiroya Fujisaki. Edited by K. Hirose, S. Kiritani & G. Fant. Published by Elsevier North Holland (1993).

Abstract

This chapter summarizes results from recent studies on the role of long-term memory in speech perception and spoken word recognition. Experiments on talker variability, speaking rate and perceptual learning provide evidence of implicit memory for very fine perceptual details of speech. Listeners apparently encode specific attributes of the talker's voice and speaking rate into long-term memory. Acoustic-phonetic variability does not appear to be "lost" as a result of phonetic analysis. The process of "perceptual normalization" in speech perception may therefore entail encoding of specific instances or "episodes" of the stimulus input and the operations used in perceptual analysis. These perceptual operations may reside in a "procedural memory" for a specific talker's voice. Taken together, the present set of findings are consistent with non-analytic accounts of perception, memory and cognition which emphasize the contribution of episodic or exemplar-based encoding in long-term memory. The results from these studies also raise questions about the traditional dissociation in phonetics between the linguistic and indexical properties of speech. Listeners apparently retain non-linguistic information in long-term memory about the speaker's gender, dialect, speaking rate and emotional state, attributes of speech signals that are not traditionally considered part of phonetic or lexical representations of words. These properties influence the initial perceptual encoding and retention of spoken words and therefore should play an important role in theoretical accounts of how the nervous system maps speech signals onto linguistic representations in the mental lexicon.

Long-term Memory in Speech Perception: Some New Findings on Talker Variability, Speaking Rate and Perceptual Learning

Introduction

Those of us who work in the field of human speech perception owe a substantial intellectual debt to Professor Hiroya Fujisaki who has contributed in many important ways to our current understanding of the speech mode and the underlying perceptual mechanisms. His theoretical and empirical work in the late 1960's brought the study of speech perception directly into the main stream of cognitive psychology (Fujisaki & Kawashima, 1969). In particular, his pioneering research and modeling efforts on categorical perception inspired a large number of empirical studies on issues related to coding processes and the contribution of short-term memory to speech perception and categorization.

My own research in the early 1970's was directly motivated by his proposal of the differential roles of auditory and phonetic memory codes in the perception of consonants and vowels. The studies that I carried out at that time demonstrated that it was possible to account for categorical and non-categorical modes of perception in terms of coding and memory processes in short-term memory without recourse to the traditional theoretical accounts that were very popular at the time (Pisoni, 1973). These accounts of speech perception drew heavily on claims for the existence of a specialized perceptual mode for speech sounds that was distinct from other perceptual systems (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967).

Professor Fujisaki's efforts along with other results were largely responsible for integrating the study of speech perception with other closely related fields of cognitive psychology such as perception, memory and attention. By the mid 1970's, the field of speech perception became a legitimate topic for experimental psychologists to study (Pisoni, 1978). This was clearly an exciting time to be working in speech perception. Before these developments, speech perception was an exotic field representing the intersection of electrical engineering, speech science, linguistics, and traditional experimental psychology.

At the present time, the field of speech perception has evolved into an extremely active area of research with scientists from many different disciplines working on a common set of problems [56]. Many of the current problems revolve around issues of representation and the role of coding and memory systems in spoken language processing, topics that Professor Fujisaki has written about in some detail over the years. The recent meetings of the ICSLP in Kobe and Banff demonstrate a convergence on a "core" set of basic research problems in the field of spoken language processing--problems that are inherently multi-disciplinary in nature. As many of us know from personal experiences, Professor Fujisaki was among the very first to recognize these common issues in his research and theoretical work over the years. The success of the two ICSLP meetings is due, in part, to his vision for a unified approach to the field of spoken language processing.

In this contribution, I am delighted to have the opportunity to summarize some recent work from my laboratory that deals with the role of long-term memory in speech perception and spoken word recognition. Much of our research over the last few years has turned to questions concerning perceptual learning and the retention of information in permanent long-term memory. This trend contrasts with the earlier work in the 1970's which was concerned almost entirely with short-term memory. We have also focused much of our current research on problems of spoken word recognition in contrast to earlier studies which were concerned with phoneme perception. We draw a distinction

between phoneme perception and spoken word recognition. While phoneme perception is assumed to be a component of the word recognition process, the two are not equivalent. Word recognition entails access to phonological information stored in long-term memory, whereas phoneme perception relies almost exclusively on the recognition of acoustic cues contained in the speech signal.

Our interests are now directed at the interface between speech perception and spoken language comprehension which naturally has led us to problems of lexical access and the structure and organization of sound patterns in the mental lexicon (Pisoni, Nusbaum, Luce & Slowiaczek, 1985). Findings from a variety of studies suggest that very fine details in the speech signal are preserved in the human memory system for relatively long periods of time (Goldinger, 1992). This information appears to be used in a variety of ways to facilitate perceptual encoding, retention and retrieval of information from memory. Many of our recent investigations have been concerned with assessing the effects of different sources of variability in speech perception (Sommers, Nygaard & Pisoni, 1992; Nygaard, Sommers & Pisoni, 1992). The results of these studies have encouraged us to reassess our beliefs about several long-standing issues such as acoustic-phonetic invariance and the problems of perceptual normalization in speech perception (Pisoni, 1992).

In the sections below, I will briefly summarize the results from several recent studies that deal with talker variability, speaking rate, and perceptual learning. These findings have raised a number of important new questions about the traditional dissociation between the linguistic and indexical properties of speech signals and the role that different sources of variability play in speech perception and spoken word recognition. For many years, linguists and phoneticians have considered attributes of the talker's voice-- what Ladefoged refers to a the "personal" characteristics of speech-- to be independent of the linguistic content of the talker's message (Ladefoged, 1975; Laver & Trudgill, 1979). The dissociation of these two parallel sources of information in speech may have served a useful function in the formal linguistic analysis of language when viewed as an idealized abstract system of symbols. However, the artificial dissociation has at the same time created some difficult problems for researchers who wish to gain a detailed understanding of how the nervous system encodes speech signals and represents them internally and how real speakers and listeners deal with the enormous amount of acoustic variability in speech.

Experiments on Talker Variability in Speech Perception

A series of novel experiments have been carried out to study the effects of different sources of variability on speech perception and spoken word recognition (Pisoni, 1990). Instead of reducing or eliminating variability in the stimulus materials, as most researchers had routinely done in the past, we specifically introduced variability from different talkers and different speaking rates to study their effects on perception (Pisoni, 1992). Our research on talker variability began with the observations of Mullennix, Pisoni & Martin (1989) who found that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers that were used to generate the test words in the stimulus ensemble. In one condition, all the words in a test list were produced by a single talker; in another condition, the words were produced by 15 different talkers, including male and female voices. The results which are shown in Figure 1 were very clear. Across three signal-to-noise ratios, identification performance was always better for words that were produced by a single talker than words produced by multiple talkers. Trial-to-trial variability in the speaker's voice apparently affects recognition performance. This pattern was observed for both high-density (i.e., confusable) and low-density (i.e., non-confusable) words. These findings replicated results originally found by Peters (1955) and Creelman (1957) back in the 1950's and suggested to us that the perceptual system must

engage in some form of "recalibration" each time a new voice is encountered during the set of test trials.

Insert Figure 1 about here

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

Insert Table I about here

In another series of experiments we used a speeded classification task (Palmeri, Goldinger & Pisoni, 1993) to assess whether attributes of a talker's voice were perceived independently of the phonetic form of the words (Mullennix & Pisoni, 1990). Subjects were required to attend selectively to one stimulus dimension (i.e., voice) while simultaneously ignoring another stimulus dimension (i.e., phoneme). Figure 2 shows the main findings. Across all conditions, we found increases in interference from both dimensions when the subjects were required to attend selectively to only one of the stimulus dimensions. The pattern of results suggested that words and voices were processed as integral dimensions; the perception of one dimension (i.e., phoneme) affects classification of the other dimension (i.e., voice) and vice versa, and subjects cannot selectively ignore irrelevant variation on the non-attended dimension. If both perceptual dimensions were processed separately, as we originally assumed, we should have found little if any interference from the non-attended dimension which could be selectively ignored without affecting performance on the attended dimension. Not only did we find mutual interference suggesting that the two sets of dimensions, voice and phoneme, are perceived in a mutually dependent manner but we also found that the pattern of interference was asymmetrical. It was easier for subjects to ignore irrelevant variation in the phoneme dimension when their task was to classify the voice dimension than it was to ignore the voice dimension when they had to classify the phonemes.

Insert Figure 2 about here

The results from the perceptual experiments were surprising given our prior assumption that the indexical and linguistic properties of speech were perceived independently. To study this problem further, we carried out a series of memory experiments to assess the mental representation of speech in long-term memory. Experiments on serial recall of lists of spoken words by Martin, Mullennix, Pisoni

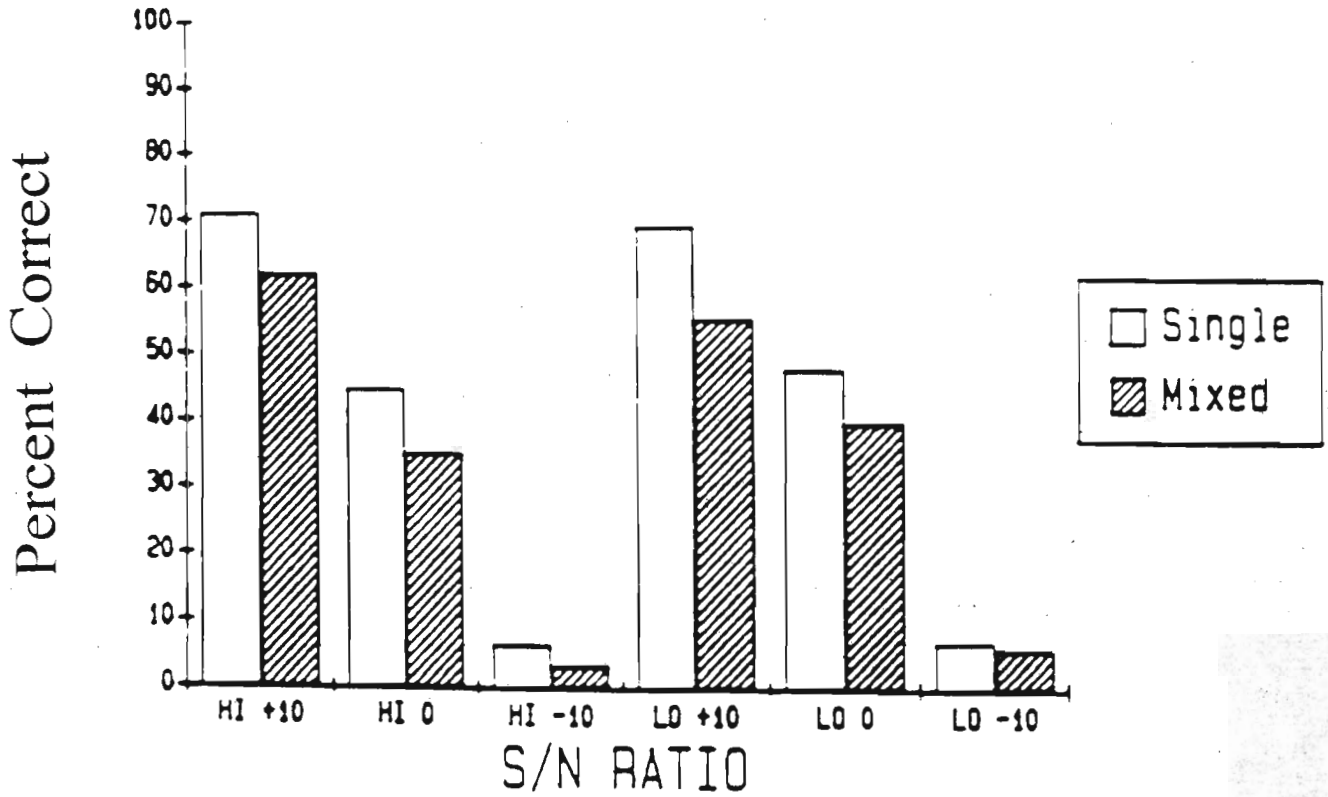


Figure 1. Overall mean percent correct performance collapsed over subjects for single- and mixed-talker conditions as a function of high- and low-density words and S/N ratio (from Mullennix et al., 1989).

Table I

Mean response latency (ms) for correct responses for single- and mixed-talker conditions as a function of lexical density (from Mullennix et al., 1989).

	Density	
	High	**Low**
Single talker	611.2	605.7
Mixed talker	677.2	679.4

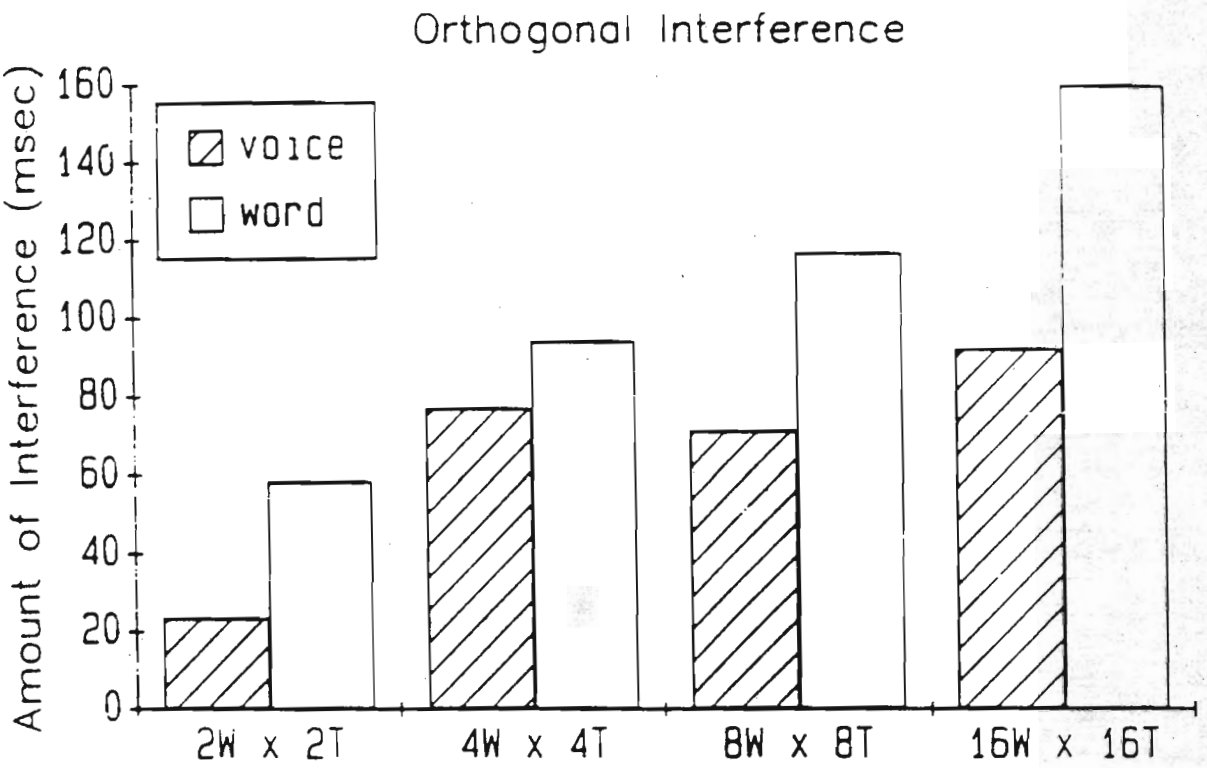


Figure 2. The amount of orthogonal interference (in milliseconds) across all stimulus variability conditions as a function of word and voice dimensions (from Mullennix and Pisoni, 1990).

and Summers (1989) and Goldinger, Pisoni & Logan (1991) demonstrated that specific details of a talker's voice are also encoded into long-term memory. Using a continuous recognition memory procedure, Palmeri et al. (1993) found that detailed episodic information about a talker's voice is also encoded in memory and is available for explicit judgments even when a great deal of competition from other voices is present in the test sequence. Palmeri et al.'s results are shown in Figure 3. The top panel shows the probability that an item was correctly recognized as a function of the number of talkers in the stimulus set. The bottom panel shows the probability of a correct recognition across different stimulus lags of intervening items. In both cases, the probability of correctly recognizing a word as "old" (filled circles) was greater if the word was repeated in the same voice than if it was repeated in a different voice of the same gender (open squares) or a different voice of a different gender (open triangles).

Finally, in another set of experiments, Goldinger (1992) found very strong evidence of implicit memory for attributes of a talker's voice which persists for a relatively long period of time after perceptual analysis has been completed. His results are shown in Figure 4. Goldinger also showed that the degree of perceptual similarity affects the magnitude of the repetition effect in memory for identical voices suggesting that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations.

Insert Figures 3 and 4 about here

Taken together, our findings on the effects of talker variability in perception and memory tasks provide support for the proposal that detailed perceptual information about a talker's voice is preserved in some type of perceptual representation system (PRS) (Schacter, 1990) and that these attributes are encoded implicitly into long-term memory. At the present time, it is not clear whether there is one composite representation in memory or whether these different sets of attributes are encoded in parallel in separate representations (Eich, 1982; Hintzman, 1986). It is also not clear whether spoken words are encoded and represented in memory as a sequence of abstract symbolic phoneme-like units along with much more detailed episodic information about specific instances and the processing operations used in perceptual analysis. These are important questions for future research on spoken word recognition.

Experiments on the Effects of Speaking Rate

Another new series of experiments has been carried out to examine the effects of speaking rate on perception and memory. These studies, which were designed to parallel the earlier experiments on talker variability, have also shown that the perceptual details associated with differences in speaking rate are not lost as a result of perceptual analysis. In one experiment, Sommers, Nygaard & Pisoni (1992) found that words produced at different speaking rates (i.e., fast, medium and slow) were identified more poorly than the same words produced at only one speaking rate. These results were compared to another condition in which differences in amplitude were varied randomly from trial to trial in the test sequences. In this case, identification performance was not affected by variability in overall level. The results from both conditions are shown in Figures 5 and 6.

Insert Figures 5 and 6 about here

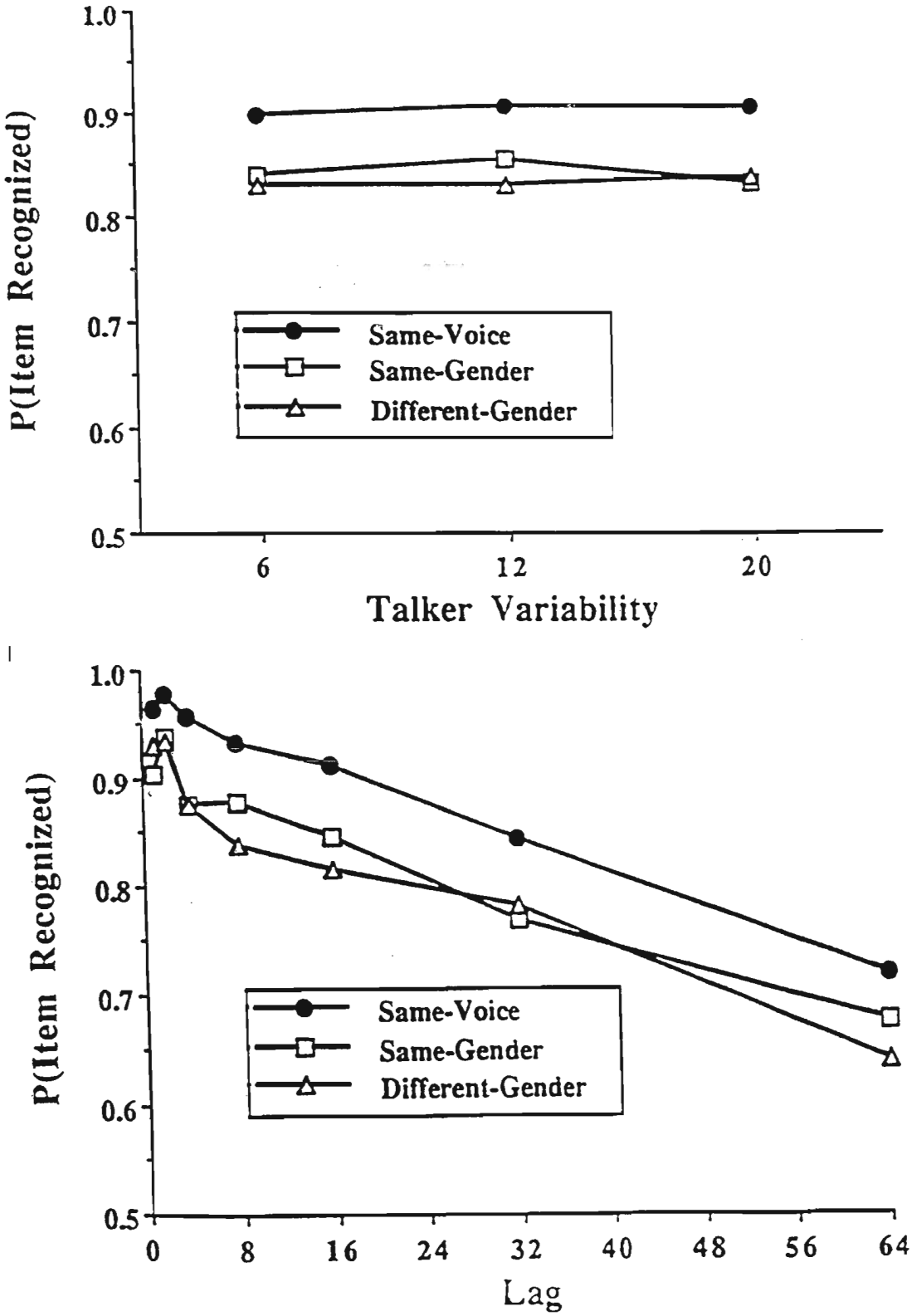


Figure 3. Probability of correctly recognizing old items in a continuous recognition memory experiment. In both panels, recognition for same-voice repetitions is compared to recognition for different-voice/same-gender and different-voice/different-gender repetitions. The upper panel displays item recognition as a function of talker variability, collapsed across values of lag; the lower panel displays item recognition as a function of lag, collapsed across levels of talker variability (from Palmeri et al., 1993).

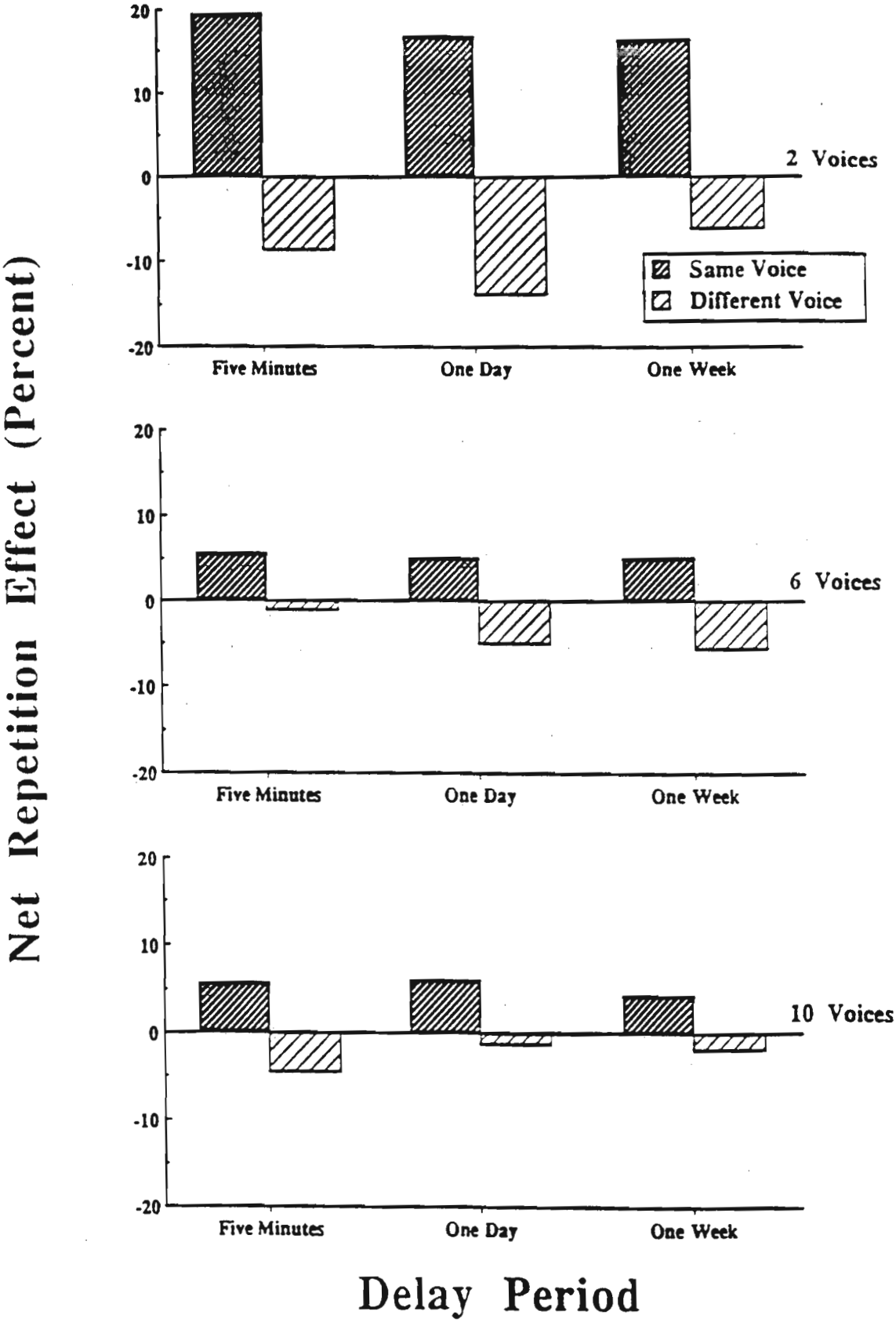


Figure 4. Net repetition effects observed in perceptual identification as a function of delay between sessions and repetition voice (from Goldinger, 1992).

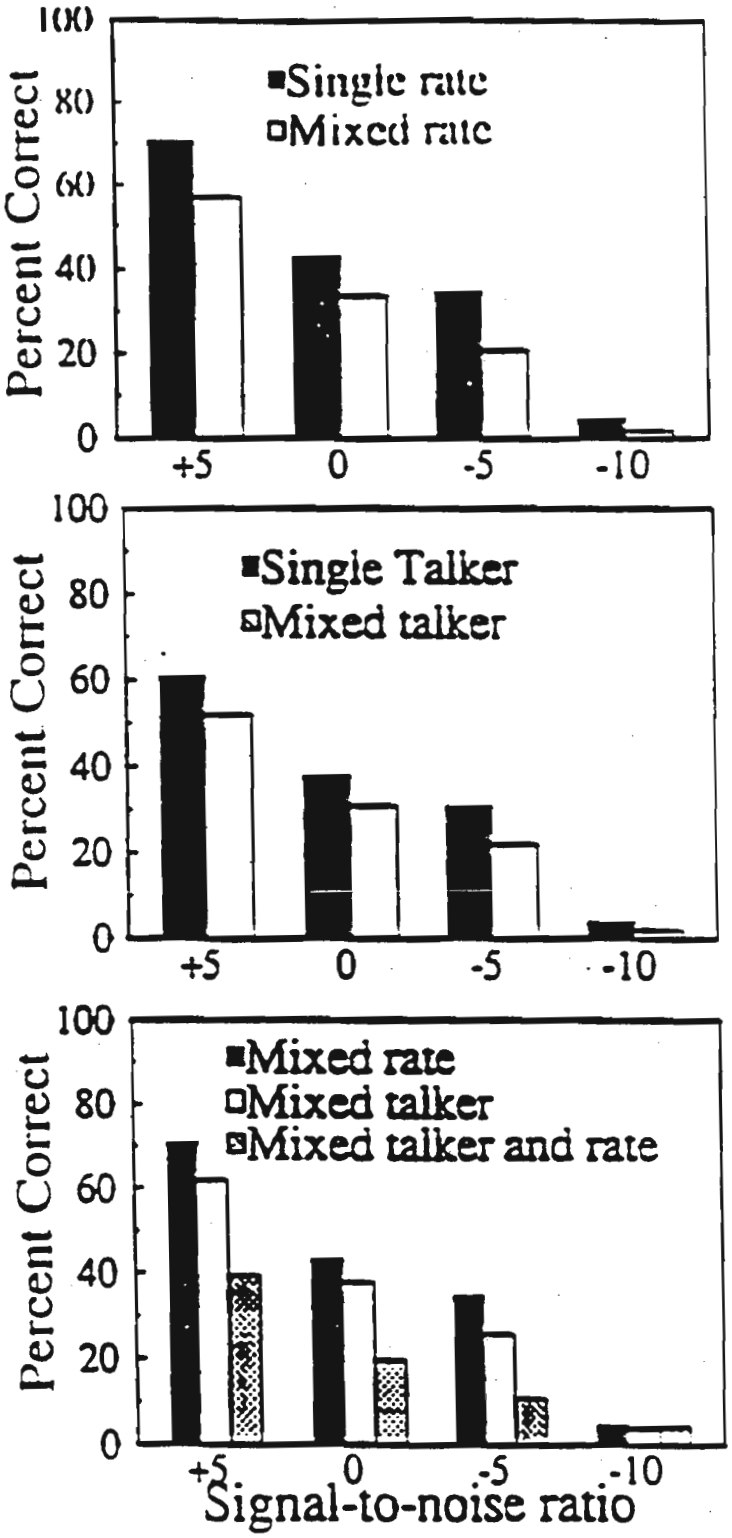


Figure 5. Effects of talker, rate, and combined talker and rate variability on perceptual identification (from Sommers et al., 1992).

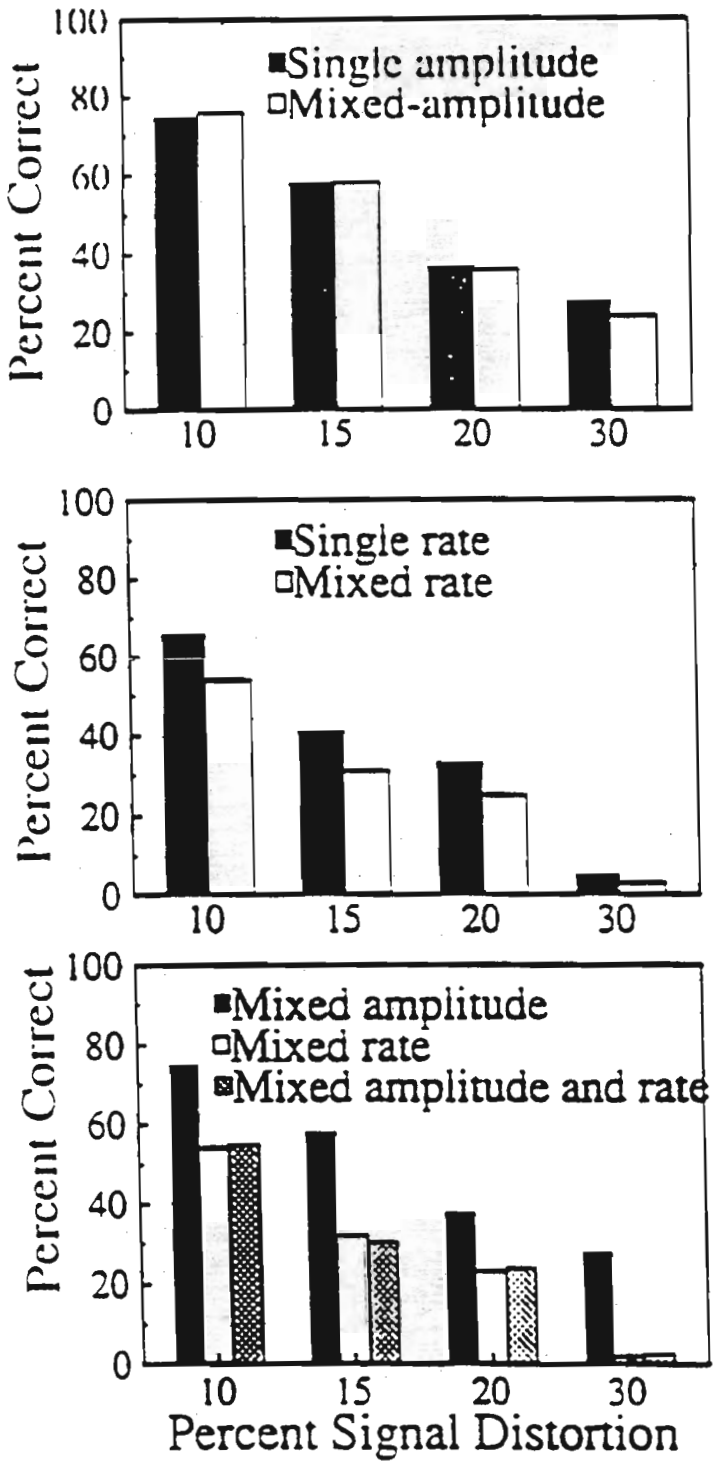


Figure 6. Effects of amplitude, rate, and combined amplitude and rate variability on perceptual identification (from Sommers et al., 1992).

Other experiments on serial recall have also been completed to examine the encoding and representation of speaking rate in memory. Nygaard, Sommers & Pisoni (1992) found that subjects recall words from lists produced at a single speaking rate better than the same words produced at several different speaking rates. Interestingly, the differences appeared in the primacy portion of the serial position curve suggesting greater difficulty in the transfer of items into long-term memory. Differences in speaking rate, like those observed for talker variability in our earlier experiments, suggest that perceptual encoding and rehearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by low-level perceptual sources of variability. If these sources of variability were somehow "filtered out" or normalized by the perceptual system at relatively early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments.

Taken together with the earlier results on talker variability, the findings on speaking rate suggest that details of the early perceptual analysis of spoken words are not lost and apparently become an integral part of the mental representation of spoken words in memory. In fact, in some cases, increased stimulus variability in an experiment may actually help listeners to encode items into long-term memory (Goldinger et al., 1991; Nygaard et al., 1992). Listeners encode speech signals in multiple ways along many perceptual dimensions and the memory system apparently preserves these perceptual details much more reliably than researchers have believed in the past.

Experiments on Variability in Perceptual Learning

We have always maintained a strong interest in issues surrounding perceptual learning and development in speech perception (Aslin & Pisoni, 1980; Walley, Pisoni & Aslin, 1981). One reason for this direction in our research is that much of the theorizing that has been done in speech perception has focused almost entirely on the mature adult with little concern for the processes of perceptual learning and developmental change. This has always seemed to be a peculiar state of affairs because it is now very well established that the linguistic environment plays an enormous role in shaping and modifying the speech perception abilities of infants and young children as they acquire their native language (Jusczyk, 1993). Theoretical accounts of speech perception should not only describe the perceptual abilities of the mature listener but they should also provide some principled explanations of how these abilities develop and how they are selectively modified by the language learning environment (Jusczyk, 1993; Studdert-Kennedy, 1980).

One of the questions that we have been interested in deals with the apparent difficulty that adult Japanese listeners have in discriminating English /r/ and /l/ (Logan, Lively & Pisoni, 1991; Lively, Pisoni & Logan, 1992; Lively, Logan & Pisoni, 1993; Strange & Dittmann, 1984). Is the failure to discriminate this contrast due to some permanent change in the perceptual abilities of native speakers of Japanese or are the basic sensory and perceptual mechanisms still intact and only temporarily modified by changes in selective attention and categorization? Many researchers working in the field have maintained the view that the effects of linguistic experience on speech perception are extremely difficult, if not impossible, to modify in a short period of time. The process of "re-learning" or "re-acquisition" of phonetic contrasts is generally assumed to be very difficult-- it is slow, effortful and considerable variability has been observed among individuals in reacquiring sound contrasts that were not present in their native language (Strange & Dittmann, 1984).

We have carried out a series of laboratory training experiments to learn more about the difficulty Japanese listeners have in identifying English words containing /r/ and /l/ (Logan et al.,

1991). In these studies we have taken some clues from the literature in cognitive psychology on the development of new perceptual categories and have designed our training procedures to capitalize on the important role that stimulus variability plays in perceptual learning (Posner & Keele, 1986). In the training phase of our experiments, we used a set of stimuli that contained a great deal of variability. The phonemes /r/ and /l/ appeared in English words in several different phonetic environments so that listeners would be exposed to different contextual variants of the same phoneme in different positions. In addition, we created a large database of words that were produced by several different talkers including both men and women in order to provide the listeners with exposure to a wide range of stimulus tokens.

A pretest-posttest design was used to assess the effects of the training procedures. Subjects were required to come to the laboratory for daily training sessions in which immediate feedback was provided after each trial. We trained a group of six Japanese listeners using a two-alternative forced-choice identification task. The stimulus materials consisted of minimal pairs of English words that contrasted /r/ and /l/ in five different phonetic environments.

On each training trial, subjects were presented with 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 response 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 error, the minimal pair remained on the monitor and the stimulus word was repeated. In addition to the daily training sessions, subjects were also given a pretest and a posttest. At the end of the experiment, we also administered two additional tests of generalization. One test contained new words produced by one of the talkers used in training; the other test contained new words produced by a novel talker.

Identification accuracy improved significantly from the pretest to the posttest. Large and reliable effects of phonetic environment also were 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 also was observed. Subjects improved more in initial consonant clusters and in intervocalic position than in word-initial and word-final positions.

The training results also showed that subjects' performance improved as a function of training. The largest gain came after one week of training. The gain in the other weeks was slightly smaller. Each of the six subjects showed improvement, although large individual differences in absolute levels of performance were observed.

The tests of generalization provided an additional way of assessing the effectiveness of the training procedures. Subjects were presented with new words spoken by a familiar talker and 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 categorization task. In our first training study, accuracy was marginally greater for words produced by the old talker compared to the new talker. However, in a replication experiment using 19 mono-lingual Japanese listeners, we found a highly significant difference in performance on the generalization tests. The results of the generalization tests demonstrate the high degree of context sensitivity present in learning to perceive these contrasts: Listeners were sensitive to the voice of the talker producing the tokens as well as the phonetic environment in which the contrasts occurred. Thus, stimulus variability is useful in perceptual learning of complex multidimensional categories like speech because it serves to make the mental representations extremely robust over different acoustic transformations such as talker, phonetic

environment and speaking rate. In a high variability training procedure, like the one used by Logan et al., listeners are not able to focus their attention on only one set of criterial cues to learn the category structure for the phonemes /r/ and /l/. Listeners have to acquire detailed knowledge about different sources of variability in order to be able to generalize to new words and new talkers.

We have also been interested in another kind of perceptual learning, the tuning or adaptation that occurs when a listener becomes familiar with the voice of a specific talker (Nygaard, Summers & Pisoni, submitted). This particular kind of perceptual learning has not received very much attention in the past despite the obvious relevance to problems of speaker normalization, acoustic-phonetic invariance and the potential application to automatic speech recognition and speaker identification (Kakehi, 1992; Fowler, in press). Our search of the research literature on talker adaptation revealed only a small number of studies on this topic and all of them appeared in obscure technical reports from the mid 1950's. Thus, we decided to carry out a perceptual learning experiment in our own laboratory.

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

The results of the intelligibility experiment are shown in Figure 7 for two groups of subjects. We found that identification performance for the trained group was reliably better than the control group at each of the signal-to-noise ratios tested. The subjects who had heard novel words produced by familiar voices were able to recognize words in noise more accurately than subjects who received the same novel words produced by unfamiliar voices. Two other groups of subjects were also run in the intelligibility experiment as controls; however, these subjects did not receive any training and were therefore not exposed to any of the voices prior to hearing the same set of words in noise. One control group received the set of words presented to the trained experimental group; the other control group received the words that were presented to the trained control subjects. The performance of these two control groups was not only same but was equivalent to the intelligibility scores obtained by the trained control group. Only subjects in the experimental group who were explicitly trained on the voices showed an advantage in recognizing novel words produced by familiar talkers.

Insert Figure 7 about here

The findings from this perceptual learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by a familiar talker. Thus, speech perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that was obtained in an entirely different experimental task-- explicit voice recognition as compared to a speech intelligibility test in which novel words were mixed in noise and subjects identified the items explicitly from an open response set.

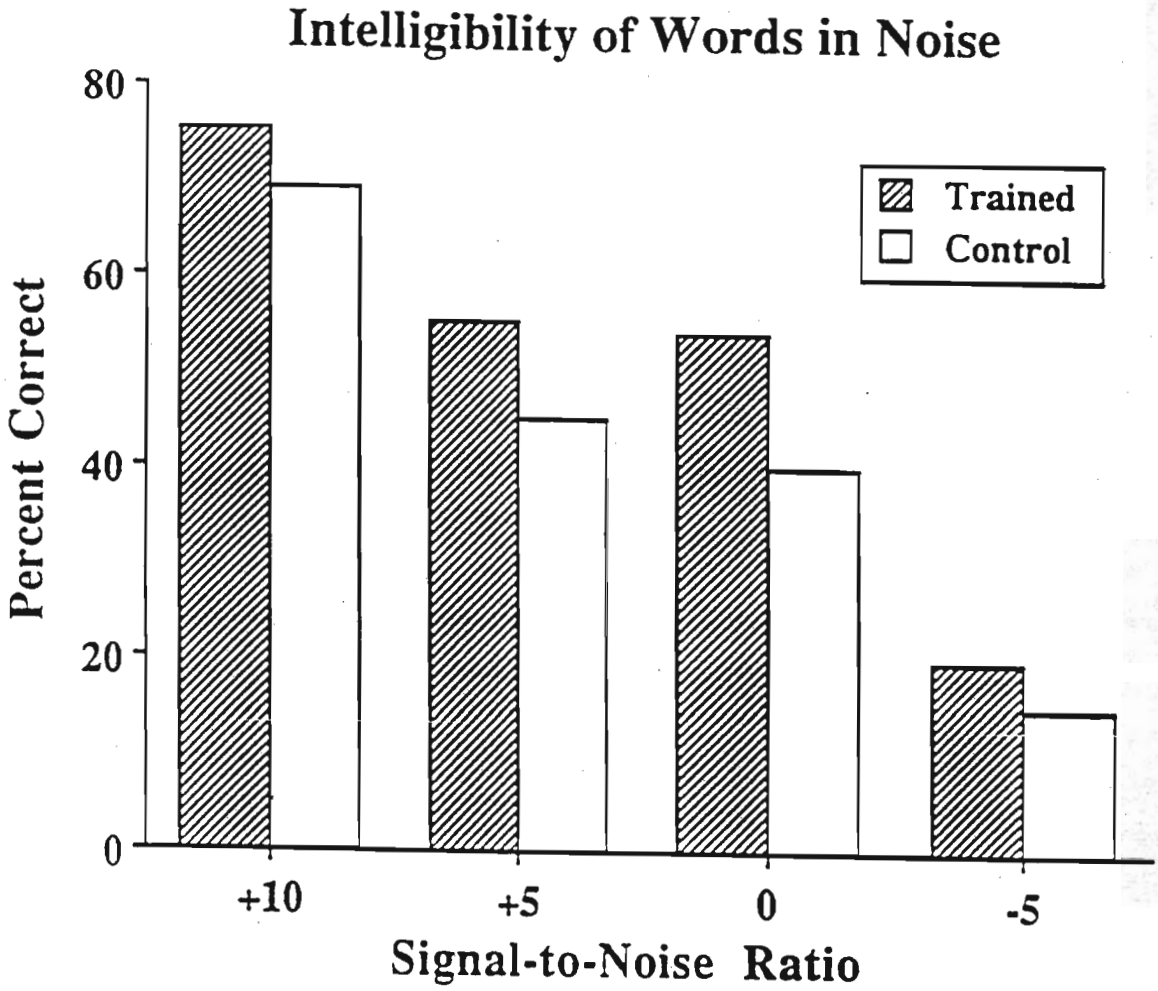


Figure 7. Mean intelligibility of words mixed in noise for trained and control subjects. Percent correct word recognition is plotted at each signal-to-noise ratio (from Nygaard et al., 1992).

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

Whatever the exact nature of this information or knowledge turns out to be, the important point here is that prior exposure to a talker's voice facilitates subsequent recognition of novel words produced by the same talkers. Such findings demonstrate a form of implicit memory for a talker's voice that is distinct from the retention of the individual items used and the specific task that was employed to familiarize the listeners with the voices (Schacter, 1992; Roediger, 1990). These findings provide additional support for the view that the internal representation of spoken words encompasses both a phonetic description of the utterance, as well as information about the structural description of the source characteristics of the specific talker. Thus, speech perception appears to be carried out in a "talker-contingent" manner; indexical and linguistic properties of the speech signal are apparently closely interrelated and are not dissociated in perceptual analysis as many researchers previously thought. We believe these talker-contingent effects may provide a new way to deal with some of the old problems in speech perception that have been so difficult to resolve in the past.

Abstractionist vs. Episodic Approaches to Speech Perception

The results we have obtained over the last few years raise a number of important questions about the theoretical assumptions or metatheory of speech perception which has been shared for many years by almost all researchers working in the field (Pisoni & Luce, 1986). Within cognitive psychology, the traditional view of speech perception can be considered among the best examples of what have been called abstractionist approaches to the problems of categorization and memory (Jacoby & Brooks, 1984). Units of perceptual analysis in speech were assumed to be equivalent to the abstract idealized categories proposed by linguists in their formal analyses of language structure and function. The goal of speech perception studies was to find the physical invariants in the speech signal that mapped onto the phonetic categories of speech (Studdert-Kennedy, 1976). Emphasis was directed at separating stable, relevant features from the highly variable, irrelevant features of the signal. An important assumption of this traditional approach to perception and cognition was the process of abstraction and the reduction of information in the signal to a more efficient and economical symbolic code (Posner, 1969; Neisser, 1976). Unfortunately, it became apparent very early on in speech perception research that idealized linguistic units, such as phonemes or phoneme-like units, were highly dependent on phonetic context and moreover that a wide variety of factors influenced their physical realization in the speech signal (Stevens, 1971; Klatt, 1986). Nevertheless, the search for acoustic invariance has continued in one way or another and still remains a central problem in the field today.

Recently, a number of studies on categorization and memory in cognitive psychology have provided evidence for the encoding and retention of episodic information and the details of perceptual analysis (Jacoby & Brooks, 1984; Brooks, 1978; Tulving & Schacter, 1990; Schacter, 1990). According to this approach, stimulus variability is considered to be "lawful" and informative to perceptual analysis

(Elman & McClelland, 1986). Memory involves encoding specific instances, as well as, the processing operations used in recognition (Kolers, 1973; Kolers, 1976b). The major emphasis of this view is on particulars, rather than abstract generalizations or symbolic coding of the stimulus input into idealized categories. Thus, the problems of variability and invariance found in speech perception can be approached in a different way by non-analytic or instance-based accounts of perception and memory with the emphasis on encoding of exemplars and specific instances of the stimulus environment rather than the search for physical invariants for abstract symbolic categories.

We believe that the findings from studies on nonanalytic cognition can be generalized to theoretical questions about the nature of perception and memory for speech signals and to assumptions about abstractionist representations based on formal linguistic analyses. When the criteria used for postulating episodic or non-analytic representations are examined carefully, it immediately becomes clear that speech signals display a number of distinctive properties that make them especially good candidates for this approach to perception and memory (Jacoby & Brooks, 1984; Brooks, 1978). These criteria which are summarized below can be applied directly to speech perception and spoken language processing.

High Stimulus Variability.

Speech signals display a great deal of variability primarily because of factors that influence the production of spoken language. Among these are within- and between-talker variability, changes in speaking rate and dialect, differences in social contexts, syntactic, semantic and pragmatic effects, as well as a wide variety of effects due to the ambient environment such as background noise, reverberation and microphone characteristics (Klatt, 1986). These diverse sources of variability consistently produce large changes in the acoustic-phonetic properties of speech and they need to be accommodated in theoretical accounts of speech perception.

Complex Category Relations.

The use of phonemes as perceptual categories in speech perception entails a set of complex assumptions about category membership which are based on formal linguistic criteria involving principles such as complementary distribution, free variation and phonetic similarity. The relationship between allophones and phonemes acknowledges explicitly the context-sensitive nature of the category relations that are used to define classes of speech sounds that function in similar ways in different phonetic environments. In addition, there is evidence for "trading relations" among cues to particular phonetic contrasts in speech. Acoustically different cues to the same contrast interact as a function of context.

Incomplete Information.

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

High Analytic Difficulty.

Speech sounds are inherently multidimensional in nature. They encode a large number of quasi-independent articulatory attributes that are mapped on to the phonological categories of a specific language. Because of the complexity of speech categories and the high acoustic-phonetic variability, the

category structure of speech is not amenable to simple hypothesis testing. As a consequence, it has been extremely difficult to formalize a set of explicit rules that can successfully map speech cues onto a set of idealized phoneme categories. Phoneme categories are also highly automatized. The category structure of a language is learned in a tacit and incidental way by young children. Because the criterial dimensional structures of speech are not typically available to consciousness, it is difficult to make many aspects of speech perception explicit to either children, adults, or machines.

Three Domains of Speech.

Among category systems, speech appears to be unique in several respects because of the mapping between production and perception. Speech exists simultaneously in three very different domains: the acoustic domain, the articulatory domain and the perceptual domain. While the relations among these three domains is complex, they are not arbitrary because the sound contrasts used in a language function within a common linguistic signaling system that is assumed to encompass both production and perception. Thus, the phonetic distinctions generated in speech production by the vocal tract are precisely those same acoustic differences that are important in perceptual analysis (Stevens, 1972). Any theoretical account of speech perception must also take into consideration aspects of speech production and acoustics. The perceptual spaces mapped out in speech production have to be very closely correlated with the same ones used in speech perception.

In learning the sound system of a language, the child must not only develop abilities to discriminate and identify sounds, but he/she must also be able to control the motor mechanisms used in articulation to generate precisely the same phonetic contrasts in speech production that he/she has become attuned to in perception. One reason that the developing perceptual system might preserve very fine phonetic details as well as characteristics of the talker's voice would be to allow a young child to accurately imitate and reproduce speech patterns heard in the surrounding language learning environment (Studdert-Kennedy, 1983). This skill would provide the child with an enormous benefit in acquiring the phonology of the local dialect from speakers he/she is exposed to early in life.

Discussion

It has become common over the last 25 years to argue that speech perception is a highly unique process that requires specialized neural processing mechanisms to carry out perceptual analysis (Liberman et al., 1967). These theoretical accounts of speech perception have typically emphasized the differences in perception between speech and other perceptual processes. Relatively few researchers working in the field of speech perception have tried to identify commonalities among other perceptual systems and draw parallels with speech. Our recent findings on the encoding of different sources of variability in speech and the role of long-term memory for specific instances are compatible with a rapidly growing body of research in cognitive psychology on implicit memory phenomena and non-analytic modes of processing (Schacter, 1992; Brooks, 1978).

Traditional memory research has been concerned with "explicit memory" in which the subject is required to consciously access and manipulate recently presented information from memory using "direct tests" such as recall or recognition. This line of memory research has a long history in experimental psychology and it is an area that most speech researchers are familiar with. In contrast, the recent literature on "implicit memory" phenomena has provided new evidence for unconscious aspects of perception, memory and cognition (Schacter, 1992; Roediger, 1990). Implicit memory refers to a form of memory that was acquired during a specific instance or episode and it is typically measured by "indirect tests" such as stem completion, cued recall, priming or changes in perceptual

identification performance. In these types of memory tests, subjects are not required to consciously recollect previously acquired information. In fact, in many cases, especially in processing spoken language, subjects may be unable to access the information deliberately or even bring it to consciousness (Studdert-Kennedy, 1974).

Studies of implicit memory have uncovered important new information about the effects of prior experience on perception and memory. In addition to traditional abstractionist modes of cognition which tend to emphasize symbolic coding of the stimulus input, numerous recent experiments have provided evidence for a parallel non-analytic memory system that preserves specific instances of stimulation as perceptual episodes or exemplars which are also stored in memory. These perceptual episodes have been shown to affect later processing activities. We believe that it is this implicit perceptual memory system that encodes the indexical information in speech about talker's gender, dialect and speaking rate. And, we believe that it is this memory system that encodes and preserves the perceptual operations or procedural knowledge that listeners acquire about specific voices that facilitates later recognition of novel words by familiar speakers.

Our findings demonstrating that spoken word recognition is talker-contingent and that familiar voices are encoded differently than novel voices, raises a new set of questions concerning the long-standing dissociation between the linguistic properties of speech-- the features, phonemes and words used to convey the linguistic message and the indexical properties of speech-- those personal or paralinguistic attributes of the speech signal which provide the listener with information about the form of the message-- the speaker's gender, dialect, social class, and emotional state among other things. In the past, these two sources of information were separated for purposes of linguistic analysis of the message. The present set of findings suggest this may have been an incorrect assumption for speech perception.

Relative to the research carried out on the linguistic properties of speech, which has a history dating back to the late 1940's, much less is known about perception of the acoustic correlates of the indexical or paralinguistic functions of speech (Ladefoged, 1975: Laver & Trudgill, 1979). While there have been a number of recent studies on explicit voice recognition and identification by human listeners (Papcun, Kreiman & Davis, 1989), very little research has been carried out on problems surrounding the "implicit" or "unconscious" encoding of attributes of voices and how this form of memory might affect the recognition process associated with the linguistic attributes of spoken words (Nygaard et al., Submitted). A question that naturally arises in this context is whether or not familiar voices are processed differently than unfamiliar or novel voices. Perhaps familiar voices are simply recognized more efficiently than novel voices and are perceived in fundamentally the same way by the same neural mechanisms as unfamiliar voices. The available evidence in the literature has shown, however, that familiar and unfamiliar voices are processed differentially by the two hemispheres of the brain and that selective impairment resulting from brain language can affect the perception of familiar and novel voices in very different ways (Kreiman & Van Lancker, 1988; Van Lancker, Cummings, Kreiman & Dobkin, 1988; Van Lancker, Kreiman & Cummings, 1989).

Most researchers working in speech perception adopted a common set of assumptions about the units of linguistic analysis and the goals of perceptual processing of speech signals. The primary objective was to extract the speaker's message from the acoustic waveform without regard to the source (Studdert-Kennedy, 1974). The present set of findings suggest that while the dissociation between indexical and linguistic properties of speech may have been a useful dichotomy for theoretical linguists who approach language as a highly abstract formalized symbolic system, the same set of

assumptions may no longer be useful for speech scientists who are interested in describing and modeling how the human nervous system encodes speech signals into representations in long-term memory.

Our recent findings on variability suggest that fine phonetic details about the form and structure of the signal are not lost as a consequence of perceptual analysis as widely assumed by researchers years ago. Attributes of the talker's voice are also not lost or normalized away, at least not immediately after perceptual analysis has been completed. In contrast to the theoretical views that were very popular a few years ago, the present findings have raised some new questions about how researchers have approached the problems of variability, invariance and perceptual normalization in the past. For example, there is now sufficient evidence from perceptual experimentation to suggest that the fundamental perceptual categories of speech-- phonemes and phoneme-like units, are probably not as rigidly fixed or well defined physically as theorists once believed. These perceptual categories appear to be highly variable and their physical attributes have been shown to be strongly affected by a wide variety of contextual factors (Klatt, 1979). It seems very unlikely after some 45 years of research on speech that very simple physical invariants for phonemes will be uncovered from analysis of the speech signal. If invariants are uncovered they will probably be very complex time-varying cues that are highly context-dependent.

Many of the theoretical views that speech researchers have held about language were motivated by linguistic considerations of speech as an idealized symbolic system essentially free from physical variability. Indeed, variability in speech was considered by many researchers to be a source of "noise" -- an undesirable set of perturbations on what was otherwise supposed to be an idealized sequence of abstract symbols arrayed linearly in time. Unfortunately, it has taken a long time for speech researchers to realize that variability is an inherent characteristic of all biological systems including speech. Rather than view variability as noise, some theorists have recognized that variability might actually be useful and informative to human listeners who are able to encode speech signals in variety of different ways depending upon the circumstances and demands of the listening task (Elman & McClelland, 1986). The recent proposals in the human memory literature for multiple memory systems suggest that the internal representation of speech is probably much more detailed and more elaborate than previously believed from simply an abstractionist linguistic point of view. The traditional views about features, phonemes and acoustic-phonetic invariance are no longer adequate to accommodate the new findings that have been uncovered concerning context effects and variability in speech perception and spoken word recognition. In the future, it may be very useful to explore the parallels between similar perceptual systems such as face recognition and voice recognition. There is, in fact, some reason to suspect that parallel neural mechanisms may be employed in each case despite the obvious differences in modalities.

Conclusions

The results summarized in this chapter on the role of variability in speech perception are compatible with non-analytic or instance-based views of cognition which emphasize the episodic encoding of specific details of the stimulus environment. Our studies on talker and rate variability and our new experiments on perceptual learning of novel phonetic contrasts and voices have provided important information about speech perception and spoken word recognition and have served to raise a set of new questions for future research. In this section, I simply list the major conclusions and hope these will encourage others to look at some of the long-standing problems in our field in a different way in the future.

First, our findings raise questions about previous views of the mental representation of speech. In particular, we have found that very detailed information about the source characteristics of a talker's voice are encoded into long-term memory. Whatever the internal representation of speech turns out to be, it is clear that it is not isomorphic with the linguist's description of speech as an abstract idealized sequence of segments. Mental representations of speech are much more detailed and more elaborate and they contain several sources of information about the talker's voice; perhaps these representations retain a perceptual record of the processing operations used to recognize the input patterns or maybe they reflect some other set of talker-specific attributes that permit a listener to explicitly recognize the voice of a familiar talker when asked to do so directly.

Second, our findings suggest a different approach to the problem of acoustic-phonetic variability in speech perception. Variability is not a source of noise; it is lawful and provides potentially useful information about characteristics of the talker's voice and speaking rate as well as the phonetic context. These sources of information may be accessed when a listener hears novel words or sentences produced by a familiar talker. Variability may provide important talker-specific information that affects encoding fluency and processing efficiency in a variety of tasks.

Third, our findings provide additional evidence that speech categories are highly sensitive to context and that some details of the input signal are not lost or filtered out as a consequence of perceptual analysis. These results are consistent with recent proposals for the existence of multiple memory systems and the role of perceptual representation systems (PRS) in memory and learning. The present findings also suggest a somewhat different view of the process of perceptual normalization which has generally focused on the processes of abstraction and stimulus reduction in categorization of speech sounds.

Finally, the results described here suggest several directions for new models of speech perception and spoken word recognition that are motivated by a different set of criteria than traditional abstractionist approaches to perception and memory. Exemplar-based or episodic models of categorization provide a viable theoretical alternative to the problems of invariance, variability and perceptual normalization that have been difficult to resolve with current models of speech perception that were inspired by formal linguistic analyses of language. We believe that many of the current theoretical problems in the field can be approached in quite different ways when viewed within the general framework of non-analytic or instance-based models of cognition which have alternative methods of dealing with variability, context effects and perceptual learning which have been the hallmarks of human speech perception.

References

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

Brooks, L. (1978), "Nonanalytic Concept Formation and Memory for Instances." In E. Rosch and B. Lloyd (Eds.), *Cognition and Categorization.* Hillsdale, NJ: Erlbaum.

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

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

Elman, J.L. & McClelland, J.L. (1986),. "Exploiting Lawful Variability in the Speech Wave." *Invariance and Variability in Speech Processes,* Hillsdale, NJ: Erlbaum, pp. 360-380.

Fowler, C.A. (In Press), "Listener-talker Attunements in Speech." In T. Tighe, B. Moore, and J. Santroch (Eds.), *Human Development and Communication Sciences.* Hillsdale, NJ: Erlbaum.

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

Goldinger, S.D. (1992), "Words and Voices: Implicit and Explicit Memory for Spoken Words." *Research on Speech Perception Technical Report No. 7,* Indiana University, Bloomington, IN.

Goldinger, S.D., Pisoni, D.B. & Logan, J.S. (1991), "On the Locus of Talker Variability Effects in Recall of Spoken Word Lists." *Journal of Experimental Psychology: Learning, Memory, and Cognition,* Vol. 17, 1, pp. 152-162.

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

Jacoby, L.L. & Brooks, L.R. (1984), "Nonanalytic Cognition: Memory, Perception, and Concept Learning." In G. Bower (Ed.), *The Psychology of Learning and Motivation, New York: Academic Press,* pp. 1-47.

Jusczyk, P. (1993), "Infant speech perception and the development of the mental lexicon." In H.C. Nusbaum and J.C. Goodman (Eds.), *The transition from speech sounds to spoken words: The development of speech perception.* Cambridge, MA: MIT Press.

Kakehi, K. (1992), "Adaptability to Differences Between Talkers in Japanese Monosyllabic Perception." In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure,* Tokyo, Japan: IOS Press, Inc.

Klatt, D.H. (1986), "The Problem of Variability in Speech Recognition and in Models of Speech Perception." In J.S. Perkell and D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum.

Klatt, D.H. (1979), "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access." *Journal of Phonetics*, Vol. 7, pp. 279-312.

Kolers, P.A. (1976b), "Pattern Analyzing Memory." *Science*, Vol. 191, pp. 1280-1281.

Kolers, P.A. (1973), "Remembering Operations." *Memory & Cognition*, Vol. 1, pp. 347-355.

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

Ladefoged, P. (1975), *A Course in Phonetics*. New York: Harcourt Brace Jovanovich, Inc.

Laver, J. and Trudgill, P. (1979), "Phonetic and linguistic markers in speech." In K.R. Scherer and H. Giles (Eds.) *Social Markers in Speech*. Cambridge: Cambridge University Press, pp. 1-31.

Liberman, A.M., Cooper F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967), "Perception of the Speech Code." *Psychological Review*, Vol. 74, pp. 431-461.

Lively, S.E., Pisoni, D.B. & Logan, J.S. (1992), "Some effects of training Japanese listeners to identify English /r/ and /l/." In Y. Tohkura (Ed.) *Speech Perception, Production and Linguistic Structure, Tokyo: Ohmsha Publishing Co. Ltd*, pp. 175-196.

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

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*, Vol. 89, 2, pp. 874-886.

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*, Vol. 15, pp. 676-684.

Mullennix, J.W. & Pisoni, D.B. (1990), "Stimulus Variability and Processing Dependencies in Speech Perception." *Perception & Psychophysics*, Vol. 47, 4, pp. 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*, vol. 85, pp. 365-378.

Neisser, U. (1976), *Cognitive Psychology*, New York: Appleton-Century-Crofts.

Nygaard, L.C., Sommers, M.S. & Pisoni, D.B. (1992), "Effects of speaking rate and talker variability on the representation of spoken words in memory." *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, 12-17 October 1992.

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

Nygaard, L.C., Sommers, M.S. & Pisoni, D.B. (1992), "Effects of Speaking Rate and Talker Variability on the Recall of Spoken Words." *Journal of the Acoustical Society of America*, Vol. 91, 4, pp. 2340.

Palmeri, T.J., Goldinger, S.D. & Pisoni, D.B. (1993), "Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words." *Journal of Experimental Psychology: Learning, Memory and Cognition,* Vol. 19 (2), pp. 1-20.

Papcun, G.,. Kreiman, J. and Davis, A. (1989). "Long-term memory for unfamiliar voices." *Journal of the Acoustical Society of America*, Vol. 85, pp. 913-925.

Peters, R.W. (1955), "The Relative Intelligibility of Single-voice and Multiple-voice Messages Under Various Conditions of Noise." *Joint Project Report No. 56, U.S. Naval School of Aviation Medicine*, pp. 1-9. Pensacola, FL.

Pisoni, D.B. (1992), "Some comments on invariance, variability and perceptual normalization in speech perception." *Proceedings 1992 International Conference on Spoken Language Processing, Banff, Canada, 12-17 October 1992.*

Pisoni, D.B. (1992), "Some Comments on Talker Normalization in Speech Perception." In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*, Tokyo, Japan: IOS Press, Inc.

Pisoni, D.B. (1990), "Effects of Talker Variability on Speech Perception: Implications for Current Research and Theory." *Proceedings of 1990 International Conference on Spoken Language Processing*, Kobe, Japan, pp. 1399-1407.

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

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

Pisoni, D.B. & Luce, P.A. (1986), "Speech Perception: Research, Theory, and the Principal Issues." In E.C. Schwab and H.C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines*, New York: Academic Press, pp. 1-50.

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

Pisoni D.B., Nusbaum, H.C., Luce, P.A. & Slowiaczek, L.M. (1985), "Speech perception, word recognition and the structure of the lexicon." *Speech Communication*, Vol. 4, pp. 75-95.

Posner, M.I. (1969), "Abstraction and the process of recognition." In J.T. Spence and G.H. Bower (Eds.), *The Psychology of Learning and Motivation: Advances in Learning and Motivation,* New York: Academic Press.

Posner, M. & Keele, S. (1986), "On the genesis of abstract ideas." *Journal of Experimental Psychology*, Vol. 77, pp. 353-363.

Roediger, H.L. (1990), "Implicit Memory: Retention Without Remembering." *American Psychologist*, Vol. 45, 9, pp. 1043-1056.

Schacter, D.L. (1992), "Understanding Implicit Memory: A Cognitive Neuroscience Approach." *American Psychologist*, Vol. 47, 4, pp. 559-569.

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

Sommers, M.S., Nygaard, L.C. & Pisoni, D.B. (1992), "The Effects of Speaking Rate and Amplitude Variability on Perceptual Identification." *Journal of the Acoustical Society of America*, Vol. 91, 4, pp. 2340.

Sommers, M.S., Nygaard, L.C. & Pisoni, D.B. (1992), "Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude." *Proceedings 1992 International Conference on Spoken Language Processing,* Banff, Canada, 12-17 October 1992.

Stevens, K.N. (1971), "Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds." *Proceedings of the Seventh International Congress of Phonetic Sciences*. The Hague: Mouton.

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

Strange, W. & Dittmann, S. (1984), "The effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, Vol. 36, 2, pp. 131-145.

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

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

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

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

Tulving, E. & Schacter, D.L. (1990), "Priming and Human Memory Systems." *Science*, Vol. 247, pp. 301-306.

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

Van Lancker, D.R., Kreiman, J. and Cummings, J. (1989), "Voice perception deficits: Neuroanatomical correlates of phonagnosia." *Journal of Clinical and Experimental Neuropsychology*, Vol. 11 (5), pp. 665-674.

Walley, A.C., Pisoni, D.B. & Aslin, R.N. (1981), "The role of early experience in the development of speech perception." In R.N. Aslin, J. Alberts and M.R. Petersen (Eds.), *The Development of Perception: Psychobiological Perspectives, New York: Academic Press*, pp. 2119-255.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Stimulus Variability and Spoken Word Recognition: Effects of Variability in Speaking Rate and Overall Amplitude[1]

Mitchell S. Sommers, Lynne C. Nygaard, and David B. Pisoni

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

[1]Portions of this research were supported by NIH Research Grant DC-00111-16 and NIDCD Training Grant DC-00012-13 to Indiana University in Bloomington, IN. A shorter version of this paper was presented at the 123rd meeting of the Acoustical Society of America, Salt Lake City, Utah (May 1992).

Abstract

These studies investigated the effects of several sources of naturally-occurring variability in speech, both in isolation and in combination, on the recognition of spoken words. Identification accuracy was poorer for word lists containing tokens produced by multiple talkers or at multiple speaking rates compared to the corresponding single-rate or single-talker conditions. Simultaneous variations in both speaking rate and source characteristics produced greater reductions in perceptual identification than either source alone. In contrast, variability due to overall amplitude did not significantly alter subjects' ability to correctly identify stimulus items. These findings suggest that the acoustic waveform is subjected to one or more transformations that act upon item-specific information in the signal prior to arriving at phonetic decisions. Implications of the results for models of speech perception are discussed.

Stimulus Variability and Spoken Word Recognition:Effects of Variability in Speaking Rate and Overall Amplitude

The spectral and temporal properties of speech signals that distinguish phonetic categories can be substantially altered by factors such as phonetic context (Liberman et al., 1967), stress level (Klatt, 1976), vocal-tract size and shape (Fant, 1973; Joos, 1948; Peterson and Barney, 1952), and speaking rate (Miller, 1981). Consequently, the perception of both vowels and consonants can be altered by changes in the speech waveform that arise from differences in one or more of the above factors (Johnson and Strange, 1982; Ladefoged and Broadbent, 1957; Peterson and Barney, 1952; Pickett and Decker, 1960; Port, 1976; Summerfield and Haggard, 1972; Verbrugge et al., 1976; Weenink, 1986). Ladefoged and Broadbent (1957), for example, reported that identification of a target vowel embedded in a carrier phrase could be altered by changing the perceived source characteristics of items immediately preceding it. Port (1976) provided evidence for a similar effect with speaking rate by demonstrating that distinctions between long and short vowels depended upon the prevailing articulation rate; listeners perceived long vowels at shorter overall durations for faster speaking rates. Summerfield (1974) extended these findings to consonantal distinctions by demonstrating that voiced-voiceless judgments for syllable-initial stops are also perceived in a rate-dependent manner; the same voice-onset-time (VOT) value can signal a voiceless, rather than a voiced stop at faster speaking rates.

Taken together, these studies illustrate one of the principle difficulties facing speech recognition systems: inherent variability in the acoustic realization of phonetic items, due to factors such as alterations in speaking rate and source characteristics, result in a one-to-many mapping between phonetic percepts and acoustic waveforms. The extent of the problem is readily apparent in the general inability, to date, of researchers to design completely speaker-independent speech recognition systems Gerstman, 1968). Specifically, one of the principle difficulties in implementing automatic speech recognition devices (ASRDs) is the lack of acoustic-phonetic invariance within the speech signal (Pisoni, 1985). However, in contrast to computer-based recognition devices, human listeners have little or no difficulty in understanding speech produced at different articulation rates or by a variety of talkers. Traditionally, theories of speech perception have accounted for this ability to maintain perceptual constancy, despite differences in the acoustic waveform, by positing a stage of perceptual normalization during which item-specific characteristics, such as rate and talker information, are adjusted or removed so that the resulting stimulus can be matched to canonical forms stored in long-term memory (Shankweiler et al., 1977).

The differential abilities of human and computer-based speech recognition systems to accommodate acoustic-phonetic variability is one factor that has recently led several investigators (Martin et al., 1989; Mullennix et al., 1989; Mullennix and Pisoni, 1990) to examine the effects of stimulus variability on phonetic processing. Automatic recognition systems that are capable of accurately encoding items produced at multiple speaking rates or by a number of different talkers do so at the cost of a significant increase in computational requirements (either increased processing time or faster and more efficient processing capabilities). Studies of stimulus variability and spoken word recognition with human listeners (Mullennix et al., 1989; Mullennix and Pisoni, 1990; Uchanski et al., 1992) have therefore focused on establishing whether normalizing for acoustic differences in speech signals produces a similar increase in processing requirements for humans. The increased demands might be reflected in poorer identification scores or other perceptual costs associated with increased stimulus variability. Such investigations of variability represent a significant departure from traditional studies of speech perception which have tended to consider variability as "noise" that is removed or

discarded prior to perceptual judgments (Shankweiler et al., 1974) rather than an aspect of the signal that can affect both the processing and representation of speech.

Several experiments examining vowel perception (Assmann et al., 1982; Ladefoged and Broadbent, 1957; Summerfield and Haggard, 1972) have demonstrated reduced identification scores and slower reaction times for stimuli produced by multiple, as opposed to single, talkers [however, see Verbrugge et al., (1976) for arguments against the effects of talker variability]. More recent experiments (Martin et al., 1989; Mullennix et al., 1989; Mullennix and Pisoni, 1990) have reported that trial-to-trial variations in source characteristics (monosyllabic words produced by different talkers) resulted in poorer word recognition scores and slower processing times relative to conditions in which the identical items were produced by single talkers. Mullennix et al. (1989) suggested that the introduction of talker variability increased the amount of processing required to identify the items by placing additional demands on the normalization system. Identification performance was reduced because the diversion of processing resources to mechanisms mediating normalization meant that fewer resources would be available for phonetic identification. Furthermore, given that changes in source characteristics impact principally on the spectral characteristics of speech waveforms, the reduced perceptual performance with mixed-talker lists likely reflects the operation of mechanisms that normalize or adjust for relative spectral differences in speech stimuli.

This suggestion raises the possibility that, in addition to adjusting for differences in spectral characteristics, listeners may also normalize for differences in temporal aspects of speech signals. As noted earlier, one temporal property that has been demonstrated to influence phoneme perception is speaking rate (Miller and Liberman, 1979; Port, 1976; Port and Dalby, 1982; Summerfield and Haggard, 1972). Miller and Liberman (1979), for example, synthesized a consonant-vowel (CV) continuum that ranged from /ba/ to /wa/ by systematically incrementing the duration of initial formant transitions; stimuli with shorter transition durations were perceived as /ba/ while those with longer transitions were perceived as /wa/. Miller and Liberman (1979) reported that the category boundary for the /ba/-/wa/ continuum could be altered by changing the duration of the steady-state portion of the syllable. For members of the continuum near the category boundary, therefore, a given transition duration could be perceived as either /ba/ or /wa/ depending on the duration (perceived rate) of the stimulus. More recently, Miller and Voalitis (1992) have shown that the internal category structure of syllables can vary as a function of speaking rate. Taken together, these findings suggest that speech signals are perceived in a rate-dependent manner and that listeners must adjust or normalize for differences in speaking rate to maintain perceptual constancy.

Only a few studies, however, (Johnson and Strange, 1982; Verbrugge and Shankweiler, 1977) have examined whether rate normalization affects phonetic identification. Johnson and Strange (1982) found improved identification of rapidly articulated long vowels when the stimuli were presented in sentences produced at the same (rapid) speaking rate as the vowel than when the vowel and sentence were spoken at different rates (rapid rate for vowel and normal rate for sentence). Thus, vowel recognition was superior when listeners were not required to devote additional processing resource to normalize for differences in speaking rate between precursor sentence and the vowel target.

The goal of the present set of experiments was to address a number of questions regarding the effects of speaking rate on spoken word recognition. First, the studies were designed to determine whether variations in speaking rate produce reductions in word identification scores similar to those found for talker variability. Secondly, the experiments evaluated whether combining different sources of variability would reduce identification performance to a greater extent than when stimuli varied

along only a single dimension. If variability in speech signals increases the processing demands required for normalization, then combining two different sources of variability might have a greater affect on word recognition than either source alone. Lastly, we wanted to establish whether variability along any stimulus dimension reduced identification performance or whether there were some sources of variability that would not affect identification scores.

EXPERIMENT 1

The purpose of experiment 1 was threefold. First, we wanted to replicate the findings of Mullennix et al. (1989) who demonstrated decreased word recognition scores for mixed-talker, compared to single-talker, word lists, using a different corpus of stimuli. Mullennix et al. (1989) used a restricted set of stimulus items taken from the Modified Rhyme Test (MRT) and the first experiment was designed to determine whether their findings would generalize to other sets of speech tokens. Secondly, we investigated whether variability along a different stimulus dimension, speaking rate, would produce decrements in identification performance similar to those observed for talker variability. Finally, the study was designed to examine whether the combined effects of rate and talker variability produced greater decrements in word recognition scores than variations in either source alone.

METHOD

Subjects

Three-hundred and ninety graduate and undergraduate students from the Psychology Department at Indiana University served as subjects. Ten of the students recorded the stimulus materials used throughout the investigations and the remainder served as listeners in perceptual experiments. All subjects were native speakers of English and reported no history of speech or hearing disorders at the time of testing.

Stimulus Materials

The stimuli used in Experiment 1 consisted of 100 different words taken from two fifty-item phonetically-balanced (PB) word lists (ANSI, 1971). All 100 words were embedded in the carrier phrase "Please say the word _____". The sentences were presented to subjects on a CRT screen located in a sound-attenuated booth (IAC 401A).

Ten different talkers (6 males and 4 females) produced each of the 100 sentences at three different speaking rates. At the beginning of each session, subjects were told that they would have to produce sentences at three different speaking rates (fast, medium and slow). Other than indicating that the different rates should be distinct, no additional instructions regarding speaking rate were given. Sentences were recorded in blocks of 100 with speaking rate remaining constant within a block. The order in which subjects recorded the different rates was determined randomly.

The utterances were transduced with a Shure (SM98) audio microphone, digitized on-line (12-bit analog-to-digital converter (DT2801) at a sampling rate of 10 kHz), low-pass filtered at 4.8 kHz , and stored on disk. Target words were edited from carrier phrases using a digital waveform editor. Average durations for the isolated words were 905, 533, and 375 ms for the slow, medium and fast items, respectively. The RMS amplitude levels of all stimuli were equated using a software package designed to modify speech waveforms.

Procedure

General Identification Paradigm

On each trial, subjects sat facing a CRT terminal and a warning stimulus (a string of letters saying "GET READY FOR NEXT TRIAL") was presented in the center of the screen. The warning stimulus remained on for 500 ms. A 500-ms silent interval separated the warning signal and stimulus presentation. Stimuli were presented binaurally over matched and calibrated TDH-39 headphones at approximately 80 dB SPL. Subjects were instructed to type the word they thought they heard onto the CRT screen. No feedback was given to subjects following a response. A 2-s ISI began after the last person had responded. Only exact phonetic matches to the item presented were counted as correct responses. Stimulus output and data collection were controlled on-line by a PDP-11/34a computer. Stimuli were output via a 12-bit D/A converter at a 10-kHz sampling rate and were low-pass filtered at 4.8 kHz.

Intelligibility Measures

To ensure comparable intelligibility of items across both talkers and speaking rates, all 100 words produced by a given talker at a single speaking rate were presented to listeners in the quiet for identification. The average intelligibility of all items produced by individual talkers ranged between 85 and 96 percent. Excluding data from one female talker whose fast items were identified at 78 percent[2] increased average intelligibility to between 91 and 96 percent. Identification scores did not differ significantly as function of either speaking rate or individual talker.

Saliency of Speaking Rate Differences

To determine whether listeners perceived three distinct speaking rates for each talker, a separate group of 50 subjects was asked to judge the speaking rate of all 300 words (100 items x 3 speaking rates) produced by each of the ten talkers. Items were presented in the quiet and subjects were instructed to press one of three buttons indicating whether they thought the word was produced at a fast, medium or slow speaking rate. Correct rate judgments (i.e. judgments matching the speaking rate designated during stimulus recording) averaged 82, 81, and 76 percent for the slow, medium and fast items respectively. Correct rate judgments did not differ significantly for the three speaking rates.

Effects of Variability Along a Single Dimension

Given the high average intelligibility ratings obtained for test items in quiet, all perceptual tests measuring the effects of stimulus variability on word recognition were conducted in a background of noise. Four signal-to-noise (S/N) ratios (+5, 0, -5, and -10) were tested in each condition.

To investigate the effects of variability in talker characteristics and speaking rate on perceptual identification, subjects in the single-dimension conditions (single-talker and single-rate) were presented 100 words produced by a single talker at one speaking rate. In the single-talker condition, a total of nineteen subjects and four different voices (2 male and 2 female) were used. Ten of the subjects heard words produced by one of the two male voices and nine listeners heard items produced by one of the

[2]Only slow and medium rate items from this talker were used in the experiments.

two female talkers. Nineteen subjects were also tested in the mixed-talker condition. These listeners heard items identical to those presented to subjects in the single-talker condition but the voice of the talker on any given trial was selected randomly from among 10 different talkers (6 male and 4 female). Four of the voices used in the mixed-talker condition were those presented to subjects in the single-talker condition.

In the single-rate condition, three groups of fifteen subjects were tested. Seven subjects in each group heard the 100 items produced at a single rate (fast, medium, or slow) by a male talker and 8 subjects heard the same items produced by a female talker. For the mixed-rate condition, there were a total of 135 subjects[3] and speaking rate was varied randomly from trial to trial. Stimulus sets were constructed such that each item was presented at a given speaking rate to one-third of the subjects. In addition, half the listeners heard words produced at a given rate by a male talker and half heard the item produced by a female talker.

Combined Effects of Rate and Talker Variability

In the final part of experiment 1, an additional 60 subjects were recruited to compare identification performance when word lists varied along a single dimension (either speaking rate or talker characteristics) versus conditions in which items varied along both dimensions simultaneously. Twenty subjects were tested in each of the three conditions (mixed-talker, mixed-rate, or mixed-talker and mixed-rate). The mixed-talker and mixed-rate conditions were identical to those described previously. In the last condition (combined variability condition) both speaking rate and talker varied randomly from trial-to-trial. Thus, words could be presented at any of the three speaking rates and by any one of the ten talkers[4].

Results and Discussion

Identification scores for all conditions tested in Experiment 1 are shown in Figure 1. Each panel displays percent correct identification as a function of signal-to-noise ratio. The top panel shows data for the single- vs. mixed-rate condition. The middle panel compares scores for single- vs. mixed-talker lists and the bottom panel shows the combined effects of talker and rate variability vs. the effects of each source individually.

Insert Figure 1 about here

A four-way ANOVA conducted on the data shown in the top panel of Fig. 1, with the factors of variability (single- vs. mixed-speaking rates), speaking rate (slow, medium, or fast), S/N ratio (+5, 0, -5, or -10), and voice (male vs. female) indicated no main effect or interaction of speaking rate with

---

[3]Since a given subject in the mixed-rate condition heard only one-third of the words produced at given rate we needed to have 3 times as many subjects to make comparisons with the single-rate condition in which a total of 45 subjects heard the words at single speaking rates.

[4]With the exception of the fast items from the female speaker, mentioned above, whose fast-rate productions were identified at 78 percent in quiet.

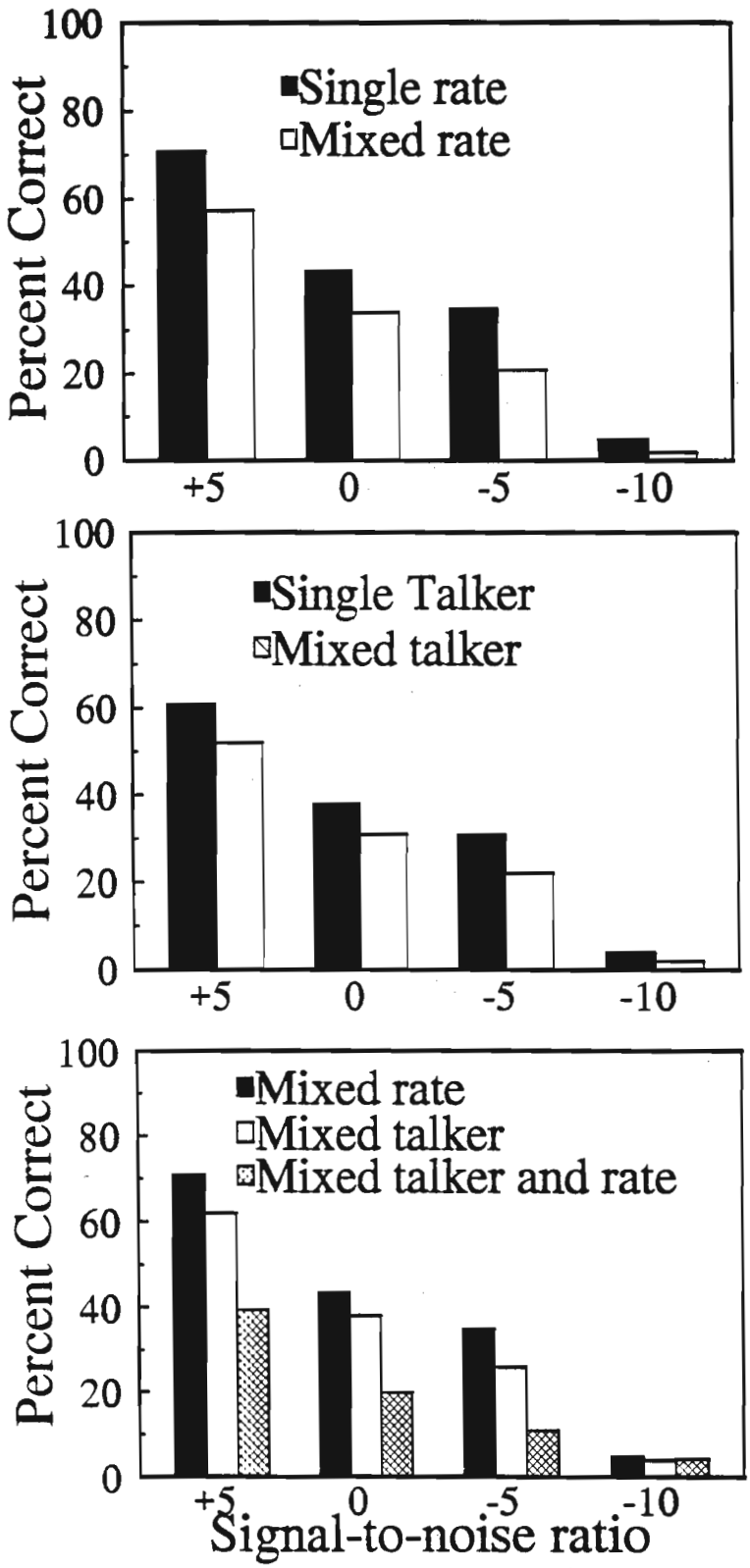


Figure 1: (**top**)-Percent correct identification as a function of signal-to-noise ratio for single-rate (filled bars) and mixed-rate (open bars) word lists. (**middle**)-same as above except bars are for single- and mixed-talker word lists. (**bottom**)-Filled and open bars are for mixed-rate and mixed talker lists, respectively. Hatched bars are for conditions in which rate and talker varied simultaneously.

the other factors. Therefore, the data at each S/N ratio have been collapsed across speaking rates. Similarly, there was no main effect or interaction of voice with the remaining variables and the data for male and female voices have therefore been combined. Significant main effects were obtained for the factors of variability [$F(1,402) = 81.95$, $p < .001$] and S/N ratio [$F(3,402) = 58.15$, $p < .001$]. Thus, items presented in mixed-rate lists were recognized less accurately than the identical items presented in single-rate lists and words presented at lower signal-to-noise ratios were identified less accurately than those presented at higher S/N ratios. None of the remaining two- or three-way interactions reached significance.

A separate three-way ANOVA for the single- versus mixed-talker contexts with the factors of variability (single- vs. mixed-talker), talker (male vs. female), and S/N ratio (+5, 0, -5, -10) was conducted on the data shown in the middle panel of Fig 1. Within each S/N ratio, the data have been collapsed across male and female talkers since there was no main effect or significant interactions of talker with the other two factors. A significant main effect of variability [$F(1,345)=1958.22$, $p < .001$)] was observed; identification was poorer for items presented in mixed-talker lists than for words presented in single-talker contexts. There was also a significant main effect of S/N ratio [$F(3,345) = 180.06$, $p<.001$]; words presented at higher S/N ratios were recognized more accurately than words presented at lower S/N ratios. The three-way interaction of variability x talker x S/N ratio was not significant.

To determine whether combining rate and talker variability produced greater decrements in identification performance than either source alone, a three-way ANOVA with the factors of variability (mixed-talker, mixed-rate, mixed-talker and rate), rate (slow, medium, or fast) and S/N ratio (+5, 0, -5, -10) was conducted on the data shown in the bottom panel of Fig. 1. Again, no main effect or interactions involving speaking rate reached statistical significance and the data have therefore been collapsed across the three rates. A significant main effect of variability was observed [$F(2,622) = 45.21$, $p<.001$]. Newman-Keuls *post hoc* analyses showed that words presented in the combined rate and talker variability condition were recognized less accurately than items presented in either the mixed-rate or mixed-talker lists. No significant difference between the mixed-rate and mixed-talker conditions was observed. A significant main effect of S/N was also found [$F(3,622)= 10.98$; $p < .001$). No other significant main effects or interactions were obtained.

The findings from experiment 1 replicate and extend earlier work (Creelman, 1957; Mullennix et al. 1989) on the effects of talker variability on spoken word recognition. As noted above, Mullennix et al. (1989) found poorer identification performance in noise for a subset of words taken from the Modified Rhyme Test (MRT). The present study demonstrated comparable effects of talker variability for words taken from phonetically-balanced word lists. More importantly, the current investigation found that variability in speaking rate has similar detrimental effects on identification of isolated words presented in noise. Trial-to-trial variations in speaking rate reduced identification accuracy relative to conditions in which the *identical* waveforms occurred in single-rate contexts. Experiment 1 also demonstrated that spoken word recognition was significantly worse when items varied simultaneously along two dimensions, speaking rate and talker characteristics, compared to conditions in which either source varied independently.

Taken together, the findings of Experiment 1 are consistent with the suggestion (Mullennix et al., 1989) that, to maintain perceptual constancy, listeners use a set of resource-demanding processes to normalize or adjust for differences in the acoustic realization of speech signals that can have significant phonetic consequences. As additional demands are placed on the normalization system, (e.g. varying

both rate and talker simultaneously), more of a limited-pool of resources must be devoted to compensating for the increased acoustic differences among stimulus items. Consequently, fewer processing resources are available for identifying the items and spoken word recognition performance is reduced.

An alternative explanation for the detrimental effects of stimulus variability on identification performance, however, is that introducing any variations in the speech signal simply diverts attention away from the phonetic content of the stimuli. That is, trial-to-trial changes in speaking rate or talker characteristics causes listeners to attend more to the varying dimension and less to the actual stimulus item. According to this hypothesis, any source of variability in the speech signal should reduce identification scores if it is perceptually salient. In contrast, explanations for decreased word recognition performance based on greater normalization demands in mixed-rate or mixed-talker conditions suggest that sources of variability which do not directly affect phonetic identification should require little or no normalization and therefore should have relatively minor effects on spoken word recognition. Experiment 2 was designed to distinguish these two explanations.

EXPERIMENT 2

Experiment 2 compared the effects of two sources of variability, speaking rate and overall amplitude, on spoken word recognition. As noted above, Miller and Liberman (1979) demonstrated that listeners normalize or adjust for changes in speaking rate. Therefore, both the normalization and attentional explanations of reduced identification performance as a function of increased stimulus variability would predict that trial-to-trial variations in speaking rate should reduce word recognition scores. Overall amplitude, in contrast, has not been shown to have direct effects on phonetic identification. According to the normalization hypothesis, therefore, variations in overall amplitude should have less of an affect on identification performance than trial-to-trial changes in speaking rate because listeners do not normalize (or at least normalize to a lesser extent) for differences in overall level. Attentional accounts of reduced word recognition scores due to stimulus variability, however, would predict similar effects of amplitude and speaking-rate variability since both are salient to listeners and both sources of variability can therefore divert attention from the phonetic content of items. Experiment 2 tested both of these predictions.

Method

Subjects

220 undergraduate students in Introductory Psychology courses at Indiana University served as subjects in perceptual experiments. All were native speakers of English and had no history of speech or hearing disorders. Each subject participated in a 1-h session and received partial course credit.

Stimulus Materials

The stimuli used to investigate the effects of variability in speaking rate on perceptual identification were identical to those of the first experiment (100 words from PB word lists produced at 3 different speaking rates). The effects of variability in overall amplitude were evaluated using the 100 medium-rate items from Experiment 1 with each word taking on three different overall amplitudes. Stimulus levels were controlled by a software package that set the maximum level in a waveform to a specified value and scaled the remaining amplitude values in the digital file relative to this maximum.

Three overall maximum levels were used in this experiment: 35, 50 and 65 dB. Thus, overall amplitude varied over a 30-dB range.

In addition to comparing the effects of variations in speaking rate and overall amplitude on spoken word recognition, Experiment 2 was to designed to determine whether the specific means of stimulus degradation used in Experiment 1, the introduction of background noise, was, at least in part, responsible for the detrimental effects of variability. Although neither of the theoretical positions (attention- or normalization-based explanations) predicted that the effects of variability would be specific to items degraded by noise, similar results with an alternative means of reducing overall performance levels would increase the generality of the findings. Therefore, stimuli in Experiment 2 were degraded using a computer program that switched the amplitude values (from positive to negative or vice-versa) in the digital waveform at randomly determined points over a specified portion of the waveform. A stimulus with a 10% degradation level, for example, consisted of the original waveform with a randomly-chosen 10% of the amplitude values opposite of those found in the original digital file. Perceptually, listeners described the stimuli as either "noisy" or "distorted". Four levels of stimulus degradation, 10, 15, 20, and 30 percent, were used for all perceptual tests in experiment 2.

Procedure

The design and procedures for Experiment 2 were similar to those of the first study. Isolated monosyllabic words were presented over headphones and subjects were required to type the word they heard onto a CRT terminal. Stimulus items were presented in five different conditions: (1) single rate--i.e. all tokens produced by a single talker at one rate; (2) mixed rate--one-third of the items randomly presented at each of the three speaking rates; (3) single amplitude--all 100 words presented at a single amplitude; (4) mixed amplitude--one-third of the stimuli randomly presented at each of the three overall amplitudes and; (5) mixed amplitude and rate--stimuli randomly presented at all three amplitudes and speaking rates). Half of the subjects in each condition heard items produced by a male talker and half listened to the identical items produced by a female talker.

In the single- and mixed-rate contexts, stimuli were presented at approximately 80 dB SPL. For the single- and mixed-amplitude conditions, presentation levels were set such that words digitized with a 50-dB peak amplitude (medium-level condition) were presented at an overall level of 80-dB SPL. Stimuli digitized at the low (35) and high (65) dB levels were therefore presented at approximately 65 and 95 dB SPL, respectively. In all conditions, subjects received equal numbers of stimuli at each of the four degradation levels.

Results and Discussion

Figure 2 displays results from the five conditions tested in Experiment 2. In each panel, percent correct identification is plotted as a function of percent stimulus degradation. The top panel compares identification performance in the single- and mixed-rate conditions using the amplitude bit-switching program as a means of degrading the stimuli. Overall performance levels were comparable to those observed for the four S/N ratios tested in Experiment 1. A four-way ANOVA was conducted on the factors of variability (single vs. mixed rate), degradation percentage (10, 15, 20, and 30), rate (fast, medium and slow), and talker (male vs. female). No significant main effects or interactions were found for the factors of talker and rate Therefore, data for the male and female talker and for all three speaking rates have been averaged. Significant main effects were observed for the factors of variability [$F(1,369) = 1683.01$, $p < .001$] and degradation level [$F(3,369) = 68.13$, $p < .001$]; identification accuracy in the single-rate condition was significantly better than in the mixed-rate context and word

recognition scores decreased with increasing amounts of stimulus degradation. None of the remaining main effects or interactions reached statistical significance. These findings suggest that the reduction in identification performance for items presented in mixed-, as opposed to single- rate contexts that was observed in both Experiments 1 and 2 is not the result of specific stimulus degradation procedures (either noise or amplitude bit-flipping) but reflects some aspect of speech processing.

Insert Figure 2 about here

The comparison between single- and mixed-amplitude conditions is shown in the middle panel of Fig. 2. The data have been averaged across talker and overall amplitude because a four-way ANOVA with the factors of variability (single-amplitude vs. mixed-amplitude), overall amplitude (35, 50 or 65), degradation percent (5, 10, 20, and 30), and talker (male vs. female) showed no significant main effects or interactions involving these two factors. Only the main effect of degradation [$F(3, 369) = 15.95$, $p < .001$] was significant; identification scores decreased with greater percentages of degradation. None of the remaining two-, three- or four-way interactions was statistically reliable. Thus, in contrast to the effects of speaking-rate and talker variability that were observed in this and the previous experiment, trial-to-trial variations in overall amplitude over a 30-dB range did not produce significant decrements in identification performance [$F(1,369) = 1.06$, $p > .3$]. The absence of a significant main effect of overall amplitude [$F(2,369)=1.33$, $p>.2$] indicated that items produced at low, medium, and high overall presentation levels did not differ in intelligibility.

The bottom panel of Fig. 2 contrasts identification scores obtained when speaking rate and overall amplitude vary independently with the effects of simultaneous variations along both dimensions. A three-way ANOVA with the factors of variability (mixed-rate, mixed-amplitude, and mixed-amplitude and rate), overall amplitude (35, 50, 65) and percent signal degradation (10, 15, 20, 30) revealed a significant main effect of variability [$F(2, 227) = 8.57$; $p < .001$]. Newman-Keuls *post hoc* analyses indicated that identification scores in the mixed-amplitude condition were significantly better than those in the mixed-rate and combined rate and amplitude conditions. Differences between the mixed-rate and combined mixed-rate and mixed-amplitude contexts did not reach statistical significance. The only other statistically reliable effect found was for degradation level [$F(3,227) = 83.25$; $p < .001$]. These findings indicate that word recognition scores are not affected by variability in overall amplitude but are significantly reduced when items vary in speaking rate. Furthermore, unlike the first experiment, where the combined effects of rate and talker variability were greater than the effects of either source alone, simultaneous variations in overall amplitude and speaking rate do not combine to produce poorer identification performance than contexts in which only speaking rate is varied.

Taken together, the results of Experiment 2 are consistent with the hypothesis that normalizing for dimensions of the speech signal that directly affect phonetic identification, such as speaking rate or source characteristics, is a resource-demanding process that can significantly reduce word recognition scores. In contrast, the normalization demands for sources of variability that do not have direct effects on phonetic identification, such as overall amplitude, are attenuated substantially. Consequently, variability along such phonetically-irrelevant dimensions do not have significant effects on word recognition performance. These results argue against any generalized attention-based explanation for the effects of stimulus variability on spoken word recognition. Specifically, it is not the case that any

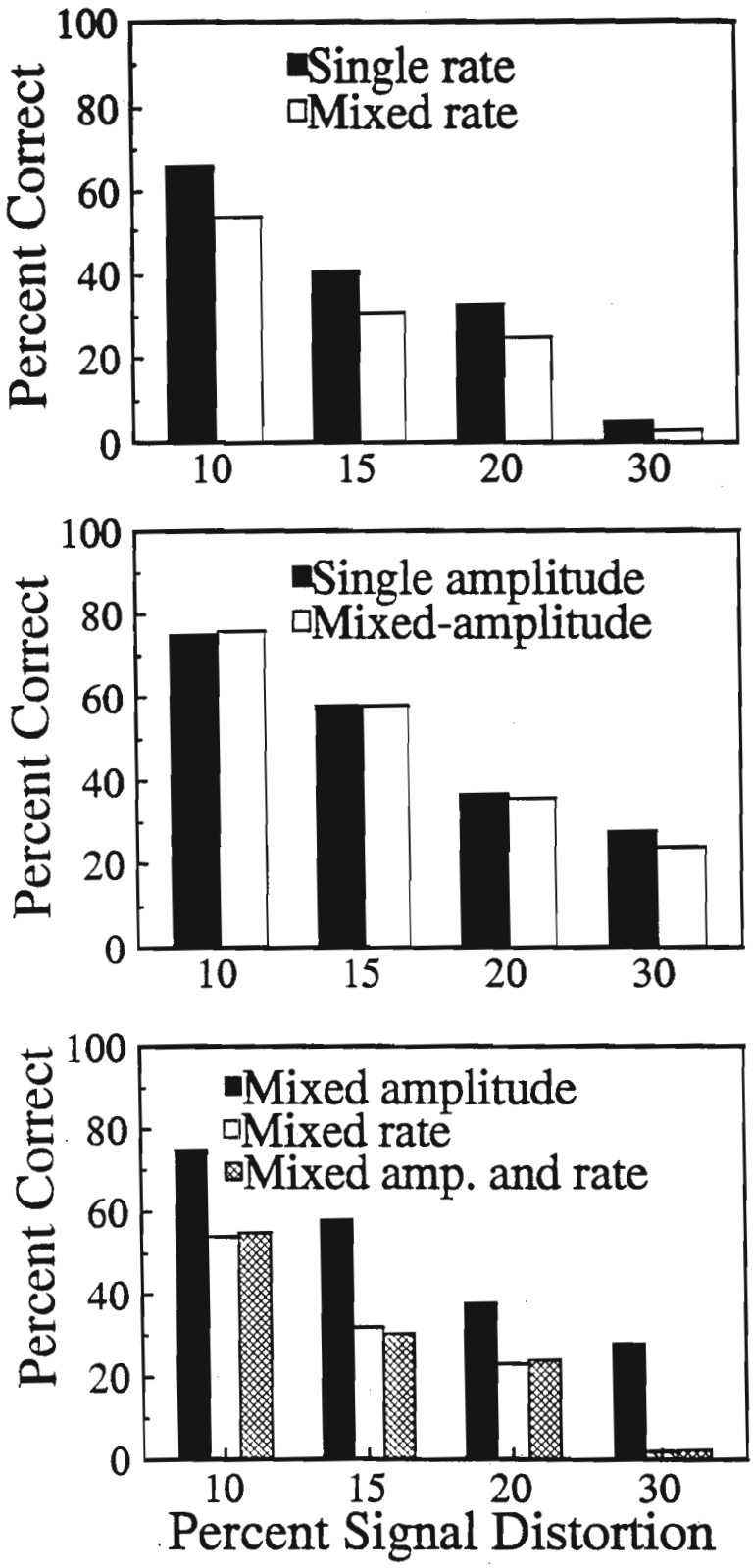


Figure 2: (**top**)-Percent correct identification as a function of degradation percent for single-rate (filled bars) and mixed-rate (open bars) word lists. (**middle**)-same as above except bars are for single- and mixed-amplitude word lists. (**bottom**)-Filled and open bars are for mixed-rate and mixed amplitude lists, respectively. Hatched bars are for conditions in which rate and amplitude varied simultaneously.

source of variability within the speech signal will reduce identification performance. Rather, certain types of variability, such as overall amplitude, are either ignored or are more easily accommodated by the speech perception system.

GENERAL DISCUSSION

The results of the present investigations extend recent research (Mullennix et al., 1989; Mullennix and Pisoni, 1990) on stimulus variability and spoken word recognition in several ways. First, the findings demonstrate that trial-to-trial changes in speaking rate can reduce identification scores for speech stimuli degraded either by external noise or by digital distortion. Secondly, the data indicate that simultaneous variations along two dimensions, speaking rate and talker characteristics, can produce greater decrements in word recognition performance than either source alone. Thirdly, the similarity of identification scores in the single- and mixed-amplitude conditions provides evidence that some sources of variability within speech signals do not impair spoken word recognition.

These findings are consistent with proposals (Martin et al., 1989; Mullennix et al. 1989; Mullennix and Pisoni, 1990), that the process of maintaining perceptual constancy in speech perception is a resource-demanding capacity that can impact significantly on spoken word recognition. According to these arguments, the 10-15% poorer identification performance for mixed-rate and mixed-talker lists (compared to the corresponding single-dimension contexts) is the result of additional processing requirements that are necessary to compensate for increased amounts of phonetically-relevant acoustic variability. Several other investigators (Johnson and Strange, 1982; Nygaard et al., 1992) have also suggested that greater normalization requirements can account for reduced phonetic processing abilities. For example, as noted previously, Johnson and Strange (1982) found that identification of rapidly articulated long vowels was better when they were presented in precursor phrases which were also produced at fast speaking rates than when they occurred in sentences produced at normal speaking rates. The improvement, according to these investigators, was due to reduced processing requirements when targets and precursors were spoken at the same rate. When the vowel and sentence were both produced at fast articulation rates, fewer resources had to be allocated to compensate for rate-dependent changes in the acoustic properties of the waveform and, therefore, more could be devoted to determining vowel quality.

A recent investigation by Nygaard et al. (1992) provides further support for the hypothesis that listeners employ a resource-demanding perceptual capacity to adjust or normalize[5] for differences in speaking rate. Nygaard et al. compared serial recall for single- vs. mixed-rate lists with the same monosyllabic stimuli used in the present experiment. We found poorer serial recall in the primacy portion of 10-word lists when the items were produced at different speaking rates than when the identical stimuli were produced at the same speaking rate. The explanation proposed for these findings was that, in the mixed-rate condition, listeners had to devote additional processing resources to adjust for changes in speaking rate. With fewer resources available for rehearsal, recall in the primacy portion of the lists was therefore impaired relative to single-rate contexts.

---

[5]Although the term normalization has traditionally been used to imply a loss of information, in the present context it refers to all perceptual mechanisms used to obtain invariant phonetic percepts from acoustically variable signals. It may be the case that talker and rate variability decrease spoken word recognition because listeners are required to encode some or all of the additional information in the signal. Thus, rather than a loss of information, increased normalization demands may result from listeners extracting additional information from the waveform.

Another recent experiment (Uchanski et al., 1992), investigating the effects of token variability on vowel identification, also suggests that additional processing resources needed to compensate for stimulus variability can produce decrements in vowel identification even when the items are produced by the same talker at the same speaking rate. Uchanski et al. had talkers produce sixteen different utterances of ten vowels and presented them to listeners for identification in either a single-token (all instances of a given vowel taken from the same utterance) or mixed-token (vowels taken from all 16 utterances) context. Identification accuracy was significantly greater for the single-token than for the mixed-token condition. In accord with previous accounts of reduced identification performance as a function of increased variability, Uchanski et al. (1992) proposed that intra-subject differences in the acoustic realization of vowels was sufficient to require additional processing resources to maintain perceptual constancy.

When considered together, the results of experiments investigating the effects of talker (Mullennix et al., 1989), token (Uchanski et al., 1992) and rate (present experiment) variability suggest that listeners employ a resource-demanding capacity to compensate for acoustic differences in speech signals. Furthermore, the alternative explanation for these findings, that increased stimulus variability diverts attention from the phonetic content of items, was not supported by the results of Experiment 2. This last finding is particularly important given that a number of previous studies have demonstrated reduced identification and discrimination performance for both speech (Miller et al., 1951) and nonspeech (Watson and Kelly, 1981) as a function of stimulus uncertainty. The comparable identification performance of listeners in the mixed-rate and mixed-amplitude conditions of Experiment 2 provides evidence that reductions in spoken word recognition as a function of increased speaking-rate variability cannot be attributed exclusively to greater amounts of stimulus uncertainly. It should be noted in this regard, however, that studies comparing the perceptual consequences of variations along two or more stimulus dimensions must be interpreted with caution. Differences in the saliency of each dimension as well as the range over which variations are introduced can have a significant effect on the outcome of studies contrasting the effects of different sources of variability. In the present investigation, although the differences in speaking rate and overall amplitude were clearly perceptible to listeners, it remains uncertain whether the range over which the two dimensions varied were perceptually equivalent. Future studies, therefore, should include methodologies, such as multi-dimensional scaling, which allow for more direct comparisons of perceived similarity among different stimulus dimensions.

The differential effects of speaking rate and overall amplitude variability on spoken word recognition raises the possibility that listeners are required to compensate or normalize only for those dimensions of the speech signal that have direct phonetic consequences. One reason that trial-to-trial variations in speaking rate may produce poorer identification performance while overall amplitude variability does not is that listeners are required to normalize for differences in speaking rate, which can directly alter phoneme identification, but changes in overall amplitude, which do not have similar phonetic consequences, are not processed in an obligatory manner.

Evidence from a number of studies, using different experimental paradigms, suggests that listeners are required to process rate information. For instance, Miller (1981), in an extension of her earlier work on rate-dependent processing of speech (Miller and Liberman, 1979), asked subjects to label stimuli varying along a /ba/-/wa/ continuum as quickly as possible. As in their earlier study, the duration of the steady-state vowel portion of the stimuli was systematically varied as a means of altering perceived speaking rate. Miller (1981) argued that, under conditions emphasizing the speed of

responses, subjects could minimize response latencies by ignoring the vowel portion of the stimuli. If listeners adopted this strategy, rate-dependent changes in category boundaries should not be observed since the duration of the initial phoneme was the same for /ba/ and /wa/. However, if listeners were required to listen to the vowel portion of the stimulus to obtain information about speaking rate before making phonetic decisions, then the typical rate-dependent shifts in category boundaries should be observed. Miller (1981) reported that, despite explicit instructions to respond as quickly as possible, subjects still demonstrated rate-dependent category boundary shifts. This suggests that, for phonetic decisions that can be affected by speaking rate, listeners were required to process the entire syllable rather than just the initial phoneme since syllable duration provided an indication of articulation rate. Tomiak et al. (1991) also found evidence to support mandatory processing of rate information using a speeded classification task (Garner, 1974). They reported that listeners could not ignore irrelevant variations in speaking-rate when making phonetic judgments. These findings are consistent with the suggestion that the processes involved in speaking rate normalization are invoked automatically and in an obligatory fashion.

Results from recent studies of talker variability (Mullennix and Pisoni, 1990) indicate that talker information is also processed in a mandatory fashion. Mullennix and Pisoni (1990) employed the same speeded classification procedure used by Tomiak et al. (1992) and found that talker and phoneme information were not processed independently. Rather, listeners were unable to ignore variations in the gender of the talker when making phonetic decisions. Thus, both rate and talker information are processed in an obligatory and automatic fashion and variability along both dimensions can reduce identification accuracy.

Unfortunately, comparable evidence regarding the mandatory processing of overall amplitude is not available. Therefore, suggestions that the differential effects of rate, talker and amplitude variability on phonetic identification are the result of mandatory processing of the first two dimensions but nonobligatory processing of overall level must remain speculative. The proposal, however, is consistent with the literature concerning contextual effects on phonetic identification. Information about speaking rate and source characteristics has been demonstrated to have direct effects on phonetic identification (Fant, 1973; Miller and Liberman, 1979; Peterson and Barney, 1952) and therefore listeners should be required to normalize for differences along these two dimensions to maintain perceptual constancy. In contrast, overall amplitude does not have direct effects on phoneme identification and, consequently, absolute level changes may not be processed in an obligatory fashion. Additional studies regarding the mandatory processing of properties of the speech signal, such as overall level and absolute pitch, which do not have direct effects on phonetic identification, will be necessary before more definite conclusions can be drawn about the relationship between mandatory processing of a given dimension and its potential for affecting spoken word recognition. Similarly, investigations of obligatory perceptual compensation for stimulus characteristics, such as token variability, which have been shown to affect spoken word recognition should provide additional data relevant to this proposal.

In view of the evidence that reductions in identification scores for word lists produced at multiple speaking rates are due to a resource-demanding normalization process, it is important to determine at what level the mechanisms mediating this normalization might operate. Miller (1987) has suggested that rate normalization occurs during initial acoustic-phonetic encoding. The advantage of adjusting for rate variations during early, low-level, stages of speech processing, according to Miller, is that the necessary compensations would occur automatically and would therefore not require conscious attention. According to this proposal, the detrimental effects of speaking-rate variability on

word recognition can be accounted for by increased acoustic-phonetic encoding difficulties in mixed- compared to single- rate contexts. Consistent with this explanation are recent findings by Sommers and Humes (1993) on the effects of speaking rate and speaking rate variability on elderly listeners. Older subjects, with little or no hearing impairment in the frequency regions most important for speech perception[6], demonstrated a significant reduction in word recognition in quiet for items produced at fast speaking rates compared to similar words spoken at either slow or medium rates. Furthermore, when variations in speaking-rate were introduced, recognition scores for words produced at fast rates declined to levels below those obtained in the corresponding single-rate context. For slow- and medium-rate words, however, no reduction in performance was observed as a function of increased rate variability. Thus, under conditions in which encoding was relatively easy (the medium and slow articulation rates) identification was unaffected by stimulus variability. However, under more demanding encoding conditions (i.e. words produced at fast speaking rates), stimulus variability produced significant decrements in identification performance.

Similar relationships between encoding difficulty and stimulus variability have also been reported for variations in talker characteristics. For example, Mullennix et al. (1989) reported a significant interaction between stimulus degradation level and the effects of variability; as encoding became more difficult, due to increased signal degradation, a significant increase in the effects of variability on word recognition scores was observed. Humes et al. (1992) also reported an interaction between encoding difficulty and the effects of talker variability using elderly hearing-impaired subjects. They found significant positive correlations between degree of hearing impairment and the effects of stimulus variability; as encoding difficulty increased due to poorer audibility of stimulus items, there was a significant increase in the negative effects of talker variability on identification scores. Considered together, these findings are consistent the proposal that the mechanisms mediating initial acoustic-phonetic encoding and those responsible for maintaining perceptual constancy are interrelated.

Other evidence, however, indicates a role of higher level (i.e. post-encoding) processes. For example, the present investigation found no significant interaction between S/N ratio and the effects of speaking rate variability. If the reductions in word recognition occurred exclusively because normalization was more difficult with increased encoding demands, then we would have expected significant interactions between signal-to-noise ratio and the effects of variability. The absence of such interactions for both talker and rate variability implicates post-encoding mechanisms as contributing to the observed reductions in spoken word recognition. Green and Miller (1985) also provided evidence for the contribution of higher-level processes in rate normalization by demonstrating rate-dependent phonetic processing even when the information about speaking rate was signaled visually. Thus, some effects of speaking rate normalization must occur at levels above those where auditory and visual information are integrated into a common representation in memory.

Evidence regarding the level at which the mechanisms mediating rate and talker normalization operate, therefore, indicates a role for both lower-level acoustic-phonetic encoding and higher-level post-encoding processes. Although additional experiments are necessary to assess the relative contributions each makes to maintaining perceptual constancy, the present findings argue against a single locus for normalization capacities. Instead, the mechanisms used to compensate for acoustic-phonetic variability appear to distributed across at least two levels of the speech perception system.

---

[6]Hearing loss in these subject did not exceed 20 dB HL for frequencies of 4 kHz and below.

The findings obtained in the present set of experiments demonstrate the importance of variability in speech perception. In the past, little, if any, attention was devoted directly to the study of variability in speech. Indeed, substantial efforts were made to reduce or eliminate as many sources of variability as possible. The assumption underlying this approach was that stimulus variability was a source of noise that should be reduced or eliminated. The present results on speaking rate, taken together with previous studies from our laboratory , show that this long-standing assumption about variability in speech is wrong and that we need to gain a better understanding of the role that variability plays in speech perception and spoken word recognition.

References

ANSI (1971). Method for measurement of monosyllabic word intelligibility. American National Standard, S3.2-1960 (R1971).

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

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

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

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

Gerstman, L. (1968). Classification of self-normalized vowel. IEEE-AU, **16**, 78-80.

Green, K.P. & Miller, J.L. (1985). On the role of visual rate information phonetic perception. *Perception and Psychophysics*, **38**, 269-276.

Humes. L.E., Davidson, S., Pisoni, D.B., Christopherson, L., & Mullennix, J.W. (1992). Effects of stimulus uncertainty on word-recognition performance of hearing impaired elderly listeners. *M*anuscript submitted to *Journal of Acoustical Society of America.*

Johnson, T.L. & Strange, W. (1982). Perceptual constancy of vowels in rapid speech. *Journal of Acoustical Society of America*, **72**, 1761-1770.

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

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

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

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1965). Perception of the Speech Code. *Psychological Review*, **74**, 431-461.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (1989). Effects of talker variability on recall of spoken word lists. *Perception and Psychophysics*, **15**, 676-684.

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*, **85**, 365-378.

Miller, J.L. (1981). Effects of speaking rate on segmental distinctions. In P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the Study of Speech*. Hillsdale, NJ: L. Erlbaum.

Miller, J.L. (1987). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the Psychology of Language*. Hillsdale, NJ: L. Erlbaum.

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

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

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

Nygaard, L.C., Sommers, M.S. & Pisoni, D.B. (1992). Effects of speaking rate and talker variability on the representation of spoken words in memory. In *Proceedings of the 1992 International Conference of Spoken Language Processing*. Alberta, Canada: University of Alberta.

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

Pickett, J.M. and Decker, L.R. (1960). Time factors in perception of a double consonant. *Language and Speech*, 3, 11-17.

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

Port, R. F. (1976). The influence of speaking tempo on the duration of stressed vowel and medial stop in English Trochee words. Unpublished Doctoral dissertation. University of Connecticut.

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

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting and Knowing*. Potomac, MD: L. Erlbaum.

Sommers, M.S. & Humes. L.E. (1993). The effects of speaking rate and stimulus variability on the perception of spoken words by young and elderly subjects. Paper presented at the 125th meeting of the *Acoustical Society of America*, Ottawa, Ontario Canada.

Summerfield, Q. & Haggard, M.P. (1972). Speech rate effects in the perception of voicing. In *Speech synthesis and Perception* (No: 6). Psychology Laboratory, University of Cambridge.

Summerfield, Q. (1974). Towards a detailed model for the perception of voicing contrasts. In *Speech Perception*, (No: 3). Department of Psychology, Queen's University of Belfast.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1074-1095.

Tomiak, G.R., Green, K.P. & Kuhl, P.K. (1991). Phonetic coding and its relationship to talker and rate normalization. Paper presented at the 122nd meeting of the *Acoustical Society of America*, Houston, Texas.

Uchanski, R.M., Miller, K.M., Reed, C.M., & Braida, L.D. (1992). Effects of token variability on resolution for vowel sounds. In M.E.H. Schouten (Ed.), *The Auditory Processing of Speech: From Sounds to Words*. New York: Mouton de Gruyter.

Verbrugge, R.R. & Shankweiler, D. (1977). Prosodic information for vowel identity. *Journal of Acoustical Society of America*, **61**, S39.

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 Acoustical Society of America*, **60**, 198-212.

Watson, C.S. & Kelly, W. (1981). The role of stimulus uncertainty in the discrimination of auditory patterns. In D. J. Getty and J.H. Howard (Eds.), *Auditory and Visual Pattern Recognition*. Hillsdale, N.J.: Erlbaum.

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Some Contributions of Auditory Psychophysics to Theoretical Issues in Speech Perception[1]

Mitchell S. Sommers and David B. Pisoni

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

[1]Preparation of this chapter was supported by NIH Research Grant DC-00111-16 and NIDCD Training Grant DC-00012-13 to Indiana University in Bloomington, IN.

Abstract

This review evaluates the contributions that auditory psychophysics has made to understanding four long-standing issues in speech perception: the lack of acoustic-phonetic invariance, perceptual normalization of speech, the internal representation of speech, and the minimal units of perceptual analysis. While the analyses suggest that psychophysical investigations have not provided comprehensive accounts for any of these issues, the available evidence indicates that psychoacoustic research has contributed significantly to our understanding of several research problems in each area. The results are discussed with regard to future empirical and theoretical directions for integrating auditory psychophysics and spoken language processing.

Some Contributions of Auditory Psychophysics to Theoretical Issues in Speech Perception

INTRODUCTION

Most contemporary models of spoken language processing (see Klatt, 1989 for a review) share the assumption, either implicitly or explicitly, that at least two distinct perceptual stages are required to transform speech signals into linguistically-meaningful phonetic percepts. The first consists of an analysis by the peripheral auditory system that results in a detailed spectral and temporal representation of the input waveform. The second, often referred to as the "phonetic stage", converts the spectro-temporal auditory representation into a form that can be used to derive both linguistic and indexical (i.e., personal) attributes of the speaker's message. The purpose of the present review is to evaluate the contribution that auditory psychophysics, which is concerned almost exclusively with the former of these processing stages, has made to understanding speech perception.

Both of these topics, auditory psychophysics and speech perception, have enjoyed extensive research histories. The volume and diversity of research in each area, however, makes any comprehensive integration of the two fields beyond the scope of a single review article (for more complete reviews see volumes edited by Schouten, 1987; 1992; Yost & Watson, 1988). In order to focus the review, the present discussion will be limited to assessing the contribution that research in auditory psychophysics has made to four long-standing theoretical issues in speech perception (Klatt, 1979; Pisoni, 1985; Luce & Pisoni, 1987; Studdert-Kennedy, 1976). These four issues are: (1) the lack of acoustic-phonetic invariance; (2) perceptual normalization; (3) the internal representation of speech sounds; and (4) the units of perceptual analysis.

The choice of these four issues as the focus of the review will, of necessity, exclude many other important theoretical questions in the area of speech perception and spoken language processing (see Klatt, 1979; 1989; Luce & Pisoni, 1987). The criteria for selecting these four topics were that, not only have they been identified as issues that all theories of speech perception must eventually address but they are also amenable to psychophysical analysis. This latter point is particularly important because a number of issues in the field of speech perception and spoken language processing do not lend themselves easily to investigations using psychoacoustic methodologies. For example, questions regarding the storage and retrieval of lexical items, mechanisms of lexical search, and the contribution of cognitive factors to spoken language comprehension, are more appropriately addressed by psycholinguistic than psychoacoustic experimentation. These topics concern issues that are more directly relevant to phonological and lexical processes rather than sensory encoding of speech signals by the auditory system.

The emphasis on these four issues should not be taken as minimizing the importance of psychoacoustic investigations for understanding speech processing. As noted earlier, almost all models of speech perception posit that the outputs of auditory analyses serve as the input for phonetic processing. Therefore, theories of spoken language processing are, to some extent, both determined and limited by the basic psychophysical capabilities of the human auditory system. Psychophysical accounts of categorical perception (Howell & Rosen, 1984; Pastore, 1981; Pisoni, 1977), trading relations (Hillenbrand, 1984; Parker, 1986) talker and rate normalization (Bladon, Henton, & Pickering, 1984; Diehl, Souther & Convis, 1980) and phonetic contrasts (Diehl, Kluender, & Walsh, 1990; Lautner & Hirsh, 1985; Pastore, 1987) have provided converging evidence that a number of phenomena observed in speech perception can be accounted for by known properties of the peripheral auditory system. The present review will extend these kinds of analyses and evaluate the contribution

that auditory psychophysics has made to the four issues outlined above. While this approach is unlikely to result in comprehensive accounts of any of the topics, our review will consider empirical findings concerning both the adequacy and limitations of psychoacoustic explanations for speech perception.

Suitability of the Auditory System for Processing Speech Signals

A prerequisite for any psychophysical account of speech perception is to provide evidence that the auditory system is capable of detecting and discriminating signals with acoustic properties characteristic of speech sounds (Stevens, 1981). In the absence of such demonstrations, it is unlikely that auditory psychophysics will be able to contribute significantly to questions of speech processing. Fortunately, a number of studies have demonstrated strong correlations between basic auditory capabilities and the acoustic properties of speech signals. For example, speech intelligibility is largely determined by information contained within the frequency range of 1-4 kHz and this is precisely the spectral region where humans exhibit greatest absolute sensitivity (Dadson & King, 1952). It is also the region where listeners show the highest relative acuity for frequency discriminations (Wier, Jesteadt, & Green, 1977) and for resolving the frequency components of complex stimuli (Zwicker, 1970)[2]. Other studies (see Moore, 1989 for a review) have demonstrated that temporal resolution for both narrow- and wide-band stimuli is at or near maximum in the 1-4 kHz region. Taken together, these findings suggest that the human auditory system is well-designed for detecting and resolving the spectral and temporal components of speech signals[3].

Additional support for the convergence of auditory and linguistic processing comes from studies of the acoustic cues underlying phonetic distinctions. Both spectral (Blumstein & Stevens, 1979; Dorman & Raphael, 1980; Fant, 1960; Peterson & Barney, 1952) and temporal (Kewley-Port, 1983; Lisker & Abramson, 1970) properties of the speech waveform have been shown to be important for signaling phonetic differences. For example, vowel quality and consonantal place of articulation are both, at least partially, mediated by detecting and encoding spectral changes in the speech waveform (Blumstein & Stevens, 1979; Fant, 1960; Peterson & Barney, 1952; Shepard, 1972). Temporal factors, on the other hand, have been demonstrated (Liberman, Delattre, Cooper & Gertsman, 1956; Lisker & Abramson, 1970; Port & Dalby, 1982) to be a principle cue for distinguishing phonetic features such as voicing and nasality. The auditory system, therefore, must possess both good temporal and spectral resolution to adequately encode the complete range of phonetic contrasts. However, in most filter-based systems, such as the auditory system (Fletcher, 1940), there is a tradeoff between precise temporal and spectral analysis; good spectral resolution is attained only at the expense of poorer temporal processing capabilities and vice-versa. Such a system would clearly be inadequate for purposes of speech perception.

The human auditory system resolves this conflict in a number of ways. First, auditory-filter bandwidth increases as a function of frequency. Thus, good spectral resolution is maintained at low frequencies, where it is necessary to resolve the lowest resonances (formants) of the vocal-tract, while good temporal resolution is available in higher frequency regions (Searle, Jacobson, & Kimberley, 1980). Secondly, the low-frequency components of speech partially excite regions of the cochlea that mediate high-frequency transduction. This allows precise temporal resolution of low-frequency information because of the relatively wide bandwidths of auditory filters in this spectral region.

---

[2]As measured by the critical band or critical ratio.

[3]Stevens (1981) and Lieberman (1979) have both suggested that languages have evolved to take advantage of the natural sensitivities of the auditory system. Therefore, this correspondence between auditory capabilities and the acoustic properties of speech signals is not unexpected.

Finally, compared to other filter systems, the cochlea has a remarkably high damping coefficient (Bailey, 1983). This serves to reduce the overly-long ringing that is characteristic of many systems and functions to reduce the blurring of temporal information. Taken together, these findings suggest that the human auditory system appears to be extremely well adapted for coding the detailed spectral and temporal information necessary for the perception of speech.

In the remaining sections, we will evaluate the extent to which psychophysical accounts can address some of the long-standing theoretical issues in speech perception. A brief description of each problem will be given first followed by a review of the relevant psychophysical data that can be used to assess the problem.

LACK OF ACOUSTIC-PHONETIC INVARIANCE

The lack of acoustic-phonetic invariance in speech refers to findings from a considerable number of studies (Klatt, 1976; Liberman et al., 1967; Peterson and Barney, 1952) showing that there is no consistent mapping between acoustic signal and phonetic percept. Rather, most investigations of acoustic-phonetic invariance have demonstrated that the acoustic realization of a given speech sound is affected by factors such as speaking rate, phonetic context, and vocal-tract characteristics (Klatt, 1976; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Peterson & Barney, 1952). Psychophysical investigations of acoustic-phonetic invariance have generally taken one of two related approaches in trying to identify invariant correlates of phonetic distinctions. The first has focused on acoustic analyses of the speech signal itself and has attempted to identify a set of acoustic correlates in the speech waveform that uniquely and consistently specify a given phoneme independent of context (Blumstein & Stevens, 1979; Cooper, 1950; Cooper, Delattre, Liberman, Borst & Gertsman, 1952; Delattre, Liberman & Cooper, 1955; Kewley-Port, 1983; Kewley-Port, Pisoni & Studdert-Kennedy, 1983; Stevens and Blumstein, 1981). Psychoacoustic considerations are essential to this approach because, in addition to establishing the invariant nature of the proposed speech cue(s), these analyses must also demonstrate that the auditory system is sufficiently sensitive to use differences in the hypothesized properties as a reliable basis for phonetic classifications.

A second, but related, methodology for investigating acoustic-phonetic invariance has focused on properties of the auditory system rather than the speech signal. The goal here is to find correlations between auditory sensitivity and the acoustic cues to phonetic distinctions. Proponents of this approach argue that previous attempts to establish invariant properties for speech (Cooper, 1950; Cooper et al., 1952; Peterson and Barney, 1952) have been largely unsuccessful because they failed to consider the relative sensitivity of the auditory system. Specifically, the claim is that natural discontinuities in auditory sensitivity exist and that languages have evolved to exploit boundary regions where sensitivity and acuity are greatest (Stevens, 1972; 1981). These natural sensitivities may, therefore, serve as the basis for a set of invariant acoustic cues to phonetic distinctions. Diehl et al. (1990), for instance, have shown that only a fraction of the acoustic features that are physiologically achievable have been incorporated into linguistic systems. They argue that languages have universally adopted this small subset of features for distinguishing phonemes because they offer maximum discriminability given the capabilities of the auditory system. Investigations of auditory sensitivity and parallel analyses of speech waveforms have been used to identify this restricted set of acoustic cues and determine the extent to which they remain invariant across contexts (see Summerfield & Bailey, 1977; Summerfield & Haggard, 1977).

Investigations of Acoustic-Phonetic Invariance in Speech Signals

Early research on speech perception (Cooper, 1950; Delattre et al., 1955) was concerned almost exclusively with the former of the two approaches outlined above, namely, the search for invariant acoustic cues within the speech waveform itself. The rationale for these investigations was that if a set of invariant cues for phonetic distinctions could be found in the signal, then the mapping from acoustic waveform to phonetic category could be accomplished by a relatively simple set of pattern recognition mechanisms. Initial studies on this problem began with development of two instruments--the sound spectrograph which allowed researchers to visually examine a time-varying spectral display of the acoustic waveform (Delattre et al. 1955; Liberman, Delattre, Cooper, and Gertsman, 1954) and the pattern playback (Cooper, 1950) which permitted an experimenter to manipulate a set of acoustic parameters to produce synthetic speech stimuli for perceptual tests.

Figure 1 displays a typical set of synthetic stimuli used in early experiments on the relationship between acoustic properties and phonetic identifications (Liberman et al., 1967). The figure shows highly schematized versions of the formant transitions and steady-state values sufficient for synthesizing the phoneme /d/ before a number of different vowels. Several important findings on the relations between acoustic cues and phoneme perception can be illustrated by these examples.

Insert Figure 1 about here

First, consider the endpoint stimuli /di/ and /du/. Perceptual experiments have shown that the identical /d/ percept is signaled by rising formant transitions before /i/ and falling transitions before /u/. Other perceptual investigations have shown that it was not possible to isolate a segment of the signals that, when played to listeners, would give a distinct /d/ percept. As the stimuli were progressively segmented from the right, listeners first perceived a /d/ plus vowel and then a nonspeech sound (Liberman et al., 1967). At the time, these were considered unusual findings because the prevailing theories suggested that speech was composed of discrete units with invariant and reliable acoustic properties (Licklider, 1952). The task of speech researchers, according to this view, was simply to specify the properties underlying each phonetic unit and determine how the units were combined. The findings reviewed by Liberman et al. (1967) demonstrated that such an approach was fatally flawed. Results from other investigations have shown that the acoustic cues to phonemes also change as a function of phonetic environment, vocal-tract size (Peterson & Barney, 1952) and speaking rate (Miller & Liberman, 1979) among other factors.

These initial unsuccessful attempts at demonstrating acoustic-phonetic invariance led to the development of the motor theory (Liberman et al., 1967), in which speech-specific specialized neural mechanisms were proposed to account for perceptual constancy. The reasoning behind this proposal was that, given the absence of a consistent mapping between acoustic properties and phonetic percepts, listeners must depend on mechanisms specialized for speech perception to accommodate the high degree of acoustic variability in speech signals. The motor theory (Liberman et al., 1967; Liberman & Mattingly, 1985) proposes that listeners perceive speech by referencing the underlying articulatory gestures with processing mechanisms devoted exclusively to speech.

This is not to suggest, however, that the search for acoustic invariants has been completely unsuccessful. Indeed, Several investigations (Blumstein & Stevens, 1979; Kewley-Port, 1983; Kewley-Port et al., 1983; Stevens & Blumstein, 1981), using digital signal processing techniques, have found

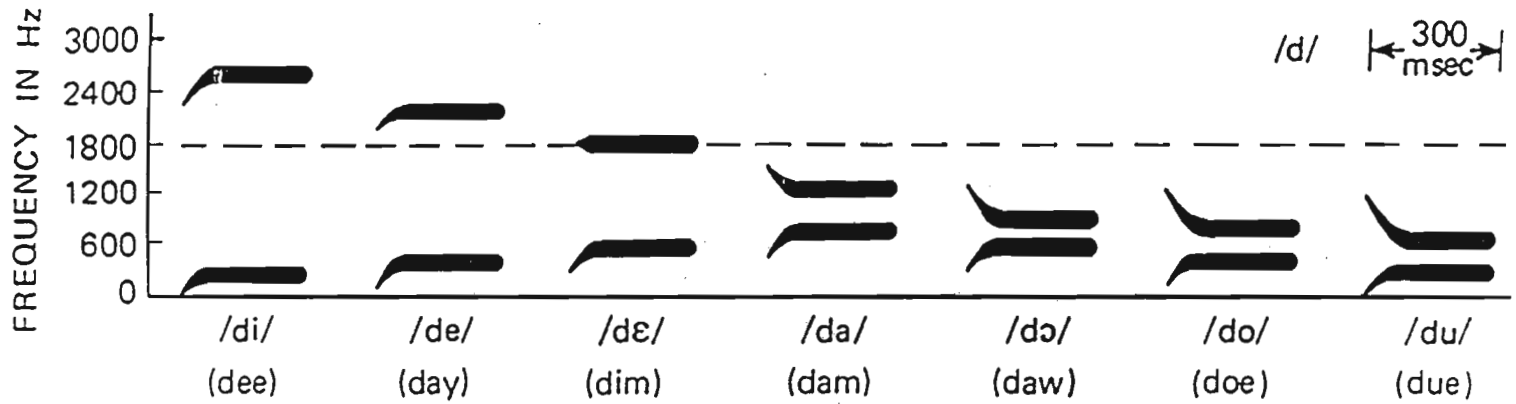


Figure 1: Schematized spectrograms of first (F1) and second (F2) formants of /d/-vowel CV syllables. (Adapted from Liberman et al., 1967).

reliable invariant cues to consonantal distinctions within the short-time amplitude spectrum of the waveform. Figure 2, taken from Blumstein & Stevens (1979), provides an example of this approach for identifying cues to place of articulation in stop consonants.

Insert Figure 2 about here

The six spectra are from synthetic versions of the voiced, /b/, /d/, /g/, and voiceless, /p/, /t/, /k/, stop consonants sampled with a 26-ms time window following the release of the consonant into a following vowel[4]. The figure shows that the spectra for the labial, syllable-initial, stop consonants, /b/ and /p/, can be characterized as diffuse-flat or diffuse falling. That is, the first three formants are relatively evenly spaced (diffuse) and the spectrum has an overall flat or downward tilt (falling). The alveolar consonants, in contrast, are characterized by a diffuse-rising spectrum while the velar consonants have a compact spectral shape (second and third formants close in frequency) and a prominent mid-frequency peak. Acoustic measurements (Blumstein & Stevens, 1979; Stevens & Blumstein, 1981) indicate that these overall spectral shapes are maintained independently of vowel context and vocal-tract characteristics and therefore provide a potentially invariant cue to place of articulation in stop consonants. Perceptual investigations (Blumstein and Stevens, 1980) have demonstrated that listeners can achieve over 85 percent correct identification for place of articulation, independent of vowel-context, using only gross spectral information sampled within the first 26-ms following consonantal release[5].

Additional studies on stop consonants (Kewley-Port, 1983; Kewley-Port et al., 1983) have shown that dynamic, rather than static, properties of consonantal spectra are more important for distinguishing place of articulation. Kewley-Port (1983) proposed that three features: spectral-tilt at onset, timing between release burst and onset of low-frequency energy, and presence/absence of a mid-frequency peak, as measured in continuous 5-ms samples of the signal, serve as more reliable invariant cues to stop-consonant place of articulation than those suggested by Blumstein and Stevens (1979). To support this hypothesis, Kewley-Port et al. (1983) demonstrated that listeners identified place of articulation significantly better from isolated CV stimuli that preserved dynamic acoustic properties in the spectra than from those based on static onset characteristics alone. Moreover, Kewley-Port and Luce (1984) have shown that the time-varying features providing invariant cues to place of articulation in syllable-initial stop consonants are maintained when the signals are produced in fluent speech and at several different speaking rates. Additional findings from Kurowski & Blumstein (1987), demonstrating that short-term temporal characteristics of the spectra may provide invariant cues to place of articulation for nasal consonants, also support the importance of dynamic spectral properties as cues to phonetic distinctions.

Although perceptual investigations (Blumstein & Stevens, 1980, Kewley-Port et al., 1983) have found that listeners can use differences in the gross spectral properties as "context-independent" cues to place of articulation, the question remains whether the auditory system is sufficiently sensitive

---

[4]The spectra have been smoothed using a linear prediction (LPC) algorithm. Stop-consonant discrimination based on smoothed spectra are consistent with the finite bandwidths of auditory filters at low frequencies.

[5]Blumstein and Stevens (1979) and Stevens and Blumstein (1981) have suggested that overall spectral shape provides the principle cue for place of articulation in stop consonants. Other acoustic correlates of the distinctions, such as formant-transition direction or formant-onset frequency, which are not context-independent, can be used as secondary cues in cases where the primary cue is weakened or ambiguous.

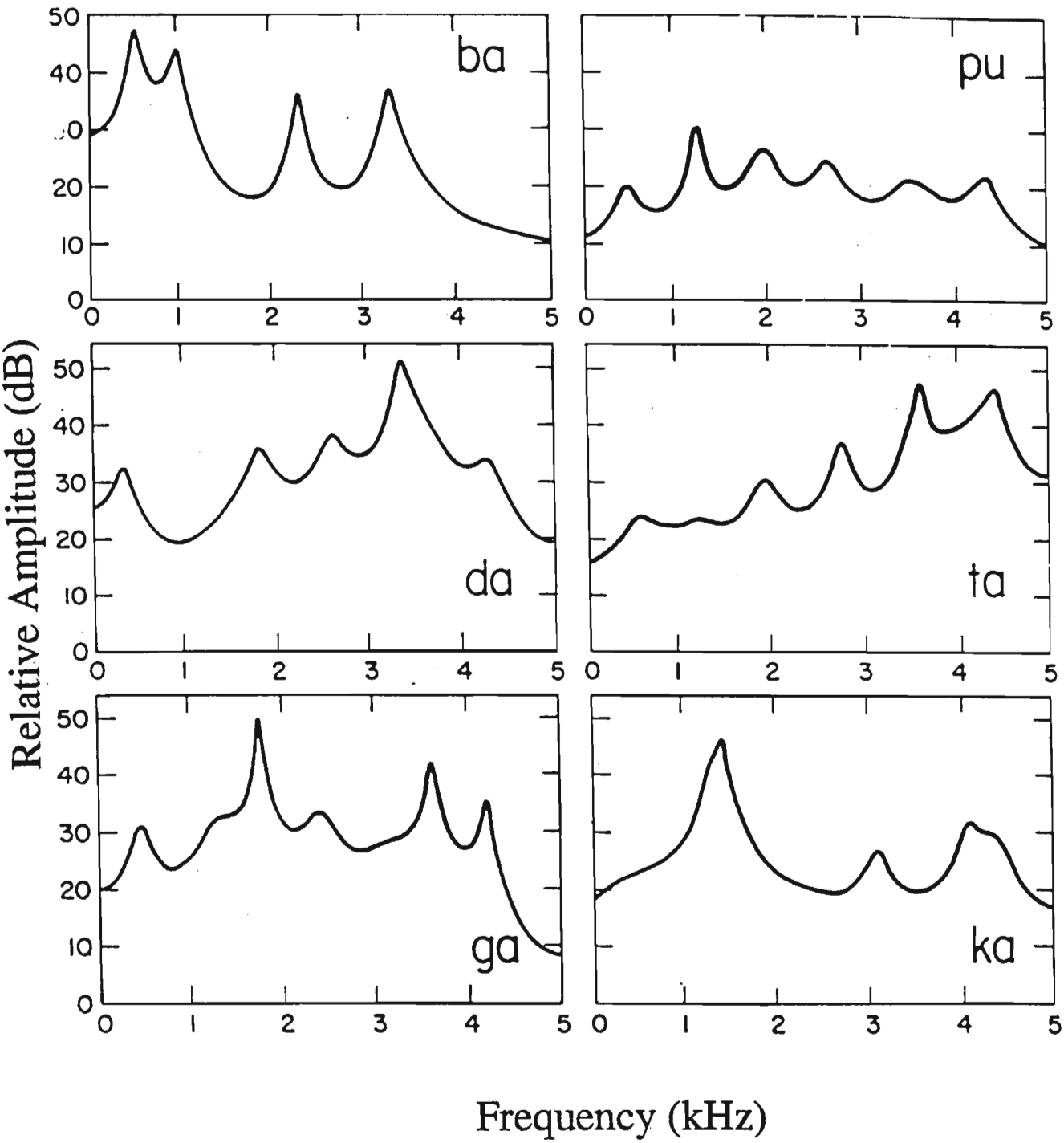


Figure 2: Smoothed spectra of syllable-initial stop consonants sampled with a 25.6 ms time window. The spectra are for the first 26 ms after consonantal release (From Blumstein and Stevens, 1979).

to such spectral alterations for them to serve as the exclusive means of discriminating and categorizing speech sounds. Recent "profile analysis" experiments by Green and his colleagues (Green, Kidd and Picardi, 1983; Green, 1988) provide strong evidence that human sensitivity to changes in spectral shape is at least two to three times greater than necessary for discriminating the static and dynamic changes that have been proposed as invariant cues to place of articulation (Blumstein and Stevens; 1979; Kewley-Port, 1983). In a typical profile analysis experiment, listeners are required to detect an intensity increment in one or more components of complex stimuli with equal-amplitude, logarithmically-spaced, components. Subjects must make their decisions regarding the presence or absence of the increment on the basis of spectral shape, rather than by detecting level changes in individual components, because the overall amplitudes of the stimuli are randomly varied over a 20-40 dB range.

Green & Kidd (1983) reported that listeners were able to detect spectral-shape changes that were produced by incrementing the level of the stimulus component nearest to 1 kHz by approximately 1.1 dB[6]. Spectral shape discrimination was even better (average thresholds for the middle component of a profile complex of approximately 0.8 dB) when there were simultaneous changes in several components of the stimuli. This level of sensitivity for detecting spectral shape alterations would provide considerably greater acuity than necessary for distinguishing place of articulation differences. For example, examination of figure 2 shows that the difference in spectral shape between syllable-initial /b/ and /d/ within the 1-kHz region is approximately 4 dB. This difference is on the order of four times greater than listeners' thresholds for detecting spectral shape changes in this frequency region (Green & Kidd, 1983). Furthermore, the spectral changes proposed as context-independent cues to place of articulation occur over a wide spectral region of the stimulus. Given that thresholds for spectral shape discrimination are better when there are simultaneous changes in a number of frequency regions (Green, 1988), this estimate of four-fold greater sensitivity than necessary is a very conservative one. Profile analysis experiments, therefore, provide additional evidence that listeners are capable of utilizing spectral shape differences as a basis for discriminating and categorizing place of articulation in syllable-initial stop consonants.

The studies summarized above support the proposal that the perception of place of articulation distinctions in stop consonants may be mediated, at least in part, by detecting invariant characteristics of the spectral envelope. However, there are also several difficulties with this account. First, Blumstein and Stevens (1980) found that 85% of a sample of 1800 naturally-produced CV tokens were accurately classified with respect to place of articulation using the overall spectral shapes that they proposed. This level of performance is significantly worse than has been demonstrated for natural CV stimuli where categorizing speech sounds according to place of articulation generally exceeds 95%. Kewley-Port et al. (1983) directly compared identification performance for place of articulation in synthetic CV stimuli when: 1) place of articulation was cued by static spectral characteristics (as suggested by Stevens & Blumstein, 1981); 2) place differences were cued by static and dynamic properties (as suggested in Kewley-Port, 1983); or 3) listeners heard natural tokens of CV stimuli. Although performance in condition 2 (static and dynamic cues) was significantly better than in the first condition (static alone), identification accuracy, even with temporal information present, was significantly worse than with natural CV stimuli (see also Walley & Carrell, 1983). Thus, identification performance based exclusively on differences in gross spectral shape was not sufficient to account for human abilities to identify place of articulation differences in stops.

---

[6]Green's metric for profile analysis is the level of the increment at threshold relative to the level of the component to which it was added. The values reported here, however, are given in the more conventional dB scale to facilitate comparison with other studies.

A second objection to Stevens and Blumstein's proposals that place of articulation distinctions are invariantly signaled by changes in overall spectral characteristics is that it remains unclear whether listeners are capable of extracting the necessary detailed spectral information at the high rates at which speech is transmitted (Liberman et al., 1967). For example, Kewley-Port and Luce (1984) obtained only 86% correct identification from visually-presented running spectra of stop-consonant-vowel syllables read in sentence context by two males and two females. As noted earlier, this is considerably below human performance for natural CV tokens which typically averages over 95%. Furthermore, Kewley-Port and Luce did not impose time limitations for judging the place of articulation differences. It is unlikely, therefore, that performance under conditions of real-time speech analysis would approach even the modest levels found by Kewley-Port and Luce (1984) in their studies of connected speech. Additional perceptual studies investigating identification performance at rates comparable to that observed for fluent speech are necessary before reaching conclusions about the relative importance of short-term spectral and temporal cues for phonetic distinctions.

Studies of Natural Auditory Sensitivities

As noted earlier, a second area of psychophysical research relevant to issues of acoustic-phonetic invariance are a series of studies that have examined the relationship between auditory sensitivity and the cues for phonetic distinctions. Demonstrations of a correspondence between psychoacoustic capabilities and cues for phonetic classifications suggest that speech perception may be partly mediated by detecting and discriminating a relatively restricted set of invariant acoustic properties to which the auditory system is most sensitive. These arguments are based on demonstrations that languages have evolved to take advantage of both auditory and articulatory constraints (Diehl et al., 1990; Liljencrants and Lindblom, 1972). Diehl et al. (1990) suggested that the acoustic cues signaling phonetic distinctions have evolved from a much larger inventory of possible acoustic properties specifically because they are the ones most discriminable by the human auditory system. If this is the case, then listeners should base phonetic judgments on this set of highly discriminable acoustic properties independent of context. Measuring relative sensitivity for an array of potential cues would therefore provide a means of determining which are likely to serve as such context-independent phonetic features. Stevens (1972) has proposed a "quantal theory of speech perception" which argues that, for several articulatory gestures, a range of values exists over which the gestures can change without producing significant acoustic variations. Furthermore, a corresponding set of values exists where relatively small variations will have large acoustic consequences. Stevens claims that languages have evolved such that phonetic categories fall within the former of these regions and phonetic boundaries occur at the latter independent of context.

Several investigations (Cutting & Rosner, 1974; Miller, Wier, Pastore, Kelly, & Dooling, 1976, Pisoni, 1977) have documented similarities in the categorization of speech and nonspeech continua that may be a consequence of natural auditory sensitivities. The rationale for these experiments is that if listeners categorize acoustically-similar speech and nonspeech signals in a comparable manner, then the categorization may be mediated by basic properties of the auditory system as opposed to speech-specific mechanisms. For example, when presented with a series of synthetic CV stimuli in which the interval between consonantal release and voicing onset, the voice-onset-time (VOT), is varied from 0-80 ms, listeners identify stimuli with VOTs less than approximately 20-30 ms as a voiced consonant and those with VOTs greater than this as voiceless (Lisker & Abramson, 1970). Is the voiced-voiceless distinction signaled by an invariant acoustic property that results from a discontinuity in temporal resolution at approximately 20-30 ms? Pisoni (1977) addressed this issue by synthesizing two-tone non-speech patterns that varied in the relative onset times of the two component

sinusoids. He found that discrimination was maximal, independent of the frequency of the leading tone, when onset asynchrony was approximately 20 ms. Miller et al. (1976) found a discrimination peak for detecting onset asynchronies in noise-buzz sequences at approximately 16 ms. These findings are consistent with the proposal that the voiced-voiceless distinction in speech is signaled by an acoustic invariant; stimuli with VOTs less than 20 ms are perceived as voiced while those with VOTs greater than 20 ms are perceived as voiceless. Furthermore, this invariance arises because the auditory system is maximally sensitive to differences in the onset of two acoustic events, either speech or nonspeech, of approximately 20 ms (Hirsh, 1959).

A similar correspondence between auditory sensitivity and phonetic classifications has been reported by Jamieson (1987) for categorizing CV syllables differing in the manner of articulation of the initial consonant. Formant transitions between appropriate frequencies are perceived as the semi-vowel /wa/ if the transitions last longer than 40-60 ms but are perceived as the stop consonant /ba/ if the duration of the initial transition is shorter than this value. Jamieson & Slawinska (1983) measured discrimination thresholds for nonspeech analogs of these stimuli that consisted of a frequency glide followed by a steady-state tone. They found that discrimination of glide duration was maximum for standards of approximately 50 ms. Note that this is not what would be predicted by Weber's law which would suggest maximum discriminability for shorter-duration glides. Thus, these results suggest another possible acoustic-phonetic invariant; formant transitions less than 40 ms are perceived as one manner of articulation (/ba/) while those greater than 60 ms are heard as a different manner (/wa/).

Despite the relationship between certain natural sensitivities and the cues for phonetic distinctions, a number of difficulties can be found with general accounts of speech perception that are based on acoustic-phonetic invariances as determined by auditory capabilities. First, the results of several investigations (Lisker & Abramson, 1971, Miller, 1981) suggest that the acoustic properties proposed above as invariants for voicing and manner distinctions (VOT and formant transition durations, respectively) are not context independent. For example, Lisker & Abramson (1971) demonstrated that VOT boundaries change as function of both place of articulation and linguistic experience. Figure 3 shows the results of one of their perceptual experiments.

Insert Figure 3 about here

The figure displays VOT category boundary values as a function of place of articulation for three different languages. Examination of the figure shows that the VOT boundary for velars can be almost twice as long as the boundary for labials. Furthermore, within a given place of articulation, the VOT boundary varies significantly for English, Spanish and Thai listeners. Brady and Darwin (1978) reported that the VOT boundary between syllable-initial alveolar stops (/d/ vs. /t/) is not a constant 20 ms, as would be required by a theory of acoustic invariance, but changes as a function of the range of stimuli presented during a testing session. Miller (1981) and Miller and Liberman (1979) have also shown that the transition duration for the boundary between /ba/ and /wa/ changes as a function of speaking rate. Note that these findings do not argue against psychophysical discontinuities within the auditory system. They do, however, suggest that such discontinuities are unlikely to be the principle basis for these phonetic distinctions which appear to be dependent on properties of the surrounding phonetic context..

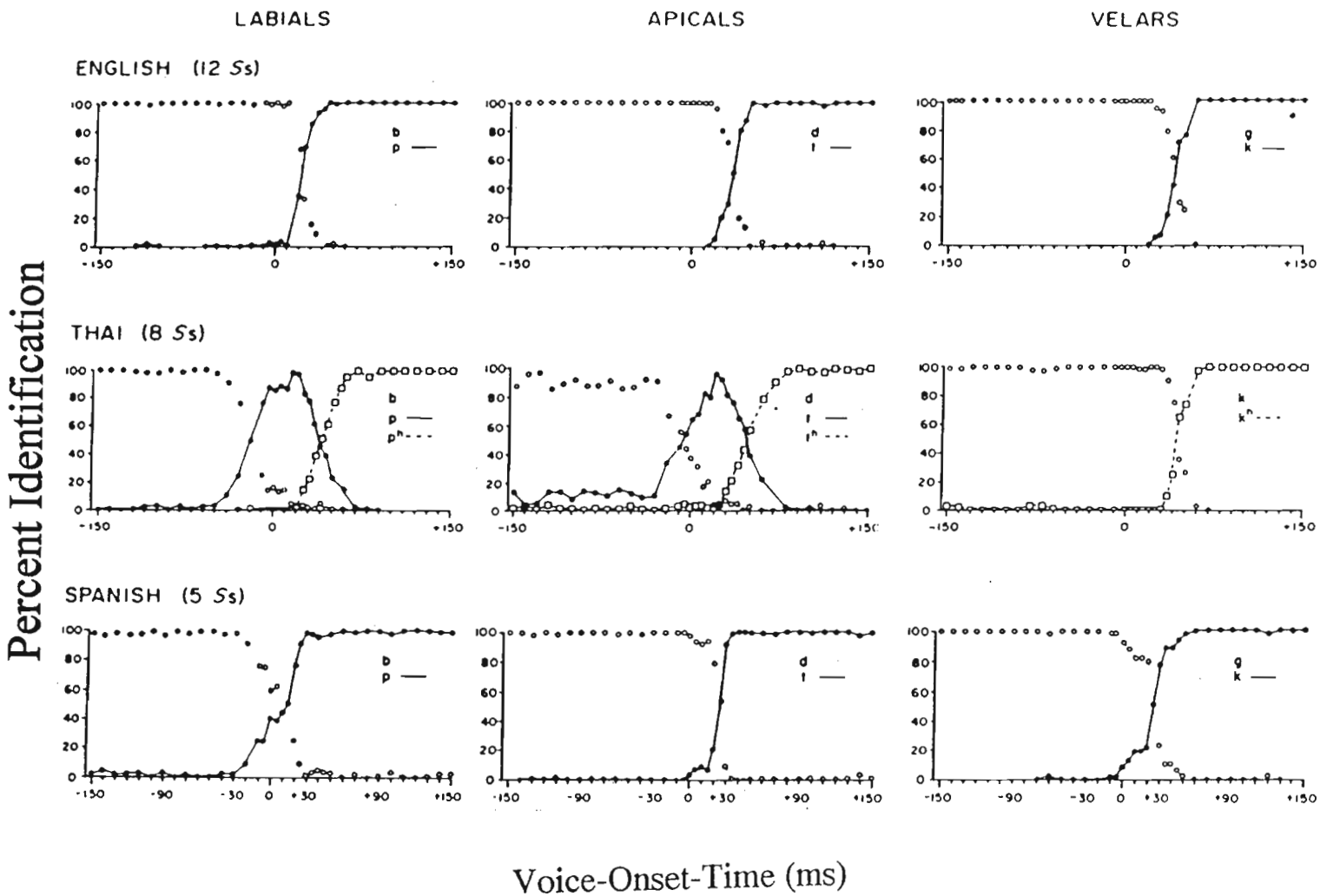


Figure 3: Labeling functions for syllable-initial stop consonants in English, Spanish, and Thai as a function of place of articulation. (Adapted from Lisker & Abramson, 1964).

Correlations Between Changes in Psychoacoustic Capacities and Speech Perception

A second set of findings that presents potential difficulties for any acoustically-based theory of speech perception is the poor to modest correlations between reductions in psychoacoustic abilities and changes in speech perception performance. If phonetic distinctions are mediated by detecting and recognizing invariant acoustic properties, then reductions in the ability to extract those properties, as has been observed in listeners with reduced frequency or temporal resolution, should result in decreased speech perception abilities. In general, however, strong correlations between reduced spectral and temporal resolution and changes in speech intelligibility scores have not been demonstrated (Dubno & Dirks, 1989; Glasberg & Moore, 1989; Humes & Christopherson, 1991; Turner & van Tasell, 1984; Tyler, Summerfield, Wood, & Fernandes, 1982).

Tyler et al. (1982), for instance, found reduced temporal and spectral resolution in a group of mild to moderately hearing-impaired listeners. Identification of a VOT continuum in these subjects, which could be expected to reveal any dependence of discrimination on spectro-temporal characteristics of the signal, was not significantly different than for a group of normal-hearing listeners. In another study, Dubno & Dirks (1989) measured auditory filters and consonant recognition in hearing-impaired listeners at levels that gave identical predicted performance on the articulation index (AI). Thus, listeners in this study were tested under conditions in which the spectral regions most important for speech perception were approximately equally audible. Differences between observed and predicted AI scores were not correlated with any of the auditory-filter characteristics. The results of this latter investigation are particularly troublesome for accounts of speech perception based exclusively on acoustic-phonetic invariance. In some cases, auditory-filter widths for hearing-impaired listeners were 5-6 times those found for normal-hearing listeners yet little if any association was found between filter width and consonant recognition performance.

These findings for hearing-impaired listeners are difficult to reconcile with proposals that acoustic-phonetic invariance in the gross spectral properties of speech signals (Stevens & Blumstein, 1981; Kewley-Port, 1983) serve as the basis for phonetic distinctions because the wider auditory-filter widths would have severely distorted overall spectral shape. Van Tasell, Hagen, Koblas, and Penner (1982) directly investigated the use of short-term spectral cues for identifying stop-consonant place of articulation by hearing-impaired listeners. Their results demonstrated that post-training identification performance, based on gross spectral properties sampled immediately after consonantal release, did not differ between normal-hearing and hearing-impaired listeners. Van Tasell et al. (1982) suggested that these findings indicated that various forms of signal distortion resulting from sensorineural hearing loss can have negligible effects on the use of short-term spectral cues to stop-consonant place of articulation.

This conclusion, however, may be premature in light of several findings regarding the effects of hearing loss on changes in frequency resolution (Peters & Moore, 1992; Sommers & Humes, in press). The average hearing loss for subjects in the Van Tasell et al. (1982) study was 34 dB. Peters and Moore reported that, for frequencies below 1 kHz, frequency selectivity, as measured by the widths of auditory filters, remained relatively unchanged for hearing losses less than approximately 30-35 dB. Sommers and Humes (in press) found no significant difference in the equivalent rectangular bandwidths (ERBs) of auditory filters at 2 kHz for hearing losses less than approximately 20 dB. Furthermore, there was only a slight increase in ERBs for losses as great as 30 dB. These findings suggest that the reduction in frequency resolution for subjects in the Van Tasell et al. (1982) investigation, where hearing losses averaged 34 dB, may have been relatively small and, consequently, did not produce substantial changes in the ability to extract short-term spectral properties of the signal.

The Dubno and Dirks (1989) findings of similar stop-consonant identification performance despite auditory filters that were five to six times greater than normal, therefore, remain problematical for accounts of speech perception based exclusively on identifying acoustic-phonetic invariances in the speech signal.

PERCEPTUAL NORMALIZATION IN SPEECH

The difficulties in finding context-independent correlates for phonetic categories has led many speech researchers (see Klatt, 1989) to propose a stage of perceptual processing during which the acoustic differences in speech signals due to factors such as phonetic environment, vocal-tract size, and speaking rate are adjusted or "normalized" to an idealized symbolic representational format that can be used in recognition. Although there has been a conspicuous absence of empirical research on perceptual normalization, most theories have considered it part of the phonetic, as opposed to psychoacoustic, analysis of speech (Liberman et al., 1967; Miller & Liberman, 1979). Normalization, according to these accounts, is conceptualized as a speech-specific processing stage that functions to maintain perceptual constancy despite extensive variation in the physical signal. A few investigators (Diehl et al., 1980; Jamieson, 1987; Diehl & Walsh, 1989; Pisoni, Carrell & Gans, 1983), however, have argued that many of the findings used to support a phonetic basis of normalization can be accounted for by considering results from auditory psychophysics. The purpose of this section, therefore, is to determine whether there may be a psychophysical basis for some of the normalization effects that have been reported in the literature.

One of the most well-documented examples of perceptual normalization comes from a series of experiments conducted by Miller and her colleagues (Miller & Liberman, 1979; Miller, 1981) on the perception of speaking rate. Miller, Grosjean & Lomanto (1984) reported an almost three fold change in articulation rate during the course of short conversations. Miller & Baer (1983) investigated whether these large differences in speaking rate altered the acoustic fine structure of speech signals. They examined the effects of speaking rate on the acoustic properties signaling the distinction between the stop-consonant /b/ and the semi-vowel /w/. The principle acoustic cues signaling this phonetic distinction are the amplitude and slope of the initial formant transitions; tokens with abrupt amplitude onsets and short transition durations are perceived as /b/ while those with more gradual onsets and longer transition durations are perceived as /w/. Miller & Baer (1983) asked speakers to produce tokens of the CV syllables /ba/ and /wa/ at several different speaking rates. Acoustic analyses of the productions demonstrated that, as speaking rate decreased, the duration of the initial transitions for /w/ became longer.

In a series of perceptual experiments, Miller & Liberman (1979) reported that listeners compensate for this "rate-dependent" change in a phonetically important acoustic property. They constructed a synthetic stimulus continuum that varied from /ba/ to /wa/ by successively incrementing the duration of the initial formant transitions. Perceived speaking rate was altered by increasing the duration of the steady-state portion of the vowel in the syllable. In labeling this continuum, listeners shifted the position of the category boundary as function of syllable duration (speaking rate); as syllable duration increased, the category boundary moved to longer initial transition durations. Miller & Liberman (1979) interpreted these results as evidence that listeners analyze the acoustic properties of speech relative to the prevailing speaking rate and compensate or normalize for changes in rate prior to making phonetic decisions.

Pisoni, Carrell & Gans (1983), however, reported results from a series of experiments suggesting that shifts in category boundaries with changes in speaking rate are not mediated by processing mechanisms devoted exclusively to speaking-rate normalization. They constructed nonspeech analogs of the stimuli used by Miller and Liberman (1979) and obtained results that paralleled those found with speech signals. Listeners were asked to identify stimuli that differed in the duration of a rapid spectrum change at onset as either "abrupt" or "gradual"; stimuli with short rise-times (small transition durations) were labeled as abrupt while those with relatively longer rise-times (larger transition durations) were labeled as gradual. To examine the effects of rate changes on the abrupt/gradual category boundary, Pisoni et al. varied the duration of a steady-state tone following the transition. Their results paralleled those of Miller and Liberman (1979); with shorter steady-state durations (faster speaking rates), they found a shift in the abrupt/gradual category boundary toward shorter transition durations.

Although the findings from Pisoni et al. (1983) demonstrated that the category boundary shifts observed by Miller & Liberman (1979) do not necessarily reflect the operation of speech-specific perceptual mechanisms, they did not establish specific factors that might mediate the effects of rate variation. Jamieson (1987) has suggested that the shifts in category boundaries for the /ba/-/wa/ continuum as a function of speaking rate can be attributed to backward masking of the formant transitions by the vowel. As the duration of the steady-state vowel increases, there is a greater amount of backward masking which makes the transitions perceptually shorter. This reduction in perceived transition duration results in more /ba/ responses since shorter transition durations are characteristic of a stop consonant. Jamieson, Johnson, & Rvachew, (1986) provided evidence to support this hypothesis by demonstrating that shifts in the category boundary, identical to those found by Miller & Liberman (1979), could be obtained for a /bad/-/wad/ continuum by reducing the amplitude of the vowel. According to the masking hypothesis, reducing the amplitude of the steady-state vowel should decrease masking of the transitions and therefore lead to a shift in the category boundary toward shorter transition durations (more /wad/ responses). The results confirmed this hypothesis and are therefore consistent with a psychophysically-based account of speaking rate normalization.

Diehl & Walsh (1989) have also recently proposed a psychophysical account of rate normalization. They suggested that the change in category boundary as a function of vowel duration that was observed by Miller & Liberman (1979) is due to the general auditory principle of durational contrast. According to this principle, the perceived duration of a segment, either speech or nonspeech, is affected contrastively by the duration of adjacent segments. Thus, a stimulus will be perceived as longer if it is followed by a shorter duration segment than if it is followed by a longer duration stimulus. The effects observed by Miller & Liberman (1979) can be accounted for, according to their proposal, as follows: an increase in vowel duration will decrease the perceived length of the initial formant transitions and, subsequently, more tokens from a /ba/-/wa/ continuum will be perceived as beginning with the syllable-initial stop. Diehl & Walsh (1989) provided support for this explanation by demonstrating category boundary shifts similar to those of Miller & Liberman (1979) with nonspeech stimuli. The signals in their experiment were sine-wave replicas of the first formant of Miller & Liberman's (1979) stimuli. The stimuli contained an initial linear increase in frequency from 250 to 750 Hz and then a variable-duration steady-state. A continuum was created by varying the duration of the transition from 15-65 ms and, as in the Pisoni et al. (1983) study described above, listeners were asked to categorize the stimuli as having either an abrupt (short duration transitions) or gradual (longer duration transitions) onset.

The results of this experiment paralleled those obtained by Miller and Liberman (1979) with speech stimuli. For longer steady-state durations, a shift in the abrupt/gradual transition boundary (towards more gradual onset responses) was observed relative to that found with shorter steady-state portions. Taken together, the results of psychoacoustic studies of rate normalization (Diehl & Walsh 1989; Jamieson et al., 1987; Pisoni et al. 1983) call into question Miller and Liberman's proposal that changes in the /ba/-/wa/ category boundary as a function of speaking rate are a result of "speech-specific" perceptual processes which compensate or "normalize" for differences in articulation rate. Instead, the effects may reflect the operation of more general psychophysical mechanisms that operate on both speech and nonspeech signals alike.

One question that must be addressed before a psychophysically-based account of speaking-rate normalization can be accepted, however, is the degree to which such explanations hold for natural stimuli produced within the context of continuous speech. All of the findings discussed previously were obtained with synthetic tokens presented in isolated consonant-vowel contexts. Furthermore, the dependent measure in each case was a shift in the category boundary for either a speech or nonspeech continuum. Recent evidence from studies examining the effects of speaking rate on spoken word recognition (rather than category boundary shifts) with naturally produced stimuli suggest that psychophysical accounts may not completely explain the underlying processes used in perceptual normalization.

A fundamental difference between these more recent investigations and previous studies of normalization concerns the level at which variability within the speech signal is resolved. Psychophysical accounts of normalization (Diehl & Walsh, 1989; Jamieson, 1987) propose that variations in the acoustic realization of phonetic segments, due to factors such as vocal-tract size and speaking rate, are removed or made equivalent during sensory encoding. As a consequence, investigations that have adopted this "sensory" approach, have had little or no concern for how variability might affect spoken word recognition since acoustic differences are assumed to be eliminated prior to phonetic analysis. An alternative view (Sommers, Nygaard and Pisoni, 1992) proposes that normalization is an integral component of phonetic analyses and can have substantial influence on spoken word recognition performance.

In a series of studies, Sommers et al., (1992) demonstrated that identification of monosyllabic words in noise was poorer when the words were presented at several different speaking rates than when the identical items were spoken at only one speaking rate. Both Sommers et al. (1992) and Mullennix, Pisoni, and Martin (1989) have also shown similar results for talker variability; identification performance is significantly poorer when items are produced by multiple talkers than when they are spoken by a single talker. Sommers et al. (1992) suggested that the decreased identification performance for multiple-talker and multiple-rate lists, relative to the corresponding single-rate conditions, was due to greater demands on the normalization system which reduced the processing resources available for phonetic identification. That is, normalization is a resource-demanding process and an integral component of the speech perception system. When demands are increased for one component of the system, the effects cascade to other components including processes used to identify spoken words.

According to this hypothesis *reducing* normalization requirements should improve word recognition performance. Specifically, as the need for continuous re-normalization is decreased, more processing resources can be devoted to word recognition as opposed to compensating for acoustic variability. The results of a recent talker identification study (Nygaard, Sommers & Pisoni, 1992)

provide support for this prediction. Nygaard et al. (1992) trained two groups of subjects to identify the voices of ten talkers (5 male and 5 female) from isolated monosyllabic words. After nine days of training, one group of subjects was asked to identify words produced by the same talkers whose voices they had learned to distinguish. The second group was asked to identify the same items produced by a set of ten unfamiliar voices. Nygaard et al. (1992) argued that learning to identify a talker's voice reduces the perceptual resources necessary to normalize that voice in the future. Therefore, they predicted that word identification would be better for those listeners identifying items produced by familiar voices (i.e. ones they had learned to identify in training). Figure 4 displays the results of this experiment.

Insert Figure 4 about here

At all four signal-to-noise ratios, word recognition scores were significantly higher for subjects who heard items produced by familiar as opposed to unfamiliar voices. These findings suggest that, in addition to determining the phonetic content of speech tokens, listeners extract and retain information such as talker characteristics. Furthermore, the results suggest that this additional information about a talker's voice can be used to facilitate recognition of novel words produced by the same talker.

One way of accounting for the present results on stimulus variability and psychophysical explanations for speaking-rate normalization (Diehl & Walsh, 1989; Pisoni et al., 1983) is to propose more than one process of perceptual normalization. Both psychophysical and phonetic processes may contribute to converting the incoming speech waveform to a representation that can be used for recognition. Psychophysical analyses would be relatively "automatic" in that their operation would depend on the bio-mechanical and physiological properties of the peripheral auditory system and, therefore, would be relatively unaffected by increased stimulus variability. For example, Kewley-Port and Watson (1991) found that thresholds for formant frequency discrimination were not significantly affected by increased variability in the number of vowels presented or the number of consonantal contexts in which they appeared. Sommers et al. (1992) investigated the effects of varying the overall amplitude of natural speech stimuli on word recognition performance. They found that, unlike the case of speaking--rate where variability reduced identification scores, items produced in mixed-amplitude lists were identified equally well as items presented in single-amplitude lists. Compensating for variability due to factors such as overall stimulus amplitude or number of consonantal environments, which do not have direct effects on the acoustic cues signaling phonetic distinctions, may be examples of psychophysically-based normalization procedures which are "automatic" and thus relatively unaffected by variability. Phonetic aspects of normalization, in contrast, would be resource-demanding procedures which are invoked to compensate for variability along dimensions that have direct phonetic consequences. The decrease in word recognition performance associated with mixed-talker (Mullennix et al., 1989) and mixed-rate (Sommers et al., 1992) lists may represent two instances of this latter type of normalization process. One of the important problems confronting researchers working on perceptual constancy in speech perception would be to differentiate between these two classes of normalization mechanisms.

While additional research on the issue of perceptual normalization and speech perception is clearly indicated, the findings discussed in the present section have begun to address essential issues regarding perceptual constancy and spoken word recognition. They suggest, for example, that listeners do not process all sources of variability in an equivalent manner. Some types of variability, such as

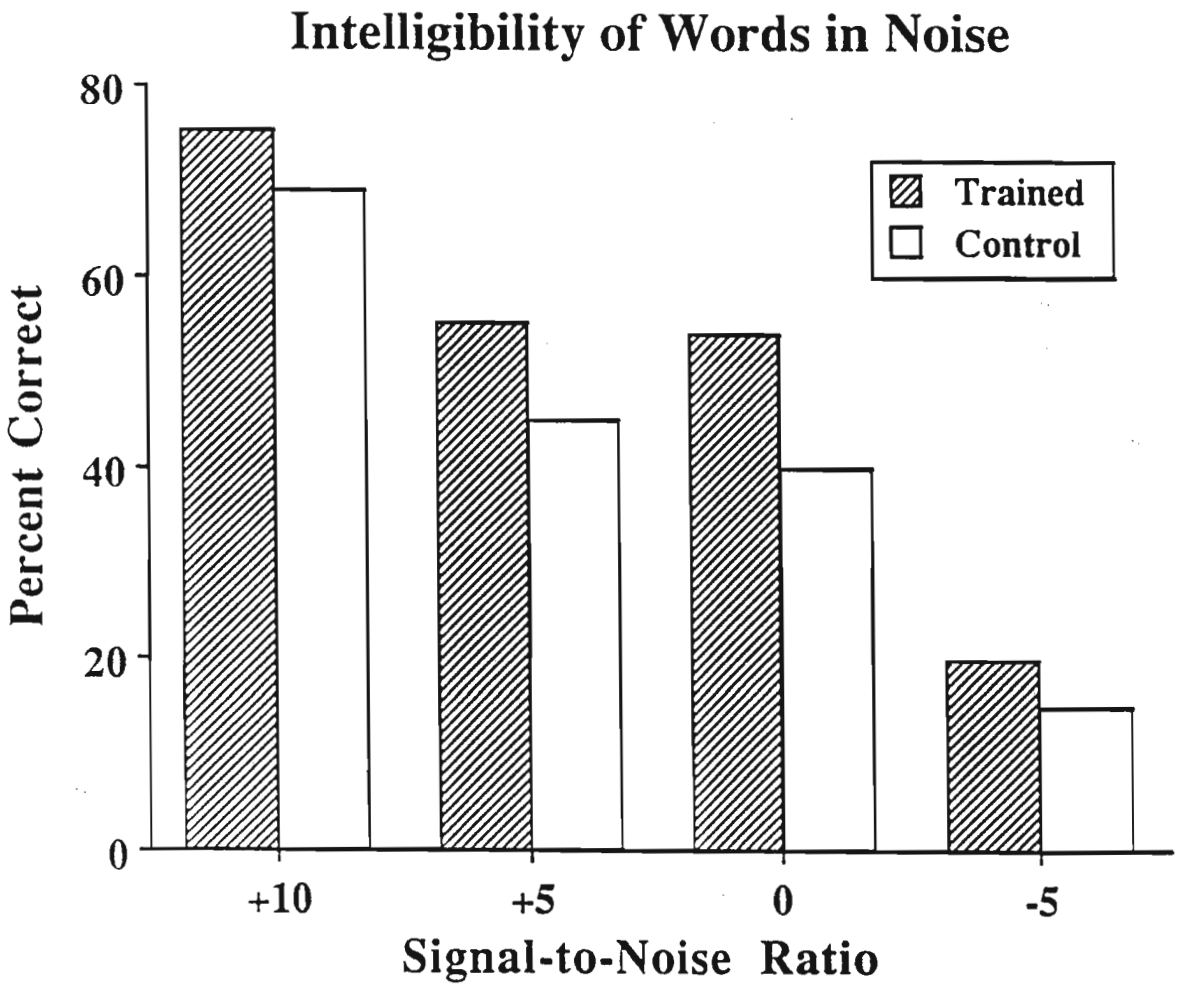


Figure 4: Percent correct identification as a function of signal-to-noise ratio. Hatched bars show identification performance for subjects identifying monosyllabic words produced by familiar voices. Open bars indicate performance for subjects identifying words produced by unfamiliar talkers (From Nygaard, et al., 1993).

changes in overall amplitude, have relatively little impact on spoken word recognition while others such as talker differences can have a profound influence on identification performance. Furthermore, the results argue that, rather than being considered as noise which the system must eliminate, stimulus variability within speech signals may serve an important function in identifying spoken words. Future research, therefore, needs to be directed at determining the specific role of psychoacoustic factors in mediating these two different types of normalization procedures.

PSYCHOPHYSICS AND THE INTERNAL REPRESENTATION OF SPEECH

The preceding account suggests that speech signals may undergo several transformations between acoustic waveform and phonetic percept (Miller and Liberman, 1979; Sommers et al., 1992). One of the central issues in speech perception has been to specify the types of representations that are constructed by the speech processing system. For many years, theorists have argued (Chomsky & Halle, 1968; Liberman et al., 1967; Stevens, 1986) that the speech signal is most appropriately represented as an idealized sequence of discrete linguistic segments or features. Other researchers (Klatt, 1979; 1986) have suggested that the detailed spectral representation that results from peripheral auditory processing may be used directly to determine the phonetic content of speech signals. The principle advantage of a detailed spectral representation, according to Klatt, is that it minimizes information loss due to intermediate recoding from the original signal. Klatt's (1979; 1989) Lexical Access From Spectra (LAFS) model of lexical access is an example of such an auditory representation. The LAFS model proposes that listeners represent spoken words as stored spectral templates similar to those proposed by Stevens & Blumstein (1981). In an extension to the original LAFS model, Klatt (1989) has incorporated both static and dynamic (specifically, smoothed estimates of local spectral changes) information into the auditory representations. Perceptual decisions in the model are based on a best-match criterion between the input representation, consisting of one or more acoustic spectra, and stored spectral templates[7]. Other investigators such as Seneff (1988) and Shamma (1988) have proposed similar psychoacoustic-based models of speech sound representations. Seneff's (1988) model is particularly noteworthy because it includes transformations of the signal that are introduced by well-established properties of basilar-membrane and hair-cell mechanics and takes into account how auditory nerve responses might affect the representation. It is difficult, however, to evaluate the validity of the proposed representations in these models because perceptual experiments with human listeners have not been conducted. Furthermore, the LAFS (Klatt, 1989) and Seneff (1988) representations were developed for implementation on automatic speech recognition devices (ASRDs). Therefore, not only is it necessary to demonstrate that the representations proposed by these models yield adequate perceptual results but the data must parallel human speech performance.

A second psychophysical approach that has been used to investigate the internal representation of speech sounds (Bacon & Brandt, 1982; Houtgast, 1974; Tyler & Lindblom, 1982) has employed both simultaneous- and forward-masking paradigms. Synthetic vowels are used as maskers for sinusoidal signals and thresholds are measured as a function of signal frequency. The resulting pattern of masked thresholds provides an indication of the activity evoked by the vowel (masker) within the auditory system. Findings from a number of studies using simultaneous masking procedures (Bacon & Brandt, 1982; Tyler and Lindblom, 1982) have shown that the formant patterns of vowels are well preserved within these vowel masking patterns (VMPs). Moreover, Moore and Glasberg (1983) found

[7]Klatt (1989) has suggested that the comparison may actually be between the input spectra and the path through a network of stored spectral templates that leads to the best match.

enhanced vowel representations, in which the differences between formant peaks and valleys within the masking pattern were *greater* than in the physical stimulus, when VMPs were determined using a forward-masking procedure. These findings suggest that the internal representation of speech signals may be enhanced by one or more nonlinear processes such as suppression which are not detectable with simultaneous masking.

Perceptual experiments with vowel masking patterns have focused primarily on how changes in the internal representation affect vowel intelligibility. Van Tasell, Fabry, and Thibodeau (1987), for instance, measured VMPs and vowel confusions in normal-hearing and hearing-impaired listeners. Relative to masking patterns for listeners without sensorineural hearing losses, the VMPs of hearing-impaired listeners showed smaller dynamic ranges and poorer preservation of formant structure. Vowel confusions, however, were largely unrelated to the VMPs. Other experiments (Bacon & Brandt, 1982; Sidwell & Summerfield, 1985) have also found that the internal representation of vowel formants is not well preserved in mild to moderately hearing-impaired listeners. Yet, vowel identification in hearing-impaired subjects is generally not significantly reduced compared to normal-hearing listeners.

A number of alternative explanations can be proposed to account for the low correlations between changes in VMPs and vowel intelligibility. First, it may be that vowel masking patterns are subjected to one or more additional transformations and that the final representation used in making phonetic decisions contains sufficient information for accurate vowel discrimination. Alternatively, the representation, as determined by VMPs, may be the basis for vowel identifications but the reduced spectral contrast found in the masking patterns of hearing-impaired listeners may nevertheless be sufficient for accurate vowel discrimination. Sidwell & Summerfield (1985) and Dubno & Dorman (1987), in support of this latter proposal, have shown that vowel discrimination is not significantly impaired until formant bandwidths have been widened to between 4 and 6 times their normal values. These findings suggest that listeners can tolerate substantial reductions in formant structure without exhibiting changes in vowel intelligibility.

Excitation-pattern models of discrimination (Zwicker, 1970; Moore and Glasberg, 1983) provide an additional alternative to VMP representations of speech that might successfully address the poor correlations between vowel identification and changes in VMPs. Excitation patterns plot the output levels of auditory filters as a function of filter center frequency. Zwicker (1970) proposed that two sounds will be discriminable when their representations, based on excitation patterns, differ by a criterion amount, generally 1 dB, in one or more frequency regions. Excitation-pattern models have provided good accounts of pure-tone intensity and frequency discrimination (Moore, 1989) but have only recently been applied to speech and speech-like stimuli. For example, Sommers and Kewley-Port (1993) have used differences in excitation patterns to successfully model formant frequency discrimination in vowels. Kewley-Port (1991) has also found that detection thresholds for vowels are well predicted on the basis of excitation-pattern representations. Indirect behavioral evidence for an auditory representation of speech sounds based on excitation-pattern models also comes from studies of vowel similarity judgements (Carlson & Granstrom, 1979; Plomp, 1970). These studies demonstrated that similarity judgements for vowels were based on spectral differences consistent with an excitation-pattern model of auditory representations.

One of the major problems of modeling speech with spectral templates, masking patterns, or excitation-pattern models is that temporal aspects of the signal are not well captured by such representations. Thus, it remains unclear whether consonants, which are distinguished primarily on

the basis of temporal or dynamic properties (Kewley-Port, 1983; Kewley-Port et al., 1983; Klatt, 1989) will be adequately preserved in such representations. A number of additional models of auditory processing have been proposed (see Patterson and Cutler, 1989, for a review) that produce a spectro-temporal representation of speech sounds. Such representations preserve the dynamic relationships in both vowels and consonants and may therefore provide a better representation of speech. To date, these models have been applied exclusively as front-end processors in automatic speech recognition devices. It will therefore be important to determine how well representations based on these models predict actual performance by human listeners in speech perception experiments.

PSYCHOPHYSICS AND THE UNITS OF SPEECH PERCEPTION

The final issue to be discussed concerns the basic units of perceptual analysis for speech processing. Several investigators (Luce & Pisoni, 1987; Pisoni, 1985) have suggested that the size of the perceptual unit will vary as attention shifts from one linguistic level of the message to another. However, suggestions that listeners rely on different units of analysis for perceiving aspects of the message as diverse as prosodic and phonetic information, for example, only serves to reinforce the multi-dimensional nature of speech perception. The present discussion will focus on evidence for the *minimal* units of perception that are used in the initial recoding from acoustic waveform to phonetic percept. Determining the minimal unit of analysis for phonetic labeling has important implications for at least one of the areas discussed previously, acoustic-phonetic invariance. Theoretical accounts suggesting that invariant acoustic cues are the principle basis for phonetic distinctions might, at least initially, restrict their search for such invariants to the minimal perceptual units. Both Stevens & Blumstein (Stevens & Blumstein, 1981) and Kewley-Port (Kewley-Port, 1983), for instance, have restricted their search for invariant static and dynamic acoustic cues to syllable-sized CV segments.

Unfortunately, there is little direct psychophysical evidence regarding the basic unit of perceptual analysis for the initial stages of speech processing. Most of the relevant data, therefore, comes indirectly from considering the results of temporal resolution experiments with nonspeech stimuli and the limitations these findings place on possible perceptual units for speech. For example, Liberman et al. (1967) have claimed that listeners are capable of accurately identifying speech at rates of approximately 25-30 phonemes/sec. Thomas, Hill, Carroll, & Bienvenido (1970), however, found that to accurately identify sequences of speech sounds, the components must have a minimum duration of 125-250 ms. Therefore, the temporal resolving power of the auditory system could not support the high rates of identification that are typical for fluent speech if the phoneme were the basic unit of perceptual analysis. If, however, larger segments of speech, such as syllables or words, are considered as basic units of perception, then the temporal resolution of the auditory system would be sufficient to process speech at the rates determined by Liberman et al. (1967). Huggins (1964), for instance, found that the modal syllable duration in connected speech was approximately 150 ms which is consistent with the time required for accurate identification of sequences of speech sounds as measured by Thomas et al. (1970). Massaro (1974) has also provided evidence in support of the syllable as the basic unit of perception during initial stages of acoustic-phonetic processing. Using a recognition-masking paradigm, he demonstrated that the auditory system integrates phonetic information over approximately a 200-ms time window which is roughly equivalent to the average syllable duration measured by Huggins (1964). Several other psychoacoustic studies (Bertoncini & Mehler, 1981; Zwicker, Terhardt, & Paulus, 1979), employing different methodologies, including discrimination paradigms and auditory modeling, have found that the syllable is the most likely candidate for the earliest unit of perceptual analysis.

At least two arguments can be raised concerning the claims that the syllable is the minimal unit of acoustic-phonetic analysis. The first is that all such accounts assume serial processing of connected speech. If, however, the speech waveform is processed in a parallel fashion, then the limits of temporal resolution discussed above can no longer be used as an objection to the phoneme, or some smaller component of speech such as features, serving as the basic unit of perceptual analysis. The second criticism is that these proposals do not take into account language-specific knowledge and its potential contribution to analyzing fluent speech. The results of phoneme-monitoring experiments provide empirical support for both of these objections. For example, Savin & Bever (1970) found that listeners have longer response latencies when asked to detect words beginning with syllable-initial "b" than if they are asked to detect words beginning with the sequence of phonemes "bab". These results argue against a strictly serial (left-to-right) processing of speech. Rubin, Turvey, and Van Gelder (1976) found that phoneme-monitoring is faster for real words than for nonsense words. This finding suggests that listeners rely on language-specific information to increase the speed of linguistic processing. Unless analogs to these results can be demonstrated with nonspeech stimuli, it is likely that the limits of temporal resolution may be different for speech and nonspeech and, therefore, conclusions regarding the earliest unit of perceptual analysis need to be determined using speech stimuli. Experiments with nonspeech signals may be useful controls but they probably will not help much to resolve the fundamental issues surrounding perceptual units in speech which are intimately related to the communicative function of language.

DISCUSSION

This review has considered a number of theoretical proposals and empirical findings from auditory psychophysics that are relevant to several long-standing issues in the field of speech perception. In some instances, psychophysical investigations have made significant contributions to our understanding of these central issues in speech perception. In other cases, psychoacoustic findings provide, at best, only indirect data for addressing the relevant questions. This diverse set of results is not unexpected given that almost all theories of speech perception include both an auditory and phonetic stage of processing. Even models such as Klatt's (1979; 1989) LAFS model, in which auditory spectra serve as the principle basis for phonetic decisions, incorporate several aspects of phonetic, or top-down processing. It is therefore improbable that any of the current theoretical issues in speech perception will be adequately addressed by relying exclusively on either psychophysical or phonetic accounts. What is instructive, however, is to evaluate the contributions that each approach has made for addressing leading problems in the field of speech perception and spoken language processing and to examine how experimental methodologies and designs might be altered to increase their relevancy in the future.

For at least one of the issues discussed in the present review, the problem of acoustic-phonetic invariance, psychoacoustic findings have contributed significantly to answering a number of questions regarding context-independent cues to phonetic distinctions. By incorporating both acoustic analyses and perceptual experiments, this line of research has shown that there may be a set of invariant acoustic properties contained within the gross spectral shape of consonants and that the auditory system is sufficiently sensitive to use the proposed cues as a basis for classifying and discriminating speech sounds. With additional research, employing longer-duration natural speech tokens and improved signal analysis techniques, investigators can continue to expand the inventory of candidate acoustic properties that can serve as invariant cues to these consonantal distinctions in speech.

In contrast, one of the issues that has not been adequately addressed with psychoacoustic investigations is the basic unit of analysis. Part of this failure is due to the use of an inappropriate theoretical framework that has only recently begun to change. Specifically, early attempts at relating auditory psychophysics and speech perception used methodologies and stimuli (mostly pure-tones) that had been successfully employed to investigate basic auditory capabilities. These signals proved largely unsuccessful for studying speech perception because, in many instances, simple and complex stimuli are processed in a qualitatively different manner within the auditory system (Pisoni, 1978; Stevens and House, 1972). More recent studies (Carrell, in press; Green, 1988; Henn & Turner, 1990; Turner & Van Tasell, 1984) have begun to explore auditory processing using speech and speech-like stimuli. The recent work on profile analysis by Green (Green, 1988), for example, has altered long-held assumptions about the integration of information across critical bands in stimuli with speech-like spectral characteristics. Carrell (in press) has shown that amplitude co-modulation plays an important role in grouping stimulus components from a sentence into a single auditory object. Studies such as these are clearly necessary if auditory psychophysics is to successfully address important theoretical issues in speech perception that deal with basic units of perception, perceptual normalization and the lack of acoustic-phonetic invariance.

An additional reason that psychophysical approaches have not been more successful in addressing issues in speech perception has been a general failure, again until recently, to integrate basic and clinical research findings. While there is a large body of data available concerning changes in auditory capabilities as a consequence of hearing impairment, only a few investigations have systematically evaluated how reductions in individual psychoacoustic abilities affect specific aspects of speech perception (see, for example, Glasberg & Moore, 1989; Van Rooji and Plomp; 1990). Moreover, those studies that have examined the relationship between auditory capabilities and speech processing have generally measured correlations between psychophysical abilities, as determined with simple stimuli, and overall speech perception scores. While these studies provide important information about general auditory abilities that might be important for speech perception, they have a number of shortcomings. First, as noted earlier, auditory processing can be significantly different for simple and complex stimuli. Secondly, the studies have not been designed to examine which specific aspects of speech processing are affected by changes in auditory capabilities. For example, several investigations have failed to find correlations between reduced spectral resolution and deficits in speech perception (Turner & Van Tasell, 1984; Van Tasell et al., 1984). This raises the obvious question of which specific aspects of spoken language processing are affected by reduced frequency selectivity in these listeners. Integrating basic and clinical research findings would provide one methodology that could be used to address questions such as these in the future.

In summary, recent psychophysical investigations have begun to employ stimuli, methodologies, and theoretical approaches that are appropriate for addressing several long-standing issues in the field of speech perception and spoken language processing. It should be noted that, despite over forty years of research, there is still no unified theory of speech perception. Perhaps the absence of such a theory can, in part, be attributed to reluctance on the part of speech researchers to integrate empirical findings from several theoretical perspectives. A notable exception to this is a recent review by Patterson and Cutler (1989) who have proposed an outline for incorporating both psychophysical and psycholinguistic considerations in theories of speech processing.

Patterson and Cutler (1989) proposed that the major obstacle to integrating psychoacoustic and psycholinguistic findings into a unified theory of speech perception is that the two approaches employ distinct and mutually exclusive representations of speech sounds. Therefore, it has not been possible

to use the output of peripheral auditory analyses, as determined by psychoacoustics, as the input to models of phonetic processing, as suggested by psycholinguistics. One recent approach that provides a theoretical framework for integrating the two representations, according to these investigators, is connectionism and neural nets (Elman and Zipser, 1987; Landauer, Kamm and Singhal, 1987; Peeling and Bridle, 1986). Although a number of specific connectionist models have been proposed and implemented for speech recognition, in general, all such networks consist of three layers of units: input units which describe a range of potential acoustic properties, output units which describe a range of phonetic items, and hidden units connecting the input and output layers (Patterson and Cutler, 1989). Training the networks consists of determining a set of weights for the hidden units that provide best-fit matches between a set of auditory inputs and phonetic outputs. The appeal of the connectionist approach is that it provides a framework for incorporating both psychoacoustic and psycholinguistic representations of speech and details how one form is converted to the other. While the current generation of connectionist models are clearly inadequate and implausible representations of human spoken language processing (see Patterson and Cutler, 1989 for a review), such integrative approaches are more likely to succeed in addressing the four major theoretical issues in speech perception discussed in this review.

Speech perception and spoken language processing are extremely complex processes that involve many diverse sources of knowledge. As theoretical advances are made in the field using new techniques and methodologies, we can expect further developments and insights into the relations between auditory system function and speech coding by human listeners.

References

Bacon, S.P. and Brandt, J.F. (1982). Auditory processing of vowels by normal-hearing and hearing-impaired listeners. *Journal of Speech and Hearing Research*, **25**, 339-347.

Bailey, P.J. (1983). Hearing for speech: The information transmitted in normal and impaired-hearing. In M.E. Lutman and M.P. Haggard (Eds.), *Hearing Science and Hearing Disorders.* New York: Academic Press.

Bertoncini, J. and Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development* **4**, 247-260.

Bladon, R.A.W., Henton, C.G., and Pickering, J.B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication,* **4**, 59-69.

Blumstein, S.E. and Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America,* **66**, 1001-1017.

Blumstein, S.E. and Stevens, K.N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67,* 648-662.

Brady, S.A. and Darwin, C.J. (1978). Range effects in the perception of voicing. *Journal of the Acoustical Society of America,* **63**, 1556-1558.

Carlson, R. and Granstrom, B. (1979). Model predictions of vowel dissimilarity. *Speech Transmission Laboratory-Quarterly Progress Status Report 2-3,* 19-35.

Carrell, T. (in press). Acoustical cues to auditory object formation. In J. Charles-Luce, P.A. Luce and J.R. Sawusch (Eds.), *Theories in Spoken Language: Perception, Production and Development.* Norwood, NJ: Ablex Press

Chomsky, N. and Halle, M. (1968). *The Sound Patterns of English.* New York: Harper and Row.

Cooper, F.S. (1950). Spectrum analysis. *Journal of the Acoustical Society of America,* **22**, 761-762.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.N., and Gertsman, L.J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America,* **24**, 597-606.

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

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

Diehl, R.L. and Walsh, M.A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America,* **85**, 2154-2164.

Diehl, R.L., Kluender, K.R., Walsh, M.A. (1990). Some auditory bases of speech perception and production. In W. A. Ainsworth (Ed.), *Advances in Speech, Hearing, and Language*. New York: JAI Press.

Diehl, R.L., Souther, A.F. and Convis, C.I. (1980). Conditions on rate normalization in speech perception. *Perception and Psychophysics,* **27**, 435-443.

Dorman, M.F. and Raphael, L.J. (1980). Distribution of acoustic cues for stop consonant place of articulation in VCV syllables. *Journal of the Acoustical Society of America,* **67**, 1333-1335.

Dubno, J.R. and Dirks, D.D. (1989). Auditory filter characteristics and consonant recognition for hearing-impaired listeners. *Journal of the Acoustical Society of America,* **85**, 1666-1675.

Dubno, J.R. and Dorman, M.F. (1987). Effects of spectral flattening on vowel identification. *Journal of the Acoustical Society of America,* **82**, 1503-1511.

Elman, J.L. and Zipser, D. (1987). Learning the hidden structure of speech. *Institute for Cognitive Science*, UCSD, California Report 8701.

Fant (1960). *Acoustic Theory of Speech Production.* The Hague: Mouton.

Fletcher, H. (1940). Auditory Patterns. *Review of Modern Physics,* **12**, 47-65.

Glasberg, B.R. and Moore, B.C.J. (1989). Psychoacoustic abilities of subjects with unilateral and bilateral cochlear impairments and their relationship to their ability to understand speech. *Scandinavian Audiology*, **Supp 1 32**, 1-25.

Green, D.M. (1988). *Profile Analysis: Auditory Intensity Discrimination*. New York: Oxford University Press.

Green, D.M. and Kidd, G., Jr. (1983). Further studies of auditory profile analysis. *Journal of the Acoustical Society of America,* **73**, 1260-1265.

Green, D.M., Kidd, G., Jr. and Picardi, M.C. (1983). Successive versus simultaneous comparison in auditory intensity discrimination. *Journal of the Acoustical Society of America,* **73**, 639-643.

Henn, C.C. and Turner, C.W. (1990). Pure-tone increment detection in harmonic and inharmonic backgrounds. *Journal of the Acoustical Society of America*, **88**, 126-131.

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

Hirsh, I.J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America,* **31**, 759-767.

Houtgast, T. (1974). Auditory analysis of vowel-like sounds. *Acustica,* **31**, 320-324.

Howell, P. and Rosen, S. (1984). Natural auditory sensitivities as universal determiners of phonetic contrasts. In B. Butterworth, B. Comrie, and O. Dahl, *Explanations for Language Universals*. New York: Zoya.

Huggins, A.W.F. (1964). Distortion of the temporal pattern of speech: Interruption and alternation. *Journal of the Acoustical Society of America,* **36**, 1055-1064.

Humes, L.E. and Christopherson, L. (1991). Speech identification difficulties of hearing-impaired elderly persons: the contribution of auditory processing deficits. *Journal of Speech and Hearing Research,* **34**, 686-693.

Jamieson, D.G. (1986). Studies of possible psychoacoustic factors underlying speech perception. In Schouten, M.E.S. (ed.), *The Psychophysics of Speech Perception*. The Netherlands: Nijhoff, Dordecht.

Jamieson, D.G. and Slawinska, E.B. (1983). The discriminability of transition duration: Effects of the amplitude and duration of following steady-state. *Journal of the Acoustical Society of America,* Suppl. **1 76**, S29.

Jamieson, D.G., Johnson, T. and Rvachew, S. (1986). A role for intra-speech masking in "rate-normalization" on a stop-semi-vowel continuum. *Alberta Conference on Language*, Banff.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America,* **72**, 379-389.

Kewley-Port, D. (1991). Detection thresholds for isolated vowels. *Journal of the Acoustical Society of America,* **89**, 820-829.

Kewley-Port, D. and Luce, P.A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception and Psychophysics,* **35**, 353-359.

Kewley-Port, D. and Watson, C.S. (1991). Thresholds for formant-frequency discrimination of vowels in consonantal context. *Journal of the Acoustical Society of America,* **Suppl 1**, S79.

Kewley-Port, D., Pisoni, D.B., and Studdert-Kennedy, M. (1983). Perception of static and dynamic cues to place of articulation in syllable-initial stop consonants. *Journal of the Acoustical Society of America,* **73**, 1779-1793.

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

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. (1986). Models of phonetic recognition I. Issues that arise in attempting to specify a feature-based strategy for speech recognition. In P. Mermelstein (Ed.), *Proceedings of the Montreal Satellite Symposium on Speech Recognition,* Twelfth International Congress on Acoustics.

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

Kurowski, K. and Blumstein, S.E. (1987). Acoustic properties for place of articulation in nasal consonants. *Journal of the Acoustical Society of America,* **81**, 1917-1927.

Landauer, T.K., Kamm, C.A. and Singhal, S. (1987). Teaching a minimally structured back-propagation network to recognise speech sounds. *Bell Communications Research Report.*

Lautner, J.L. and Hirsh, I.J. (1985). Speech as temporal pattern: A psychoacoustic profile. *Speech Communication,* **4**, 41-54.

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

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

Liberman, A.M., Delattre, P.C., Cooper, F.S. and Gertsman, L.J. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology,* **52**, 127-137.

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gertsman, L.J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs,* **68**, Number 379.

Lieberman, P. (1979). Hominid evolution, supralaryngeal vocal tract physiology and the fossil evidence for reconstructions. *Brain and Language,* **7**, 101-126.

Liljencrants, J. and Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language,* **48**, 839-862.

Lisker, L. and Abramson, A. (1971). Distinctive features and laryngeal control. *Language,* **47**, 767-785.

Lisker, L. and Abramson, A.S. (1970). The voicing dimension: Some experiments in comparative phonetics. In, *Proceedings of the 6th International Congress of Phonetic Sciences*. Prague: Academia.

Luce, P.A., and Pisoni, D.B. (1987). Speech perception: New directions in research, theory and application. In H. Wintz (Ed.), *Human Communication and its Disorders: A Review.* Vol. 1. Norwood, NJ: Ablex.

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology,* **2**, 199-208.

Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J. and Dooling, R.J. (1976). Discrimination 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. (1981). Effects of speaking rate on segmental distinctions. In P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the Study of Speech.* Hillsdale, NJ: Erlbaum.

Miller, J.L. and Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America,* **73**, 1751-1755.

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

Miller, J.L., Grosjean, F., and Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica,* **41**, 215-225.

Moore, B.C.J. (1989). *An Introduction to the Psychology of Hearing.* New York: Academic Press.

Moore, B.C.J. and Glasberg, B.R. (1983). Masking patterns for synthetic vowels in simultaneous and forward masking. *Journal of the Acoustical Society of America,* **73**, 906-917.

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

Nygaard, L.C., Sommers, M.S. and Pisoni, D.B. (1992). Speech perception as a talker-contingent process. *Psychological Science.* Submitted.

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

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

Pastore, R.E. (1987). Possible acoustic bases for the perception of voicing contrasts. In, M.E.S. Schouten (Ed.) *The Psychophysics of Speech Perception.* The Netherlands: Nijhoff, Dordecht.

Patterson, R. and Cutler, A. (1989). Auditory preprocessing and speech recognition. In A..D. Baddeley and N.O. Bernsen (Eds.), *Research Directions in Cognitive Science: A European Perspective.* Vol. I. London: Erlbaum.

Peeling, S. and Bridle, J. (1986). Experiments with a learning network for a simple phonetic task. *Proceedings of the Institute for Acoustics: Speech and Hearing.* Vol. 8, Part 7, 315-322.

Peters, R.W. and Moore, B.C.J. (1992). Auditory-filter shapes at low center frequencies in young and elderly hearing-impaired subjects. *Journal of the Acoustical Society of America,* **91**, 256-266.

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

Pisoni, D.B. (1977). Identification and discrimination of 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. (1978). Speech perception. In W.K. Estes (Ed.), *Handbook of Learning and Cognitive Processes*. Vol. 6. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

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

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In, R. Plomp and G. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing*. Leiden: Sijthoff.

Port, R.F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics,* **32**, 142-152.

Rubin, P., Turvey, M.T., and Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception and Psychophysics,* **19**, 394-398.

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

Schouten, M.E.H. (Ed.) (1992). *The Auditory Processing of Speech: From Sounds to Words*. Berlin: Mouton de Gruyter.

Schouten, M.E.S. (Ed.) (1987). *The Psychophysics of Speech Perception*. The Netherlands: Nijhoff, Dordecht.

Searle, C.L., Jacobson, J.Z., and Rayment, G. (1979). Stop consonant discrimination based on human audition. *Journal of the Acoustical Society of America,* **65**, 799-809.

Seneff, S. (1988). A joint synchrony/mean-rate response model of auditory speech processing. *Journal of Phonetics,* **16**, 55-76.

Shamma, S. (1988). The acoustic features of speech sounds in a model of auditory processing: Vowels and voiceless fricatives. *Journal of Phonetics,* **16**, 77-92.

Shepard, R.N. (1972). Psychological representation of speech sounds. In, E.E. David and P.B. Denes (Eds.), *Human Communication a Unified View*. New York: McGraw-Hill.

Sidwell, A. and Summerfield, Q. (1985). The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise. *Journal of the Acoustical Society of America,* **78,** 495-506.

Sommers, M.S. and Humes, L.E. (in press). Auditory-filter shapes in normal-hearing, noise-masked normal, and elderly listeners. *Journal of the Acoustical Society of America.*

Sommers, M.S. and Kewley-Port, D. (1993). Modeling formant frequency discrimination. *Journal of the Acoustical Society of America,* Submitted.

Sommers, M.S., Nygaard, L.C. and Pisoni, D.B. (1992). Stimulus variability and the perception of spoken words: The effects of variations in speaking rate and overall amplitude. *Proceedings of the Second International Conferences on Spoken Language Processing.* Banff, Canada, 217-221.

Stevens, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In P.B. Denes and E.E. Davids (Eds.), *Human Communication: A Unified View.* New York: McGraw-Hill.

Stevens, K.N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics and psychoacoustics. In, T. Myers, J. Laver, J. Anderson (Eds.), *The Cognitive Representation of Speech.* The Netherlands: North-Holland Publishing.

Stevens, K.N. (1986). Models of phonetic recognition II: A feature-based model of speech recognition. In P. Mermelstein (Ed.), *Proceedings of the Montreal Satellite Symposium on Speech Recognition,* Twelfth International Congress on Acoustics.

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

Stevens, K.N. and House, A.S. (1972). Speech perception. In J. Tobias (Ed.), *Foundations of Modern Auditory Theory.* Vol. 2. New York: Academic Press.

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

Summerfield, Q. and Bailey, P. (1977). On the dissociation of spectral and temporal cues for stop consonant manner. Journal of the Acoustical Society of America, **61,** Suppl 1. S46.

Summerfield, Q. and Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America,* **62,** 435-448.

Thomas, I.B., Hill, P.B., Carroll, F.S. and Bienvenido, G. (1970). Temporal order in the perception of vowels. *Journal of the Acoustical Society of America,* **48,** 1010-1013.

Turner, C.W. and Van Tasell, D.J. (1984). Sensorineural hearing loss and the discrimination of vowel-like stimuli. *Journal of the Acoustical Society of America,* **75**, 562-565.

Tyler, R.S. and Lindblom, B. (1982). Preliminary study of simultaneous-masking and pulsation-threshold patterns of vowels. *Journal of the Acoustical Society of America,* **71**, 220-224.

Tyler, R.S. Summerfield, Q., Wood, E.J. and Fernandes, M. A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America,* **72**, 740-752.

Van Rooji, J.C.G.M. and Plomp, R. (1990). Auditive and Cognitive factors in speech perception by elderly listeners. II: Multivariate analyses. *Journal of the Acoustical Society of America,* **88**, 2611-2624.

Van Tasell, D.J., Fabry, D.A., and Thibodeau, L.M. (1987). Vowel identification and vowel masking patterns of hearing-impaired subjects. *Journal of the Acoustical Society of America,* **81**, 1586-1597.

Van Tasell, D.J., Hagen, L.T., Koblas, L.L., and Penner, S.G. (1982). Perception of short-term spectral cues for stop consonant place by normal and hearing-impaired subjects. *Journal of the Acoustical Society of America,* **6**, 1771-1780,

Walley, A.C. and Carrell, T.D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America,* **73**, 1011-1022.

Yost, W.A. and Watson, C.S. (Ed.) (1988). *Auditory Processing of Complex Sounds*. Hillsdale, N.J: L. Erlbaum.

Zwicker, E. (1970). Masking and psychological excitation as consequences of the ear's frequency analysis. In, R. Plomp and G. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing*. Leiden: Sijthoff.

Zwicker, E., Terhardt, E., and Paulus, E. (1979). Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America,* **65**, 487-498.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Speech Perception: New Directions in Research and Theory[1]

Lynne C. Nygaard and David B. Pisoni

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

[1]This work was supported by NIH Research Grant DC-000111-16 and NIH Training Grant DC-00012-14 to Indiana University, Bloomington, IN. We thank Scott Lively for comments and criticisms on an earlier draft of the chapter. This chapter is to appear in J.L. Miller & P.D. Eimas (Eds.), *Handbook of perception and cognition. Vol. 11: Speech, language and communication.* New York: Academic Press.

Abstract

This chapter provides a selective review of some of the fundamental problems encountered in the study of speech perception. We address several of the key issues in the field and suggest some promising alternative approaches to problems that have traditionally confronted speech researchers. Based on the accumulated knowledge in the field, we argue for a rethinking of the nature of linguistic units--how they are represented and how they are processed by the nervous system. We suggest a reinterpretation of the canonical abstract linguistic unit as the type of representation that underlies the perception of speech. Rather, the richness and detail of the neural representations used in speech perception as well as the flexible, context-dependent nature of speech perceptual analysis are emphasized.

Speech Perception: New Directions in Research and Theory

Introduction

The fundamental problem in speech perception is how to characterize the process by which listeners derive meaning from the acoustic waveform. At first glance, the solution to the problem of how we perceive speech seems deceptively simple. If one could identify stretches of the acoustic waveform that corresponded to units of perception, then the path from sound to meaning would be clear. Unfortunately, this correspondence or mapping has proven extremely difficult to find, even after some forty-five years of research on the problem.

One of the earliest and most basic findings in the study of speech perception was the observation that speech sounds were not organized in the signal like "beads on a string" (Liberman, 1957; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) to be picked out of the acoustic signal one at a time by the perceiver for further processing. Rather, just like capabilities in other domains of perceptual research, the ability to perceive speech is a highly complex process that involves converting a physical stimulus into some kind of abstract neural representation. The ultimate goal of research on speech perception is to describe and explain the structure and function of the perceptual system at each level of description--from behavioral aspects to neural instantiation.

For the most part, the issues encountered in the study of speech perception are the same as those confronting researchers in perception and cognition in general. For example, determining what information in the physical signal reaching the sensory organs specifies objects in the world underlies investigation of any field of perception. However, the study of speech has some unique problems. If one conceptualizes perception in terms of what is perceived (or what is the object of perception) and also what constitutes the perceptual medium, then it becomes immediately apparent that speech perception may be a special case of the more general problem.

Consider first a classic example in visual perception. When we perceive a chair, the object of perception is clear. Reflected light serves as the proximal stimulus, or medium of perception, and provides information about the object that is perceived--the chair. However, what is the analogous situation in speech? Do listeners perceive acoustic units (Diehl & Kluender, 1989), phonetic segments, articulatory gestures (Fowler, 1986), articulatory intentions (Liberman & Mattingly, 1989), or simply meaning (see Remez, 1986)? What exactly does the acoustic speech signal provide information about? These questions have generated a considerable degree of controversy in the field and stem in a large part from the symbolic nature of speech and language. Because the relationship between sound and meaning is arbitrary, the object of perception in speech, however it is defined, may have only an indirect relationship to the information encoded in the acoustic medium.

This relationship becomes even more complex if one considers the time-varying nature of the acoustic signal. Not only is there an arbitrary relation between sound and meaning, but the continuous time-varying acoustic signal provides simultaneous information about symbolic, discrete perceptual units such as phonemes, words, and phrases. Given these inherent differences between the acoustic medium and the results of perceptual analysis, how might speech be processed and represented by the nervous system?

Typically, over the last twenty years, the study of speech has been pursued from an information processing approach (Studdert-Kennedy, 1976). The acoustic waveform was assumed to be analyzed into a set of discrete, symbolic units which are then compared with stored representations

in memory. From this point of view, the end product of perception in speech is a string of abstract, timeless, idealized, canonical linguistic units such as distinctive features, phonetic segments, or phonemes in the traditional sense. Although generating a rich and informative body of empirical research and phenomena, this approach to speech perception has left many theoretical issues unexplored and many questions unanswered.

The goal of the present chapter is to provide a selective review of some of the fundamental problems encountered in the study of speech perception. Although a complete and exhaustive discussion of the empirical work is beyond the scope of this chapter (for other reviews, see Cutting & Pisoni, 1978; Darwin, 1976; Goldinger, Pisoni, & Luce, in press; Jusczyk, 1986; Luce & Pisoni, 1987; Miller, 1990; Pisoni, 1978; Pisoni & Luce, 1986; Studdert-Kennedy, 1974, 1976), we hope to address some key issues in the field and suggest some promising alternative approaches to the old problems that have confronted speech researchers. Our aim will be to demonstrate how the accumulated knowledge in the field leads to a rethinking of the nature of linguistic units--how they are represented and how they are processed by the nervous system. Our bias will be toward a reinterpretation of the canonical abstract linguistic unit as the type of representation that underlies the perception of speech. Rather, we hope to emphasize the richness and detail of the neural representations used in speech perception as well as the flexible, context-dependent nature of the processes that act on and use the informationally-rich speech representations. In doing so, our discussion will focus primarily on phonetic perception. We maintain a somewhat arbitrary distinction between phonetic perception and spoken word recognition only for convenience and brevity. However, we acknowledge that any complete account of speech perception will necessarily have to take into consideration all levels of perceptual processing (see Cutler, this volume) and will also have to ultimately confront the complex issues in spoken language comprehension.

Basic Issues in Speech Perception

Linearity, Lack of Invariance, and Segmentation

One of the most prominent characteristics of speech is that it is a time-varying continuous signal; yet, the impression of the listener is that of a series of discrete linguistic units--the phonemes, syllables and words that carry meaning. The problem for the theorist working in speech perception is to account for how listeners might carve up the continuous signal of speech into the appropriate units of analysis and recover the linguistic intent of the talker.

Although perceptually, linguistic units such as phonemes, syllables, and words appear to follow one after another in time, the same linearity is not found in the physical signal (Chomsky & Miller, 1963). Thus, it is often stated that the speech signal fails to meet the *linearity* condition which assumes that for each phoneme, there must be a particular stretch of sound in the utterance. For example, if phoneme X precedes phoneme Y in perception, then the stretch of sound associated with phoneme X should, according to this principle, precede the stretch of sound associated with phoneme Y in the physical signal. Unfortunately, this condition does not hold for natural speech. Properties of the acoustic signal associated with one phoneme often overlap or co-occur with the properties of adjacent segments (Delattre, Liberman, & Cooper, 1955; Liberman, Delattre, Cooper, & Gerstman, 1954). In addition, a single acoustic property may contribute to the perception of several linguistic units and conversely, several acoustic properties or segments may contribute to the perception of just a single linguistic unit. Although the overlapping of linguistic information in the acoustic signal complicates theoretical explanations of speech perception, it apparently allows for the parallel

transmission of linguistic information and consequently, for the very high rates of information transfer characteristic of speech communication (Liberman, et al., 1967).

In addition to the lack of linearity in the mapping of acoustic segments onto phonetic percepts, there also exists a lack of acoustic-phonetic invariance in speech perception. Researchers have been unable to find invariant sets of acoustic features or properties that correspond uniquely to individual linguistic units. Instead, because adjacent phonetic segments exert a considerable influence on the acoustic realization of a given phoneme, the acoustic properties that specify a particular phoneme can vary drastically as a function of phonetic context, speaking rate, and syntactic environment. Figure 1 illustrates a classic example of the role of context variability in speech perception (see Liberman, 1957). The first two formants (energy maxima in the acoustic signal resulting from natural resonances of the vocal tract) are shown for a variety of syllables containing the phoneme /d/. The formant transitions that begin each syllable provide information for the /d/ in each syllable context. However, depending on vowel environment, the formant transitions are very different acoustically and consequently, the physical realization of the phoneme /d/ is quite different. The paradox for the study of speech perception is that listeners perceive each of these syllables as beginning with a /d/ sound even though the formant transitions for place of articulation in these patterns are anything but invariant acoustically (see Liberman et al., 1967).

Insert Figure 1 about here

The complex relations between acoustic properties of the speech signal and the linguistic units that listeners perceive stems from the way in which speech is produced. The sounds of speech are largely coarticulated--that is, the articulatory gestures associated with adjacent phonetic segments overlap in time. For example, consider the consequences of lip rounding on the initial consonant cluster in the word *stoop* versus the word *stop*. When a speaker produces the word *stoop*, lip-rounding for the vowel occurs as articulation of the initial consonant cluster begins. In producing the word *stop* on the other hand, lip rounding is absent and the articulatory realization of the initial consonant cluster is consequently different acoustically. The process of coarticulation of phonetic segments results in the smearing together of information in the acoustic signal about individual phonemes. So, as gestures overlap in time, so do the corresponding acoustic properties. Hockett's (1955) famous Easter egg analogy provides an apt illustration of this phenomenon. Hockett described the production of speech as a series of Easter eggs that are sent down a conveyer belt to be pushed through a wringer. As each colored egg is smashed by the wringer, the bits of colored shell and yolk are mixed with other eggs to create a smear of the original series of eggs. The consequence of intended phonetic segments passing through the "articulatory wringer" is that the acoustic properties of the speech signal do not relate in any simple, obvious, or invariant way to the sequence of phonetic segments resulting from the listener's perceptual analysis of the signal.

The way in which speech sounds are coarticulated in production leads to yet another problem in the study of speech. Because speech sounds overlap in the acoustic signal and no one-to-one mapping exists between acoustics and phonetics, it is unclear how listeners are able to separate information about one linguistic unit from another. This difficulty in the parsing of the acoustic waveform into discrete linguistic units has been termed the *segmentation problem*. Figure 2 shows a speech spectrogram of an utterance that illustrates this problem. Although a native speaker of English has no problem identifying the phonemes, syllables, and words in this phrase, "I owe you a yo-yo,"

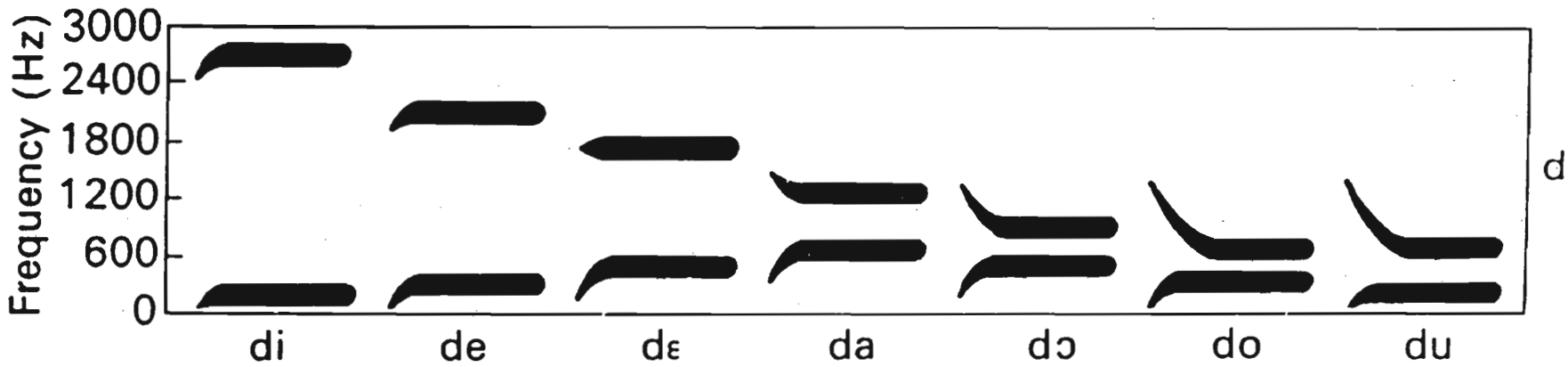


Figure 1. Synthetic tokens of the syllables beginning with /d/ in which the second formant transitions specify the consonant (adapted from Liberman, 1957).

there appear to be no obvious acoustic junctures that correspond with perceptual units. The influence of adjacent segments can be seen not only between phonemes, but also across syllable and word boundaries as well and although it is possible to identify acoustic segments in the stream of speech reliably (Fant, 1962), these acoustic segments do not always correspond to linguistic units resulting from perceptual analysis.

Insert Figure 2 about here

Units of Perceptual Analysis: Phoneme, Syllable, Word, or Beyond

The principles of linearity, invariance, and segmentation all presuppose a basic unit of analysis in speech perception and in fact, a great deal of research has been devoted to uncovering such minimal units. Given the momentary and time-varying nature of the speech signal and constraints on auditory memory, it has been reasonable to assume that information in the acoustic waveform must be recoded quickly into a more permanent symbolic representation for further analysis (Broadbent, 1965; Liberman et al., 1967). The essential problem with the idea of the primacy of one type of representation for processing over another has been that no one unit has been found to be the perceptual building block in every situation (Pisoni, 1981). Therefore, given the constraints of a task and the nature of the stimuli used in an experiment, evidence has been marshaled for a wide variety of perceptual units, including features, phonemes, syllables, words, and phrases.

Two of the most popular choices for the basic unit of representation have been the syllable and the phoneme and much debate has centered on the primacy of the former over the latter (Massaro, 1972; Savin & Bever, 1970). The variability associated with the context sensitivity of the phoneme was believed to be alleviated by focusing on the larger syllable-sized unit (Massaro, 1972). Unfortunately, the influence of the surrounding context does not stop at syllable boundaries and for that matter, at word or phrase boundaries either. Thus, to the extent that variability affects all kinds of candidate units, the issue must rest with the *psychological reality* of any one "primary" analytic unit.

Empirical results are mixed with regard to what listeners use--features, phonemes, syllables, words, or even larger units--when they are analyzing speech. Numerous studies have shown that listeners appear to initially analyze the acoustic waveform into phoneme-like units and when called upon to make a response based on this type of segmental representation, they do so readily (Cutler, Mehler, Norris, & Sequi, 1986; Norris & Cutler, 1988). Other studies have found that subjects are faster and more reliable when their responses are based on a syllable-sized unit (Mehler, 1981; Sequi, 1984). The assumption here is that faster responses for a given unit reflect primacy in processing. This issue is further complicated, however, by evidence that the minimal unit may in a large part be language-dependent. In French, for example, recent experiments suggest that the syllable is the natural unit of segmentation while in English (Mehler, 1981; Sequi, 1984), the phoneme may be the most useful segmentation unit (Cutler et al., 1986; Norris & Cutler, 1988).

Larger units than the syllable have been considered as well. Studies showing that units such as phonemes and syllables are contingent perceptually on larger perceptual entities such as words and phrases (Bever, Lackner, & Kirk, 1969; Ganong, 1980; Miller, 1962; Rubin, Turvey, & van Gelder, 1975) suggest that listeners take into account progressively larger stretches of the acoustic waveform when processing the linguistic content of an utterance for meaning. Accordingly, it appears that no one unit can be processed without consideration of the context in which it is embedded. This

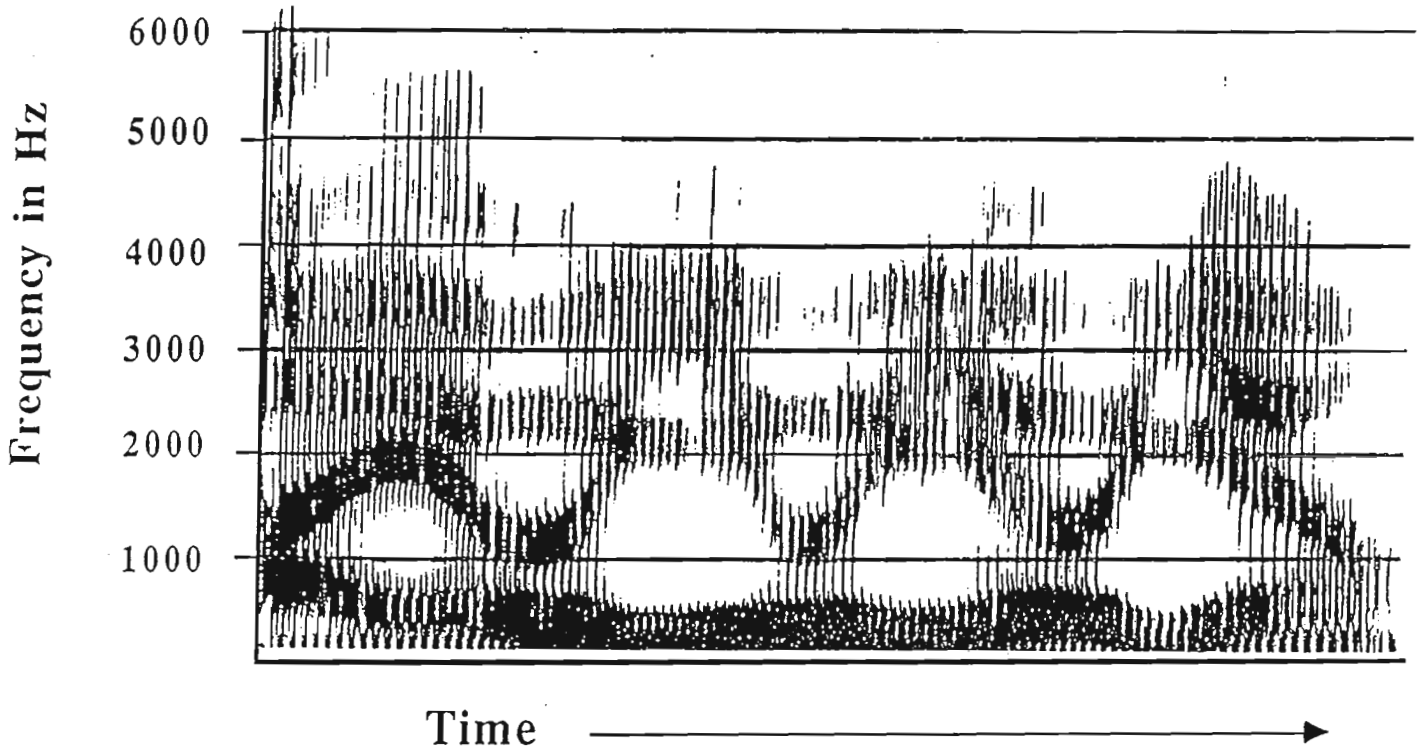


Figure 2. Spectrogram of the utterance, "I owe you a yo-yo," demonstrating that perceptual segmentation is not clearly related to acoustic segmentation.

interdependence of representations argues against a strict hierarchical view of the processing of speech in which the signal is segmented into a set of basic units that provide the foundation in every case for the next level of processing (Remez, 1987). Instead, the primacy of any particular unit may rest to a large extent on processing contingencies, attentional demands of the task, and the information available at any given time (e. g., Eimas, Hornstein, & Payton, 1990; Eimas & Nygaard, 1992).

Finally, no discussion of the basic unit of representation in speech would be complete without mentioning that most of the proposed units in speech perception can trace their inspiration to formal linguistic analysis. Representations such as context-sensitive allophones (Wickelgren, 1969, 1976), context-sensitive spectra (Klatt, 1979), and demisyllables (Fujimura & Lovins, 1978) have all been proposed as alternatives to the traditional types of symbolic representations typically assumed in speech perception research. Most of these proposed processing units stem from an interest in implementing speech recognition algorithms and many attempt to sidestep problems of context-sensitivity by precompiling variability into the analytic representation. Although these proposed units are contingent on larger perceptual units just as the more traditional linguistic representations are, the attempt implicit in these approaches to take into account the enormous context-sensitivity in speech may well prove a promising approach when dealing with the problem of variability in the acoustic speech signal.

Internal Representation of Speech Signals

In general, the type of representation that has provided the cornerstone of almost all theories of speech perception has been the abstract, canonical linguistic unit--the phonetic segment or phoneme or phoneme-like unit. Although, as we have seen, both the size of the representational unit and the exact properties comprising the representation may vary with the listening requirements, the basic assumption is that listeners must extract abstract, invariant properties of the speech signal to be compared with prototypical representations that are hypothesized to be stored in long-term memory (Kuhl, 1991; Miller & Volaitis, 1989; Oden & Massaro, 1978; Samuel, 1982; Volaitis & Miller, 1991).

Implicit in this notion of the canonical abstract linguistic unit is the assumption that information in the signal above and beyond phonetic content is a perceptual problem that the listener must solve (Shankweiler, Strange, & Verbrugge, 1976). That is, variation in the signal due to phonetic context, talker characteristics, changes in speaking rate, and even room reverberation is considered "noise" that is somehow automatically discarded in the construction of the neural representation of speech. Unfortunately, this point of view ignores the issue of how this paralinguistic or extralinguistic[2] information in the signal is processed and represented by the nervous system. In the next section, we address more fully the issue of variability and its effect on linguistic representation.

Variability and Perceptual Constancy in Speech

Variation due to phonetic context is just one of a wide variety of factors that can change the acoustic realization of speech sounds. Although much of the research in speech perception has been devoted to studying how listeners achieve consistent percepts in spite of the surrounding phonetic context, the way in which listeners cope with variability from changes in talker and speaking rate is at least as problematic for theories of speech perception. It has long been recognized that variability in

[2]The terms paralinguistic and extralinguistic are used here to refer to features of the speech signal that do not directly signal linguistic meaning. Extralinguistic factors are features such as voice that lie completely outside of signaling linguistic meaning. Paralinguistic factors are features such as speaking rate that can serve to communicate affective information, but nevertheless cannot be arranged and rearranged sequentially to convey meaning--a critical property of linguistic communication (see Laver & Trudgill, 1979).

talker characteristics and speaking rate among other factors can have profound effects on the acoustic realization of linguistic units. Substantial differences can even be found between two apparently identical utterances produced by the same speaker at the same rate of speech. The question still remains: how do listeners achieve perceptual constancy given the vast amount of acoustic variability that exists in the speech signal?

The traditional conceptualization of this problem has been to consider this type of variability in the acoustic signal as noise that must be extracted to achieve constant phonetic percepts (Shankweiler et al., 1976). Listeners are thought to compensate for changes due to talker and speaking rate through a *perceptual normalization* process in which linguistic units are evaluated relative to the prevailing rate of speech (e.g., Miller, 1987; Miller & Liberman, 1979; Summerfield, 1981) and relative to characteristics of a talker's vocal tract (e.g. Joos, 1948; Ladefoged & Broadbent, 1957; Summerfield & Haggard, 1973). Implicit in this view of normalization is the assumption that the end product of perception is an idealized canonical linguistic unit. Variation is assumed to be stripped away to arrive at the prototypical representations that underlie further linguistic analysis.[3] Indeed, there has been a considerable amount of recent research exploring the nature of perceptual compensation for variability in the speech signal (for a review on rate effects, see, Miller, 1981, 1987; for talker effects, see Miller, 1989; Nearey, 1989; Pisoni, in press).

In this section, we consider the role of acoustic variability due to talker characteristics and speaking rate on speech perception. Although our focus will be on these two factors, they do not by any means exhaust the sources of acoustic variability that can affect the acoustic fine structure of speech. Syntactic structure of the utterance (Klatt, 1975; Oller, 1973), utterance length (Klatt, 1975; Oller, 1973), room reverberation (Watkins, 1988), semantic content and even microphone characteristics, to name but a few, all introduce variation in the way in which speech is produced and realized acoustically.

Talker Variability

Many personal characteristics of a talker are reflected in their voice. Individuals differ in the size and shape of their vocal tracts (Fant, 1973; Joos, 1948; Peterson & Barney, 1952), in their idiosyncratic methods of articulation (Ladefoged, 1980), in their social class and dialect (Labov, 1966), as well as in their glottal characteristics (Carr & Trill, 1964; Carrell, 1984; Monsen & Engebretson, 1977). Given all these individual differences in the way a speaker shapes the speech signal, it is remarkable that listeners appear to effortlessly understand speech from a wide variety of speakers.

One of the first demonstrations that listeners take into account a talker's voice in perceiving speech sounds was provided by Ladefoged and Broadbent (1957). These researchers presented listeners with synthetic precursor phrases in which the formant values had been shifted up or down to simulate vocal tract differences between speakers. Following the precursor phrases, four target words (bit, bet, bat, and but) were presented and subjects' identification responses were evaluated with regard to the type of precursor phrase they received. Ladefoged and Broadbent found that subjects' vowel

[3]For the purpose of this discussion, we have adopted the standard definition of normalization as a reduction to a normal or standard state. As applied to speech analysis, if a listener normalizes a speech sound, then they are bringing it into conformity with a standard, pattern, or model. It should be noted, however, that the term normalization is also used within the field of speech perception to refer in general to a process of perceptual compensation. In this sense, there is not necessarily a reduction of information per se. Nevertheless, most accounts of speech perception have implicitly assumed that the normalization involves some kind of reduction of information and transformation of the signal into some common representational format.

identification responses depended on the relative formant values contained in the precursor phrase and concluded that listeners "calibrate" their perceptual systems based on the vowel space of each individual talker (e. g., Joos, 1948; Lieberman, Crelin, & Klatt, 1972). Thus, according to this view, vowel perception is an adaptive process that can be modified or adjusted quickly by the surrounding phonetic context.

Numerous studies since Ladefoged and Broadbent (1957) have explored the issue of vocal tract normalization. Researchers have attempted to determine what type of information in the speech signal may be used to compensate for differences in vocal tract size and shape (Gerstman, 1968; Johnson, 1990; Miller, 1989; Nearey, 1989) and they have tried to develop normalization algorithms to account for differences between speakers. It is important to note here that all these algorithms have adopted the standard view of normalization and as such, are generally based on rescaling procedures meant to reduce variation in the static properties of the speech signal (e.g., vowel formant frequencies) to a standard set of acoustic values that can be used for recognition.

Although this notion of vocal tract normalization may benefit endeavors such as machine recognition of speech, it remains unclear to what extent perceivers use analogous types of procedures and extract prototypical or idealized representations from the acoustic signal. Studies conducted by Verbrugge, Strange, Shankweiler, and Edman (1978) and Verbrugge and Rakerd (1986) suggest that calibration in terms of evaluating phonetic content with regard to a speaker's vowel space may not necessarily occur. Their research showed not only that vowel identification performance is quite good even with extensive talker variability in the stimulus set, but also that exposure to a talker's point vowels appears to provide little to aid in vowel identification performance. They concluded that when identifying vowels, listeners may use higher-order variables such as patterns of articulatory and spectral change that are assumed to be independent of a talker's vocal tract characteristics.

A growing body of research, however, suggests that the absence of effects due to talker variability in these studies may be the result of identification and discrimination tasks that minimize attentional and time constraints. When tasks are made more difficult or when processing times are measured in addition to accuracy, variability in the signal due to talker characteristics has been found to have a marked effect on listeners' performance. For example, Summerfield and Haggard (1973) have shown that reaction times are slower to recognize words when they are presented in multiple-talker versus single-talker contexts. Peters (1955), Creelman (1957), and more recently, Mullennix, Pisoni, and Martin (1988) have shown that recognizing words in noise is more difficult when listeners are presented with words produced by multiple talkers compared to words produced by only a single talker. Finally, using a Garner (1974) speeded classification task, Mullennix and Pisoni (1990) reported that subjects had difficulty ignoring irrelevant variation in a talker's voice when asked to classify words by initial phoneme. These findings suggest that variations due to changes in talker characteristics are time and resource demanding. Furthermore, the processing of talker information appears to be dependent on the perception of the phonetic content of the message. The two sources of information are not perceptually independent.

Additional research has suggested that talker variability can affect memory processes as well. At relatively fast presentation rates, Martin, Mullennix, Pisoni, and Summers (1989) and Goldinger, Pisoni, and Logan (1991) found that serial recall of spoken words is better in initial list positions when all the words in the list are produced by a single speaker compared to a condition in which each word is produced by a different speaker. Interestingly, at slower presentation rates, Goldinger et al. (1991) found that recall of words from multiple-talker lists was actually superior to single-talker lists. These

results suggest that at fast presentation rates, variation due to changes in the talker affects the initial encoding and subsequent rehearsal of items in the to-be-remembered lists. At slower presentation rates, on the other hand, listeners are able to fully process, rehearse, and encode each word along with the concomitant talker information. Consequently, listeners are able to use the additional talker information to aid in their recall task.

Further evidence that talker information is encoded and retained in memory comes from recent experiments conducted by Palmeri, Goldinger, and Pisoni (1993). Using a continuous recognition memory procedure, voice-specific information was shown to be retained along with lexical information and these attributes were found to aid later recognition. The finding that subjects are able to use talker-specific information suggests that this source of variability may not be discarded or normalized in the process of speech perception, as widely assumed in the literature. Rather, variation in a talker's voice may become part of a rich and highly detailed representation of the speaker's utterance (Geiselman & Bellezza, 1976, 1977).

Variability in Speaking Rate

Changes in speaking rate or the tempo of a talker's speech is another source of variability that can alter the acoustic structure of phonetic segments. In a normal conversation, speakers can vary considerably the speed at which they produce speech. Not only do speakers increase and decrease the number and length of pauses, but they also lengthen or shorten the acoustic underpinnings of linguistic units in the utterance (Miller, Grosjean, and Lomanto, 1984). This alteration due to speaking rate is particularly profound for phonetic distinctions that are temporal or durational in nature. For example, the timing of voicing onset in voiced versus voiceless stop consonants (Miller, Green, & Reeves, 1986; Summerfield, 1975) as well as the relative duration and extent of stop-glide contrasts (Miller & Baer, 1983) may change dramatically with changes in the prevailing rate of speech. Consequently, the theoretical issue concerning rate variability is similar to the one concerning talker variability: How do listeners maintain perceptual constancy given the changes in the temporal properties of phonetic contrasts due to speaking rate? Here again, the notion of perceptual compensation arises. Listeners are thought to utilize some kind of rate normalization process in which articulation rate is taken into consideration when evaluating the linguistic content of an utterance.

A considerable amount of research has been devoted to the effects of speaking rate on the processing of phonetic contrasts that depend on temporal or duration information. By and large, listeners appear to be sensitive both to changes in the speech signal due to rate and to the ramifications of those changes for phonetic distinctions. One of the first studies investigating effects of speaking rate on phonetic identification was conducted by Miller and Liberman (1979). They presented listeners with a synthetic /ba/-/wa/ continua in which the stop-glide distinction was cued primarily by the duration of the formant transitions. Five /ba/-/wa/ continua were synthesized in which overall syllable duration was varied from 80 to 296 msec by changing the duration of the vowel segment. The results showed that listeners required a longer transition duration to hear the glide, /w/, as the overall duration of the syllable increased. Thus, a shift was observed in the identification boundaries toward longer values of transition duration as the overall duration of the syllable became longer. According to the authors, the longer syllable duration specified a slower speaking rate and listeners adjusted their perceptual judgments accordingly.

Further research has shown that information about speaking rate preceding a phonetic contrast can affect perceptual judgments of phonetic identity as well. Summerfield (1981) conducted a series of experiments to evaluate the effect of a precursor phrase varying in rate of articulation on the

identification of voiced versus voiceless stop consonants. His results showed that phoneme identification boundaries shifted to shorter values of voice onset time as the articulation rate of the precursor phrase increased. Thus, listeners were apparently basing their classification of the initial stop consonants on information about the prevailing rate of speech in the precursor phrase.

Although it appears clear that listeners are sensitive to changes in rate of articulation, less attention has been paid to the processing consequences, in terms of attention and resources, of this type of variability. At issue here is whether the observed changes in perceptual judgments are due to compensatory processes that require time and attention or whether adjustments to differences in rate are automatic and cost-free in terms of processing. This question has been investigated recently in our laboratory in a series of experiments conducted by Sommers, Nygaard, and Pisoni (1992). They presented subjects with lists of words mixed in noise under two conditions--one in which all the words in the list were presented at a single speaking rate and one in which words in the list were presented at multiple speaking rates. The results mirrored those found earlier for talker variability (Mullennix et al., 1989). Subjects were better able to identify words mixed in noise if all the words were produced at a single speaking rate. Changes in speaking rate from word to word in a list apparently incurred some kind of processing cost that made identifying words more difficult. Again, the conclusion is that if a compensatory process exists, it must demand the attention and processing resources of the listener.

These studies and numerous others demonstrating the effects of speaking rate on perceptual judgments (see Miller, 1981, 1987, for a comprehensive review of the research on effects of speaking rate) all indicate that changes in speaking rate, both internal and external to the target segment, produce shifts in category boundaries. Recently, Miller and her colleagues (Miller & Volaitis, 1989; Volaitis & Miller, 1992) have shown that internal phonetic category *structures* that rely on temporal information are also sensitive to relative changes in rate of articulation. Their results indicate that changes in speaking rate affect the mapping of acoustic information onto the organization of phonetic category structures. Thus, listeners seem to compensate for changes in speaking rate not only by shifting or re-adjusting category boundaries, but also by re-organizing the entire structure of their phonetic categories. The implication of this finding is that any type of normalization or compensation process for speaking rate may be much more complex than just stripping away or normalizing rate information to arrive at idealized, time-invariant linguistic units. Instead, it seems that a great deal of temporal information, enough to specify category structure, is preserved in the initial encoding of the speech signal and is represented in memory.

Our brief review of research investigating talker and rate variability calls into question the traditional view of perceptual normalization. Variation in the speech signal due to factors associated with voice characteristics and speaking rate is not discarded or dissociated from the phonetic content of the signal when listeners develop linguistic representations. Information about a talker's voice affects the perception of speech, memory for spoken words, and appears to be encoded in parallel with the more symbolic phonetic information. Likewise, information about articulation rate affects the perception of phonetic contrasts and has a marked affect on phonetic category structure. This demonstrated relationship between phonetic processing and variability in the speech signal suggests a reconsideration of the traditional role of talker and rate effects in speech perception. Talker and rate variation may be important sources of information in the signal that cannot be considered independently of processes that evaluate linguistic content.

Indexical Properties of Speech

The study of speech perception has traditionally been considered separately from the study of the perception of voice (Laver, 1989; Laver & Trudgill, 1979). However, the speech signal carries a considerable amount of personal information about the talker along with linguistic content into the communicative setting. The human voice conveys information about a speaker's physical, social, and psychological characteristics (Laver, 1989; Laver & Trudgill, 1979) and these aspects, referred to as *indexical information* (Abercrombie, 1967), appear to also play an important role in speech communication. Physical characteristics of a speaker such as vocal tract size and shape provide information about a speaker's identity. So, for example, not only are most perceivers reasonably able to discriminate the speech patterns of a child from those of an adult, but they are also readily able to recognize individuals from properties of their voice alone (Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985).

Likewise, more general information about a speaker's origin and background is carried by an individual's voice. Dialect, accent, and even social class (Abercrombie, 1967) are all reflected in a speaker's articulatory patterns. Listeners know a Spanish accent from a German one and know a Southerner from a Californian simply because they are sensitive to the way in which each individual structures the speech signal according to the conventions of his or her own region, country, or position in society. Finally, the speech signal provides important information about more short-term aspects of a speaker such as emotional or psychological state. These psychological factors are readily perceived when we recognize anger, depression, or happiness in a speaker's voice.

In everyday conversation, the indexical properties of the speech signal become quite important as perceivers use this information to govern their own speaking styles and responses. From more permanent characteristics of a speaker's voice that provide information about identity to the short-term vocal changes related to emotion or "tone of voice," indexical information contributes to the overall interpretation of a speaker's utterance. How then is the perception and encoding of the indexical properties of the speech signal related to the analysis of the linguistic content of an utterance?

On the one hand, according to traditional accounts, information conveyed by a talker's voice introduces variability or noise into the signal, presumably obscuring the phonetic content of an utterance. On the other hand, listeners are able to exploit this variation in the signal to apprehend characteristics of a talker. The essence of the problem is that both types of information are conveyed in parallel along the same acoustic dimensions within the speech signal. Figure 3 illustrates this point for vowel perception. As the acoustic waveform corresponding to a spoken vowel reaches the listener's ear, information about the identity of the speaker must be disentangled from information about the phonetic identity of the vowel. Consequently, it seems that any explanation of perceptual normalization for talker will necessarily need to include an account of the processing and representation of both the indexical and the linguistic information that is carried in the speech signal.

Insert Figure 3 about here.

Indeed, a recent study from our laboratory investigating the effects of talker familiarity on the perception of spoken words suggests that indexical and linguistic information may not be processed and represented independently (Nygaard, Sommers, & Pisoni, submitted). In this experiment, subjects were asked to learn to explicitly identify a set of unfamiliar voices over a nine day period. Half the

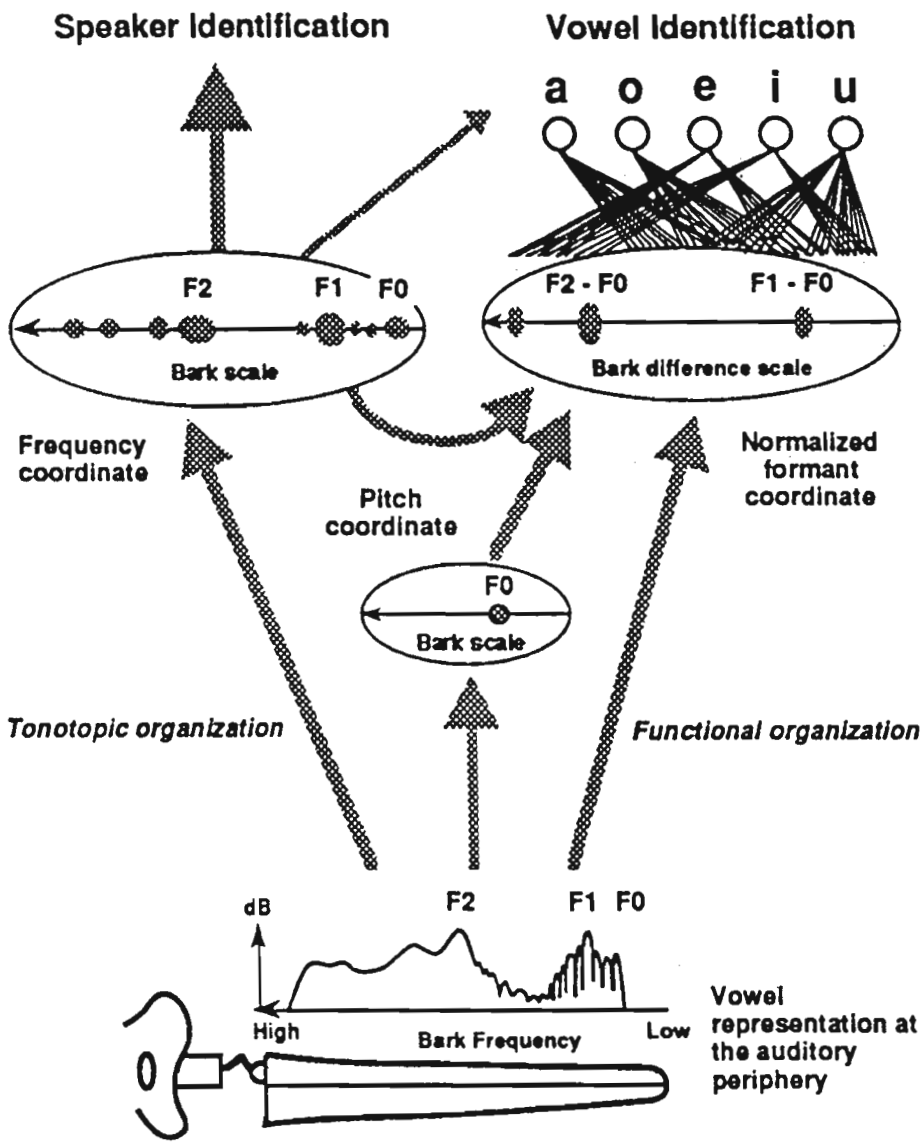


Figure 3. A diagram depicting the separation of indexical and linguistic properties of the speech signal (from Hirahara & Kato, 1992).

subjects then identified novel words mixed in noise that were produced by talkers they were previously trained on and half the subjects then identified words mixed in noise produced by new talkers that they had not been exposed to previously. The results provide evidence for transfer of learning from the explicit voice identification task to the word recognition task. Subjects who heard novel words produced by familiar voices were able to recognize words in noise more accurately than subjects who received the same novel words produced by unfamiliar voices. These findings suggest that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by that talker. Thus, these findings provide additional support for the view that the internal representation of spoken words encompasses both a phonetic description of the utterance and a structural description of the source characteristics of the specific talker. Although the exact nature of this type of representation remains to be specified, indexical and linguistic properties of the speech signal are apparently closely interrelated and are not necessarily dissociated in the perceptual analysis of speech.

Prosody and Speech Timing

Just as indexical properties constitute a rich source of information for perceivers of speech, so do the prosodic or suprasegmental attributes of the speech signal. Similarly, as in the case of indexical information, the role of these attributes in speech perception has received relatively little attention. Prosody refers the melody of speech or more precisely, to differences in pitch, intensity, duration, and the timing of segments and words in sentences. Since most research in speech perception has concentrated on the segmental analysis of phonemes, there has always been a wide gap between research conducted on the perception of isolated segments and the role of prosodic factors in the processing of connected speech (see Cohen & Nooteboom, 1975). Nevertheless, it is apparent that prosody provides a crucial connection between phonetic segments, features, and words and higher-level grammatical processes (see Darwin, 1975; Huggins, 1972; Nooteboom, Brokx, & de Rooij, 1978, for reviews). In addition, prosody provides useful information regarding lexical identity, syntactic structure, and the semantic content of a talker's utterance.

Two acoustic correlates that underlie the perception of prosody are fundamental frequency and duration. Changes in both these dimensions have been found to provide important cues to the syntactic structure of an utterance. For example, based on acoustic analysis of connected speech, Lea (1973) found that fundamental frequency usually drops at the end of a major syntactic constituent and rises near the beginning of the following constituent (see also Cooper & Sorenson, 1977). Likewise, Oller (1973) found lengthening effects for phonetic segments in word-final, phrase-final, and utterance-final positions and Klatt (1975) found vowel-lengthening effects at the end of major syntactic units (see also Klatt, 1976). Lindblom and Svensson (1973) found that listeners can parse speech and even identify words in sentences that have intact prosodic contours but are lacking detailed segmental cues (see also Svensson, 1974).

Providing information about syntactic structure is just one role that prosody plays in the perception of speech. Prosodic information has also been found to enable listeners to predict where sentence stress will fall (Cutler, 1976; Cutler & Darwin, 1981; Cutler & Fodor, 1979; Cutler & Foss, 1977) as well as to maintain perceptual coherence in continuous speech (Darwin, 1975; Nooteboom, Brokx, and de Rooij, 1978; Studdert-Kennedy, 1980).

The characterization of prosody as a potentially rich source of information in the speech signal is at odds with its contribution to variability in the acoustic realization of phonetic segments. Just as in our discussion regarding indexical information, it appears that prosodic processing and linguistic processing of the segmental attributes of speech are not independent. The challenge for researchers

working in speech perception is to reconcile the idea that variation in the signal due to factors such as talker and rate is noise with the proposal that changes along these dimensions also provide information that must be extracted as well from the speech signal.

In the preceding sections, we have reviewed two sources of variability that can affect the processing of the linguistic content of a speaker's utterance. In addition, we have outlined the types of information--indexical and prosodic--that variations in talker and rate provide to the listener in everyday communication. Taken together, the research on these factors suggests that traditional explanations of speech perception may need to reconsider their long-standing emphasis on the search for abstract, canonical linguistic units as the endpoint of perception. It appears that a great deal of information is conveyed by the same factors that also exacerbate the problem of acoustic-phonetic invariance. Therefore, any account of speech perception will necessarily need to consider the apparently contradictory nature of these different sources of variability in the speech signal. In traditional accounts of speech perception, variability was considered to be noise--something to be eliminated so that the idealized symbolic representations of speech as a sequence of segments could emerge from the highly variable acoustic signal. Current thinking on these problems suggests a very different view. Variation in speech should be considered as a rich source of information that is encoded and stored in memory along with the linguistic content of the talker's utterance. By this account, speech perception does not involve a mapping of invariant attributes or features onto idealized symbolic representations, but rather a highly detailed and specific encoding of the acoustic speech signal.

Perceptual Organization of Speech

Another area in speech perception that has been neglected relative to the emphasis on the search for idealized segmental representations is the issue of perceptual organization of speech. In normal conversation, speech communication occurs in a rich acoustic environment consisting of a mixture of other speakers' conversations and competing environmental sounds. The classic example of this problem is the situation a listener encounters at a cocktail party (Cherry, 1953). Upon arriving at the party, to strike up and maintain a conversation, the listener must be able to perceptually isolate the speech signal produced by their conversational partner from background noise consisting of music, voices, and other party noise. Thus, a listener attempting to follow a particular talker's message must somehow separate the acoustic energy attributable to that talker's utterance from acoustic energy attributable to other sounds occurring simultaneously. How does the listener carry out this task? The success of this perceptual feat depends not only upon the listener's ability to separate sources of sound, but more importantly perhaps, it depends on the listener's ability to integrate the acoustic components that comprise the particular speech signal to which the listener is attending. The acoustic elements that constitute a particular utterance must somehow "cohere" into an identifiable perceptual object.

To date, theories of speech perception have typically taken as their starting point a coherent perceptual object--implicitly assuming that perceptual organization has already taken place. Indeed, the study of speech has concentrated almost exclusively on laboratory experiments designed to evaluate the perception of speech produced by a single speaker in an acoustically sterile environment. The consequences of this approach has been the neglect of issues relating to perceptual organization. Recently, however, a growing body of theoretical and empirical work has been conducted that significantly narrows this gap. This research concentrates on the issue of perceptual coherence--how the perceptual system integrates different components of the speech signal while at the same time segregating competing acoustic input.

Gestalt Principles of Perceptual Grouping

In general, two types of explanations have been proposed to account for the listener's ability to organize acoustic energy into coherent perceptual objects. The first is based on Gestalt principles of organization as applied to auditory perception (Julesz & Hirsch, 1972; Wertheimer, 1923). In particular, Bregman (1990) has proposed that general auditory grouping mechanisms underlie the perceptual segregation of frequencies from separate sound sources and the perceptual integration of acoustic energy from the same sound source. Principles such as proximity, common fate, symmetry, and closure are hypothesized to be the basis for perceptual coherence in speech. There is considerable evidence that these grouping tendencies might describe the organization of nonspeech auditory stimuli consisting of tone and noise patterns (Bregman & Campbell, 1971; Bregman & Doehring, 1984; Dannenbring & Bregman, 1978). Less clear, however, is if these principles can explain the organization of the complex acoustic properties found in speech.

Evidence that this type of explanation may account for perceptual organization in speech comes from experiments suggesting that perceptual organization based on properties such as those proposed by Bregman (1990) can have a marked effect on phonetic coherence. For example, Ciocca & Bregman (1989) studied the effects of auditory streaming on the integration of syllable components in a dichotic listening task. Using the standard "duplex" paradigm (see Rand, 1974), listeners were presented with a three formant synthetic syllable split so that the third-formant transition was presented to one ear and the rest of the syllable was presented to the other ear. In addition, Ciocca and Bregman embedded the isolated third-formant transition into a series of capturing tones (repetitions of third-formant transitions that preceded or followed the duplex transitions). Their results showed that the capturing tones successfully created a perceptual "stream" with the isolated third-formant transition that reduced its phonetic contribution to the syllable percept. Thus, it appears that principles such as spatial and frequency proximity served as the basis for perceptual coherence in these dichotically-presented syllables.

Similarly, Darwin and his colleagues (Darwin, 1981; Darwin & Gardner, 1986; Darwin & Sutherland, 1984; Gardner, Gaskill, & Darwin, 1989) have shown that onset-time and pitch differences can also serve to segregate acoustic/phonetic components within the speech signal. For example, Darwin (1981) presented listeners with four formant composite syllables. When all four formants were excited at the same pitch and started and stopped simultaneously, the predominant percept was that of the syllable /ru/. However, when the composite's "second" formant was perceptually segregated, the predominant percept becomes /li/. Darwin found that onset asynchronies between the second formant and the rest of the composite accomplished this perceptual segregation. Likewise, sufficient mistuning of the second formant caused a shift in phonetic quality.

Although these experiments demonstrate the influence of low-level acoustic dimensions on perceptual organization, it should be noted that the results also suggest that factors associated with the phonetic integrity of the stimuli are influential as well. In both examples, residual phonetic coherence persists even when Gestalt principles of organization are violated. Often, relatively extreme differences along an acoustic dimension or even combinations of differences must exist before components of a speech sound will segregate perceptually.

Phonetic Organization

Alternative accounts of perceptual organization in speech (Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Fowler, 1986; Remez, Rubin, Berns, Nutter, & Lang, under review; Remez, Rubin, Pisoni, & Carrell, 1981) have proposed that the perceptual coherence found in speech is based

on complex principles of organization that are highly specific to the vocal source of the speech signal. Perceivers are assumed to be sensitive to patterns of spectral-temporal change that provide information about the acoustic event that has occurred. In short, these accounts claim that perceptual organization of speech is not dependent on low-level auditory grouping principles. Rather, perceivers are thought to use higher-order organizational principles that are based on "perceptual sensitivity to properties of vocally produced sound" (Remez et al., under review).[4]

Evidence for the second approach stems in a large part from a series of experiments that demonstrate phonetic coherence in spite of severe violations of general auditory grouping principles. One such demonstration comes from experiments on duplex perception (Liberman, Isenberg, & Rakerd, 1981; Mann & Liberman, 1983; Rand, 1974). As mentioned earlier, duplex perception occurs when a synthetic consonant-vowel (CV) syllable is split so that the third-formant transition is presented to one ear and the rest of the syllable (or base) is presented to the other ear. Figure 4 shows an example of these types of stimuli. When presented with these signals, listeners report hearing two distinct percepts--a complete syllable in the ear presented with the base and a nonspeech chirp in the ear presented with the isolated transition. This finding suggests that information from the isolated formant transition integrates with the base to form a unitary percept despite the difference in location of the syllable components. In addition, experiments have shown that discontinuities or differences along additional acoustic dimensions such as onset time, fundamental frequency, and amplitude (Bentin & Mann, 1990; Cutting, 1976; Nygaard & Eimas, 1990) do not necessarily disrupt the integration of the isolated phonetic components with the differently lateralized base (see also Nygaard, in press; Whalen & Liberman, 19487). Furthermore, Eimas and Miller (1992) have shown that three- and four-month old infants are susceptible to the duplex perception phenomenon as well. Not only do infants appear to integrate the disparate syllable components presented to each ear, but they also exhibit a tolerance for additional acoustic discontinuities. These findings suggest that the perceptual coherence of speech relies on some type of perceptual organization other than simple auditory grouping. Listeners seem to be sensitive to higher-order, perhaps phonetic, principles of perceptual organization rather than to their acoustic similarity. Interestingly, this underlying phonetic coherence of speech seems to be present quite early in life, suggesting that infants may be born with, or acquire very early, a sensitivity to the unique properties of vocally produced sound.

Insert Figure 4 about here

Further evidence for the tolerance of mechanisms of perceptual organization to violations of general auditory grouping principles comes from a set of studies using sine-wave analogs of speech (Remez & Rubin, in press; Remez et al., 1981). In these experiments, time-varying sinusoids were used to reproduce the center frequencies and amplitudes of the first three formants of a naturally spoken utterance. Figure 5 provides an example of a sine-wave approximation of the utterance, "The steady drip is worse than a drenching rain." Sinusoidal utterances preserve the time-varying information found in natural speech but have none of the short-time acoustic attributes that are commonly thought to underlie segmental phonetic perception. When asked to attend to the phonetic quality of these types of stimuli, subjects are able to transcribe them quite readily. Interestingly, sine-wave speech seems to violate several of the grouping principles that have been proposed, yet these sine-wave patterns do cohere phonetically.

[4]It should be noted that speech signals or vocally produced sounds are not arbitrary acoustic signals. Speech is produced by a well-defined sound source with highly constrained parameters for signal space (Stevens, 1972).

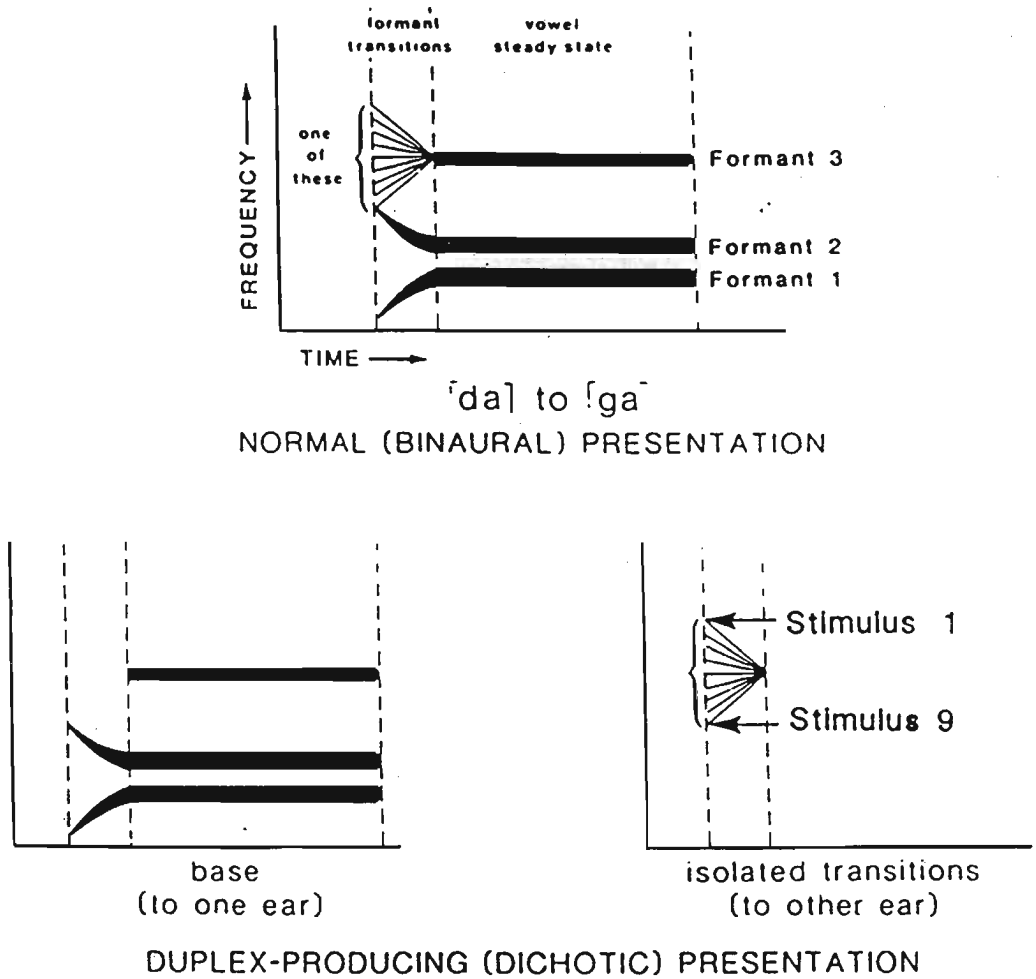


Figure 4. Stimuli used in a duplex perception experiment. The upper panel shows a synthetic speech syllable with a continuum of third-formant transitions ranging from /da/ to /ga/. The lower left panel shows the constant syllable base and the lower right panel shows the /da/ to /ga/ continuum of third-formant transitions (from Mann & Liberman, 1983).

Insert Figure 5 about here

Additional experiments conducted by Remez and his colleagues (Remez et al., under review) have shown that the phonetic coherence evident in sine-wave speech is not a result of cochlear distortion products, on the one hand, or cognitive restoration strategies, on the other hand. Rather, it is the highly constrained pattern of spectro-temporal change in these sine-wave analogs that provided the basis for phonetic coherence. Thus, these results demonstrate that general auditory grouping principles may be insufficient to account for the perceptual organization of speech. In addition, they also provide evidence for the alternative account; namely, that the perceptual organization of speech is based on characteristic spectro-temporal variation related to the dynamic properties of the vocal source used to produce speech.

In summary, the balance of evidence seems to suggest that Gestalt principles of perceptual organization may not be sufficiently complex to account for the coherence of the speech signal. While these general principles may loosely constrain possible groupings of acoustic-phonetic components and help to segregate extraneous acoustic energy, they do not seem to provide the primary basis for the formation of phonetic objects. It seems more likely that the time-varying spectral information related to the vocal source provides the underlying perceptual coherence in the speech signal and that this sensitivity to information from the vocal source may be present very early in development. Additional research will need to be done to definitively resolve this issue and to determine how evidence marshaled for either approach may fit into the larger picture of the speech processing system.

Autonomous versus Interactive Processing in Speech Perception

The nature of the perceptual mechanisms employed in speech perception is obviously highly complex and as such, must involve numerous stages of analysis and representation. As the speech signal arrives at the listener's ear, it must first undergo a peripheral auditory analysis and then be quickly recoded for further processing. Along the way, the information contained in the physical signal must make contact with the listener's linguistic and general knowledge; eventually resulting in the comprehension of the speaker's message. The question for theories of speech perception is what kind of representations are constructed and used at each level of linguistic analysis. In addition, how many and what kinds of analyses are involved in the apprehension of linguistic content? Do sources of knowledge interact flexibly with information from the incoming signal or is perceptual processing carried out in a strict hierarchical manner?

In general, two basic approaches to the nature of perceptual processing have been proposed. The first, inspired by Fodor's (1983) modularity thesis, assumes that analysis of the speech signal proceeds in a strictly bottom-up fashion. That is, discrete stages of processing are assumed--from phonetic, lexical, syntactic, to semantic--in which linguistic information is processed with regard to a limited pool of knowledge specific to that particular stage. The end product of one stage of processing is assumed to provide the input to the next level of processing and the analysis of speech is presumed to proceed in a uniformly serial manner (see Studdert-Kennedy, 1974, 1976).

Alternative accounts hypothesize that speech perception is highly interactive (Connine, 1990; Elman & McClelland, 1988; Samuel, 1990). That is, the construction of linguistic representations is determined not only by information accrued from the acoustic signal, but also from information and knowledge from higher levels of processing. These explanations emphasize the dynamic adaptive

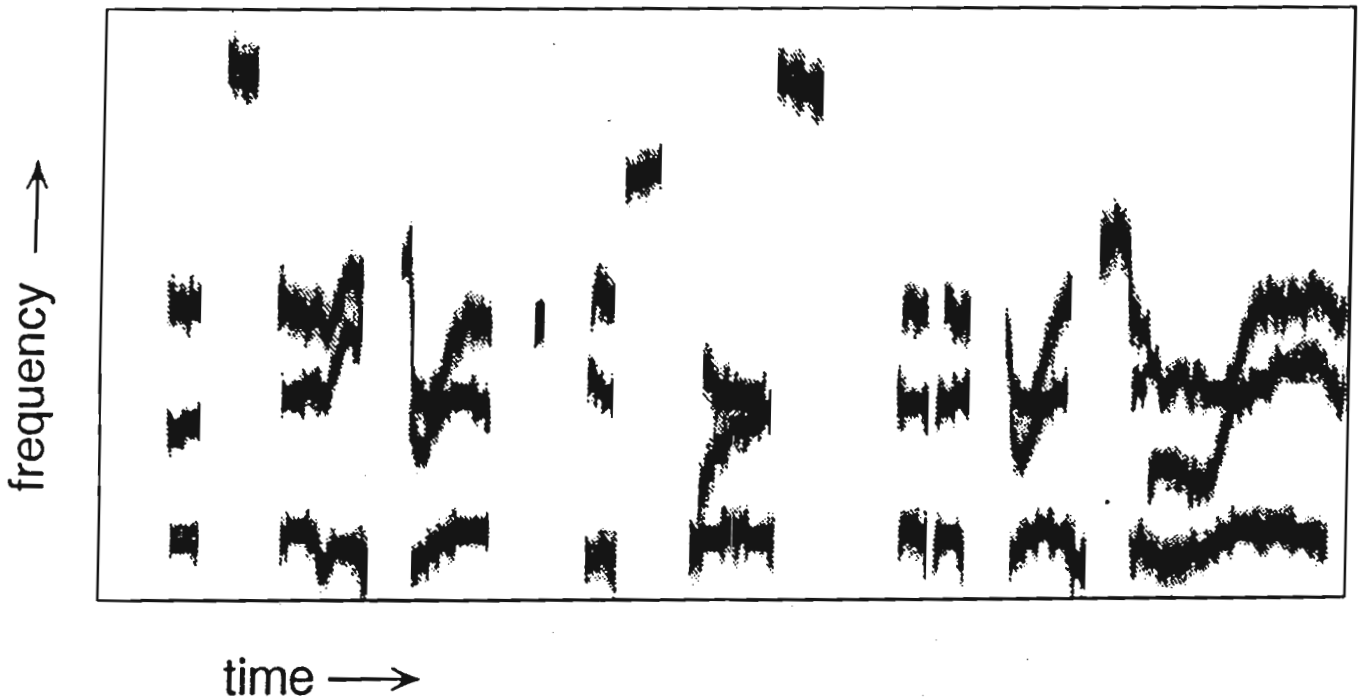


Figure 5. A sinusoidal version of the utterance, "The steady drip is worse than a drenching rain." Time-varying sinusoids are used to replicate the center frequencies of the first three formants of a natural utterance. (from Remez et al., under review).

nature of speech perception and, in the most extreme formulations, propose that many diverse types of knowledge are brought to bear on linguistic decisions at any level of analysis.

The relative usefulness of each of these approaches has been a hotly debated topic in the field of speech perception for many years and numerous studies have been conducted in an attempt to marshal evidence for each view. Although a comprehensive evaluation of the empirical work relating to this topic is beyond the scope of this chapter, we will review two representative experiments. The first is a well-known experiment conducted by Ganong (1980) in which he found that lexical information affected the placement of the perceptual boundary between two phonetic categories. When subjects were asked to identify initial consonants from a voicing continuum, stimuli with ambiguous voice-onset time (VOT) values were more often identified as having the initial consonant that formed a word compared to a non-word. This so-called "lexical effect" suggested that the status of the stimulus item as a word or nonword interacted with or influenced phonetic identification. Ganong argued that this influence of lexical information on phonetic categorization reflected top-down or interactive processes at work in speech perception and he argued that this finding provided evidence against a strictly bottom-up modular processing system.

In a related experiment, Burton, Baum, & Blumstein (1989) questioned the generality of Ganong's findings. Their experiment consisted of a replication of Ganong's study (1980) with one interesting difference--their voicing continuum used edited natural speech and was created to provide a more complete set of voicing cues. The authors argued that Ganong had found lexical effects simply because his synthetic stimulus continuum provided listeners with impoverished and contradictory voicing information. They reasoned that if the speech signal provided natural and consistent cues to voicing, perceptual processing would proceed in a strictly bottom-up fashion based entirely on the information contained in the speech signal. In fact, this is exactly what Burton et al. found. Their results showed that lexical effects disappeared when the stimulus set preserved the redundant multiple cues to voicing found in natural speech.

How are the results of these two experiments to be reconciled with regard to claims about the processing architecture in speech perception? It seems that listeners are able to flexibly alter their processing strategies based on the information available to them in the signal and the information available from their store of linguistic knowledge. On the one hand, if the physical signal provides rich unambiguous information about a phonetic contrast, listeners appear to attend primarily to the physical signal when making their phonetic identifications. On the other hand, if the physical signal is noisy, impoverished or degraded as synthetic speech often is, listeners may shift their attention to different levels of processing to assist in phonetic categorization (see Pisoni, Nusbaum, & Greene, 1985). Thus, speech perception appears to be a highly adaptive process in which listeners flexibly adjust to the demands of the task and to the properties of the signal.

Key Findings and Basic Assumptions

In this section, we now turn away from the general issues and problems confronting the field of speech perception to focus our discussion on several individual empirical findings regarding the nature of the speech perceptual process. Over the years, the study of speech perception has centered on a number of key perceptual phenomena or effects. As each phenomenon was discovered and demonstrated, researchers pursued the theoretical implications and initiated a flurry of empirical work to elucidate the underlying mechanisms. In the following section, we discuss three of the more

influential empirical findings and consider their implications for the kinds of processing and representations likely to underlie the perception of speech.

Categorical Perception

One of the earliest and most influential findings in the study of speech perception came from a classic study conducted by Liberman, Harris, Hoffman, and Griffith (1957). These researchers generated a continuum of synthetic consonant-vowel (CV) syllables that ranged from /b/ to /d/ to /g/ by changing the onset frequency of the second-formant transition in graded steps. When subjects were asked to identify the syllables from this continuum, they reported three distinct categories--/b/, /d/, and /g/. In other words, their identification functions were sharply discontinuous even though the step size between stimuli remained constant. When asked to discriminate pairs of stimuli, listeners were unable to discriminate pairs that were taken from within the same perceptual category. Discrimination for pairs that crossed a category boundary, on the other hand, was quite good. Thus, for these speech stimuli, listeners' discrimination performance was no better than their identification performance. Liberman et al. called this pattern of results *categorical perception.*

At the time, this phenomenon of categorical perception contrasted with typical results of psychophysical experiments using nonspeech stimuli. Nonspeech continua consisting of pure tones, for example, are perceived continuously, resulting in discrimination functions that are monotonic with the physical scale. The differences in perception between speech and nonspeech stimuli led Liberman et al. (1967) to propose that specialized perceptual mechanisms, distinct from those used for general audition, subserved the processing of speech.

Since Liberman et al.'s (1957) original study, a number of studies have been conducted that call into question the special nature of categorical perception for speech. For example, several experiments have demonstrated that categorical perception can be obtained with nonspeech sounds as well (e. g., Miller, Wier, Pastore, Kelly, and Dooling, 1976; Pisoni, 1977). Although the evidence is not completely unequivocal, these and other findings suggest that categorical perception may not be a phenomenon confined specifically to speech perception (see Repp, 1983a, for a review of the categorical perception literature).

Setting aside the issue of neural specialization, it is useful to emphasize here that the categorical perception findings laid the groundwork for the traditional symbolic view of linguistic representation in speech perception. Variation along the voicing continuum within a category was assumed not to be represented in the processing of the phonetic contrast. Rather, all acoustic-phonetic variation that did not mark the natural boundary between phonetic categories was assumed to be discarded as listeners quickly recoded the acoustic waveform into canonical idealized phonetic representations. Although experiments by Fujisaki & Kawashima (1969, 1970, 1971) and Pisoni (1971, 1973. 1975) showed that fine details of within-category variation are, in fact, preserved in short-term memory, albeit only for a short period of time, this view of the abstract, idealized nature of the representations underlying speech perception continues to predominate.

Trading Relations and Integration of Acoustic Cues

Another phenomena that has figured prominently in the study of speech perception is phonetic trading relations (see Repp, 1982, for a review). It has long been recognized that multiple acoustic cues or features may be associated with a single phonetic contrast (e. g., Delattre, Liberman, Cooper, & Gerstman, 1952; Denes, 1955; Harris, Hoffman, Liberman, Delattre, & Cooper, 1958; Hoffman, 1958). The multiplicity of cues creates a situation in which when the effectiveness of one cue to a

phonetic contrast is diminished, another cue can become the primary cue to signal that distinction. The assumption is that cues may trade off in the perception of phonetic contrasts because they are "phonetically equivalent" with respect to the contrast in question. For example, Fitch, Hawles, Erickson, and Liberman (1980) demonstrated that closure duration and first formant transitions are perceptually equivalent when signaling the contrast between "slit" and "split" (see also Denes, 1955). That is, even though listeners were able to use either cue in perceiving the contrast in question, they were unable to determine which cue they were using and for that matter, to distinguish perceptually between the acoustically distinct types of cues.

Like categorical perception, phonetic trading relations have been cited as evidence for a specialized mode of processing for speech. Because cue integration can occur between quite disparate types of acoustic segments (namely, spectral and temporal) and integration seems to take place over relatively long temporal intervals, Repp (1982) has argued that some kind of specialized processing system possessing abstract articulatory knowledge must be in operation. He suggests that trading relations occur because listeners perceive speech in terms of the underlying articulatory act. That is, in speech perception, listeners are assumed to possess detailed knowledge of the consequences of articulation and they use that knowledge to perceive the most plausible articulatory gestures.

Another reason that trading relations have been thought to provide evidence for a specialized speech mode of perception comes from comparisons of speech and nonspeech signals. Trading relations apparently do not occur for nonspeech sounds. For example, using sinewave speech, Best, Morrongiello, and Robson (1981) found that listeners in a perceptual set appropriate for speech exhibited the typical cue trading and integration effects in perception, while listeners who did not hear the stimuli as speech failed to exhibit the same trading relations. Thus, it seems that the integration and perceptual equivalence of multiple cues that give rise to trading relations may be specific to speech and to perception in a speech mode.

Of course, other researchers have challenged the conclusion that trading relations provide evidence for a speech-specific perceptual mode. For example, Massaro and his colleagues (Derr & Massaro, 1980; Massaro, 1972, 1987; Massaro & Cohen, 1976, 1977, Massaro & Oden, 1980; Oden & Massaro, 1978) argue that speech perception constitutes just another case of general pattern recognition, and as such, generic processes of feature extraction and integration underlie the phenomenon of trading relations as well as categorical perception.

Again, regardless of whether this finding demonstrates specialized processing mechanisms for speech, it does demonstrate the strong appeal of representations based on abstract phonetic units. Like categorical perception, the logic of trading relations presupposes that information about the nature and source of the multiple cues is somehow discarded in speech processing to arrive at prototypical or idealized patterns. For example, Repp (1983b), now reaching a somewhat different conclusion, states that trading relations may be the inevitable outcome of a general pattern matching operation in which prototypical patterns are identified and processed.

In short, phonetic trading relations, whether interpreted within the framework of specialized speech processing or within a more general pattern-recognition framework, assumes the extraction of stable, abstract, prototypical linguistic representations from highly variable acoustic signals.

Cross-Modal Cue Integration

Another finding in speech perception that has been attributed to specialized perceptual mechanisms and also underscores the presumed abstract nature of linguistic representation is the so-called "McGurk effect" (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976; Roberts & Summerfield, 1981; Summerfield, 1979, 1991). This effect is elicited when a subject is simultaneously presented with a video display of a talker articulating simple CV syllables. and with spoken syllables that are synchronized with the video display. The McGurk effect occurs when there is a mismatch between the auditory and visual information reaching the perceiver. In these cases, subjects sometimes report hearing a type of perceptual compromise. They do not seem to hear exactly what is seen or for that matter, exactly what is heard, but rather something in between. For example, if presented with a face articulating the syllable /ka/ and hearing a voice articulating the syllable /ba/, listeners will often report hearing the syllable /da/. It is as if the listener is taking the voicing cues from the auditorily presented /ba/ syllable and integrating it with information about place of articulation garnered from the visually presented /ka/ syllable to arrive at a unitary /da/ percept.

Similar to the phenomena mentioned in the preceding sections, the McGurk effect has been interpreted as support for a specialized speech mode of processing. Auditory and visual information are thought to converge at an abstract level of processing in which both types of information provide data about articulatory gestures. In contrast, as was the case with categorical perception and trading relations, researchers have attempted to provide alternative explanations for the phenomenon based on more general pattern-recognition principles. For example, Massaro and Cohen (1983) have argued that their fuzzy-logical model of perception can account for the audio-visual integration effects without assuming any speech-specific processing. In addition, Easton and Basala (1982) found that the illusion disappears if complete words, as opposed to syllables, are used and Sekiyama and Tohkura (1991) have found that the effect may not occur with Japanese stimulus materials and listeners. Thus, both the generality of these findings in different language environments and the specificity of the effects for speech is unclear at this time.

Nevertheless, the effect is apparently quite compelling. For example, Liberman (1982) reports that the procedure affects listeners' experience of *hearing* the syllable. When encountering this illusion, listeners do not seem to be able to determine the degree to which their perception of syllable identity is due to audio or visual information. Again, this type of phenomenon and the observations that accompany it raise important issues about the type of linguistic representations that might be at work. In the case of the McGurk effect, the representation is assumed to be extremely abstract (Green, Kuhl, Meltzoff, & Stevens, 1991). Not only is fine acoustic detail discarded, but also something as salient as information about auditory and visual modality seems to be discarded as well. Listeners appear to be integrating information across modalities into an abstract representation of the speech event that is independent of input modality.

In summary, two different issues have been raised regarding the phenomena reviewed above. The first issue concerns the specialized nature of perceptual processes used in speech. In each case, these phenomena--categorical perception, trading relations, and cross-model cue integration--have been cited as support for processing mechanisms that have evolved for the perception of speech and are tuned specifically to the articulatory source of the speech signal (e. g., Liberman, 1982). In each case, critics have attempted to provide a more general perceptual account of the phenomenon. Although to date, this lively debate has not provided definitive answers regarding the nature of the speech perception, it has generated an enormous body of empirical work. It should be noted, however, that the issues reviewed above by no means exhaust the evidence for and against specialized perceiving

mechanisms (see Goldinger, Pisoni, & Luce, in press, for a more complete review). Duplex perception, differences in speech and nonspeech processing, cross-language studies, the role of early language experience (see Jusczyk, this volume), hemispheric specialization (see Blumstein, this volume) and cross-species research are all very active areas of investigation that bear upon the issue of perceptual specialization in speech.

The second issue that these phenomena raise concerns the type of linguistic representations that are presumed to underlie speech perception. These findings, in contrast to those regarding stimulus variability, seem to suggest that linguistic representations are highly abstract constructs. Information about fine acoustic differences between members of a category, as well as information about the spectral and temporal properties of the cues integrating to form phonetic percepts, are assumed to be discarded in the construction of linguistic representations. Even information about the input modality appears to be lost in the processing of the phonetic content of a speaker's utterance. Again, these examples by no means exhaust the types of evidence that illustrate the abstractness of the phonetic percept. For example, evidence for phonetic integration despite acoustic differences in duplex perception (Nygaard & Eimas, 1990) and for cross-modal integration despite gender differences between the visual channel and the auditory channel (Green, Kuhl, Meltzoff, & Stevens, 1991) also suggests that phonetic representations are highly abstract, canonical entities. The question for researchers in speech perception is how the evidence for the abstract nature of linguistic units can be reconciled theoretically with the recent evidence that variability in the speech signal is informative and contributes to the perceptual analysis and representation of speech in memory.

Theoretical Approaches to Speech Perception

In the preceding sections, we have considered a number of problems and issues that confront researchers in the field of speech perception. In general, the long-standing nature of these problems illustrates the complexity of the speech perceptual system. The numerous processes involved in the comprehension of a speaker's message, from analysis of the auditory signal to the apprehension of meaning, draw upon many sources of knowledge and multiple levels of representations to arrive at the endpoint of perception. As a consequence of this complexity, the theoretical models and approaches that have been proposed to date to explain speech perception often fall short of being detailed and testable accounts. Rather, most of the theoretical approaches that have been proposed offer possible frameworks in which to place the specific phenomena and paradigms that seem to be the focus of the field.

In this section, we will briefly introduce and discuss some of the models and theories that have attempted to account for the complex process of speech perception. We will focus primarily on a few representative and influential theoretical approaches in the field, as an exhaustive review is beyond the scope of this chapter (see Klatt, 1989, for a recent comprehensive review). Our aim is to briefly introduce the basic assumptions and unique properties of each approach and to discuss how these accounts deal with the basic problems in speech perception.

Invariant Feature or Cue-Based Approaches

Each of the theoretical issues outlined in the preceding sections has to some extent centered on the problem of variability in the speech signal--whether due to phonetic context, changes in speaking rate, talker characteristics, or competing signals. One theoretical approach that has been proposed to account for listeners' success in spite of this vast amount of variation in the speech signal has simply assumed that the lack of invariance in the speech signal is more apparent than real (e.g., Cole & Scott,

1974a, 1974b; Fant, 1967). The assumption here is that invariant acoustic properties corresponding directly to individual features or phonetic segments could be uncovered if the speech signal were examined in the "correct" way. That is, it is assumed that traditional analyses of the speech signal which relied on highly simplified synthetic speech stimuli may have overlooked the presence of invariant acoustic properties in the signal. Consequently, proponents of this approach have engaged in a careful and systematic search for the invariant properties in the acoustic signal which might correspond uniquely to individual phonetic attributes (Stevens & Blumstein, 1978, 1981; Kewley-Port, 1982, 1983).

Although sharing the basic assumption that acoustic-phonetic invariance can be found, researchers have differed in their assessment of the most likely candidate acoustic features or cues. Stevens and Blumstein (1978; Blumstein & Stevens, 1979, 1980), for example, have focused primarily on gross spectrum shape at the onset of burst release as a possible invariant for place of articulation in stop consonants. Thus, their emphasis has been on *static* properties of the speech signal as the primary acoustic correlates of phonetic segments. In contrast, Kewley-Port (1982, 1983) has emphasized a more dynamic type of invariant representation. Her approach has been to examine both the auditory transformations of the speech signal and the dynamic changes within these transformations. Thus, it is hypothesized that dynamic invariants as opposed to the more static ones proposed by Stevens and Blumstein (1978) may underlie the perception of speech (see Walley & Carrell, 1983). More recently, however, both approaches have acknowledged the possibility of invariant dynamic representations (Mack & Blumstein, 1983).

The assumption of invariant acoustic properties, regardless of their static or dynamic nature, necessarily constrains the types of processing operations that can underlie the perception of speech. Thus, according to this view, speech perception consists of a strictly bottom-up process by which stable invariant acoustic cues are extracted from the speech signal for recognition. Accordingly, this type of approach makes explicit the assumption that the theoretical endpoint of perception is a series of idealized canonical linguistic units like features, phonemes, syllables, or words; because in this case, the representation is, by definition, free of any variability or noise.

In summary, the search for invariance advocated by this class of models has yielded a wealth of information regarding the acoustic structure underlying linguistic content and has, in some cases, provided promising candidates for the role of acoustic invariants. Unfortunately, however, this approach has not yet discovered a complete set of invariants that are both impervious to phonetic context and are used consistently by perceivers in recognizing the linguistic content of an utterance. Further, these models provide no account of the representation and processing of indexical and prosodic information, for example, and their interaction with the analysis of the linguistic content of an utterance. Variation in the signal, even if it is informative, is ignored by the speech processing system in the extraction of the idealized phonetic representations. Thus, although it is appealing to consider the possibilities and advantages of a direct and invariant mapping from acoustic properties to phonetic percepts, it may be unrealistic to assume that these properties alone will be used by perceivers of speech in all listening situations.

Motor Theory of Speech Perception

In contrast to models that assume acoustic-phonetic invariance in the signal, an alternative approach has been to explain speech perception by focusing on perceptual processes by which listeners might unravel the complicated articulatory encoding characteristic of speech (Liberman & Mattingly, 1989; Liberman et al., 1967). This view looks to processes *within* the listener that take into account

context sensitivity and variability when analyzing the speech signal. Thus, rather than searching for invariant features in the acoustic waveform that uniquely specify particular linguistic units, perceivers are thought to "reconstruct" or recover the phonetic intent of the talker from incomplete, impoverished, or highly encoded information provided by the speech signal.

The most influential of this class of explanation is the motor theory of speech perception. In Liberman et al.'s (1967; Liberman, Cooper, Harris, & MacNeilage, 1963) original formulation, the complicated articulatory encoding was assumed to be decoded in the perception of speech by the same processes that are involved in production. That is, articulatory movements and their sensory consequences are assumed to play a direct role in the analysis of the acoustic signal. Subsequent versions of the model have abandoned articulatory movements per se as the basis of the perception of speech in favor of neural commands to the articulators (Liberman, 1970; Liberman et al., 1967) or most recently, intended articulatory gestures (Liberman & Mattingly, 1985, 1989).

This view of speech perception was motivated to a large extent by the observation that listeners are also speakers and as such, a common type of representation must underlie both perceptual and productive activities. Because speech sounds undergo a complex process of encoding during production, it was assumed that the same processes responsible for production would be responsible for decoding the message back to the underlying phonetic segments. The complex overlapping of speech sounds in the speech signal due to coarticulation was assumed to be unraveled in perception by reference to the same rules used in production.

In the original motor theory, the crucial relationship underlying speech perception was the one between articulatory movement and phonetic segment. It was assumed that a more direct relationship might exist between these two types of representations. Regrettably, research on articulation provided little evidence for a more invariant relationship between articulatory movement and phonetic segment than between acoustic signal and phonetic segments (MacNeilage, 1970). Consequently, the revised motor theory has concentrated on the intended articulatory gestures of the talker. Speech perception is proposed to occur with reference to abstract knowledge of a speaker's vocal tract.

In addition to the assumption that the object of perception in speech is the phonetic gesture, the motor theory contends that speech perception is subserved by an innate, special-purpose phonetic module conforming to the specifications for modularity proposed by Fodor (1983). Speech perception is considered to be special because the mechanisms and processes that are used to extract the linguistic content of an utterance are separate from those used to perceive other kinds of auditory events (Mattingly & Liberman, 1988, 1989). In fact, the speech perceptual system is assumed to have evolved for the special purpose of extracting intended articulatory gestures (Mattingly & Liberman, 1988, 1989).

In terms of the nature of the neural representation of speech, the motor theory, though vastly different in other respects, assumes like models of acoustic-phonetic invariance that linguistic representations are abstract, canonical phonetic segments or the gestures that underlie these segments. The speech perceptual module is assumed to be tuned to the phonetic intent of the speaker and no account is given for how information about the vocal source is integrated with information regarding linguistic intent. It is unclear, assuming this type of conceptualization of the speech perceptual process, when information about talker identity or prosody might be extracted and how it would be used in ongoing processing of the acoustic signal.

The claims about intended articulatory gestures as the objects of perception and about the specialized nature of speech processing have spawned a great deal of controversy in the field and have as a result, generated a considerable body of research. Support for the revised motor theory has stemmed in a large part from its ability to account for a wide range of phenomena in a principled and consistent manner. However, the theory itself has received little direct unequivocal empirical support, due in a large part to the abstract nature of the proposed perceptual mechanisms. Obviously, a more precise specification of how listeners extract the underlying articulatory gestures and how those gestures are decoded into phonetic segments is needed.

Direct-Realist Approach to Speech Perception

The direct-realist approach to speech perception outlined by Fowler (1986; Fowler & Rosenblum, 1991) shares with the motor theory the basic assumption that the objects of perception in speech are articulatory gestures. However, in addition to this assumption, many crucial differences exist. Most importantly, the two theories approach the perception of the signal in fundamentally different ways. Motor theory assumes that the speech signal is highly encoded and as such, must be subjected to computations to retrieve the underlying gestures. The direct-realist approach, on the other hand, assumes that cognitive mediation is not necessary to support perception. Articulatory gestures are assumed to be readily available to the perceiver in the speech signal and consequently, are perceived directly. Another important difference between the two types of theories is the assumption of the specialized nature of speech. In contrast to the motor theory, the direct-realist approach assumes that the perception of speech is just like the perception of other events in the world, auditory or otherwise (Fowler & Rosenblum, 1991).

The direct-realist approach derives its inspiration from Gibson's (1966, 1979) ecological approach to visual perception which views perception in terms of the recognition of "events" in the world--in the case of speech, natural "phonetic events"--rather than in terms of recognition of sensory stimulation. Thus, a fundamental distinction is made between the object of perception or the event in the world and the informational medium. Returning to our example of the perception of a chair, the chair constitutes the object or event that is perceived. The structured reflected light that hits the sensory apparatus is termed the proximal stimulus and provides information about the chair that is the endpoint of perception. It is important to note that it is not the light, per se, that is perceived according to this view, but rather the object or event in the world. In terms of speech perception, Fowler (1986) has proposed that the acoustic signal provides information about articulatory events. The acoustic signal serves as the proximal stimulus and it is the articulatory gestures that are assumed to be the objects of perception.

In general, the direct-realist approach differs from all other models of speech perception principally in terms of its de-emphasis on the role of representation in the analysis of speech. Because no cognitive mediation is assumed and articulatory gestures are thought to be perceived directly, the focus of this approach is on the rich information in the speech signal that uniquely specifies phonetic events. Consequently, this view may provide one of the more promising accounts of the role that variability may play in the perception of speech. Just as articulatory gestures specific to phonetic events can be perceived directly, it seems reasonable to assume that the acoustic signal provides information regarding talker identity and prosodic structure, for example. The relationship between indexical properties, for instance, and phonetic properties of the speech signal would than be a problem of perceptual organization of the information in the speech signal rather than a problem of perceptual normalization (Fowler, 1990). Thus, according to this approach, the perception of speech is conceptualized as a general process of listener-talker attunement in which listeners perceive "the

separation of gestures for different segments or for suprasegmental information, that are layered in articulation and hence in the acoustic speech signal (page 126, Fowler, 1990)."

Although the direct-realist view is appealing in its ability to account for a wide variety of phenomena in speech perception and in its alternative conceptualization of the analysis of speech, it remains unclear empirically whether articulatory gestures are the essential objects of perception. As was mentioned in the introduction to this chapter, the choice of the proper perceptual object in speech is problematic (Remez, 1986). Because language is symbolic, the acoustic speech signal may provide information about articulatory gestures; but obviously, these articulations are not the ultimate endpoint of perception. Rather, articulatory gestures provide information themselves about the words, ideas, and meanings that constitute the essence of language perception. Nevertheless, the direct-realist approach has provided a viable alternative to the solution of the classic problems in speech perception. As more empirical work is carried out from this theoretical point of view, the role of articulatory gestures in the perception of speech should become clearer.

TRACE

In sharp contrast to all of the accounts or approaches outlined in the preceding sections, the TRACE model of speech perception (Elman, 1989; Elman & McClelland, 1986; McClelland & Elman, 1986) is an example of an interactive connectionist model applied to the problem of speech perception. As such, TRACE advocates multiple levels of representation and extensive feed-forward and feed-back connections between processing units. These functional units in TRACE are simple, highly interconnected processing units called nodes that are arranged in three different levels of representation. At each level, nodes represent the phonetic features, phonetic segments, and words that are assumed to underlie the processing of speech.

When a speech signal is presented to the network, as information passes upward through each level, nodes collect evidence for the presence of specific information--featural, phonemic, or lexical--in the input representation. Nodes fire when some threshold of activation is reached and activation is then passed along weighted links to connected nodes. In this manner, nodes act as feature detectors, firing when information in the signal is consistent with their specific sensitivity.

In addition to this feed-forward flow of information, excitatory links exist between nodes at different levels and inhibitory links exist between nodes within the same level. This key property of TRACE has important implications for the types of processing operations possible within the model. Inhibitory links between nodes at the same level of representation allow highly activated nodes to inhibit their nearest competitors resulting in a type of "winner take all" form of perceptual decision. In addition, the presence of excitatory feed-back allows for the contribution of top-down processing to the perception of phonetic segments. Not only is information accrued from the features present in the physical signal, but information from higher level lexical information influences phonetic identification as well.

The highly interactive processing and the types of representations that are assumed within the framework of TRACE contrast sharply with the motor theory and traditional accounts emphasizing acoustic-phonetic invariance. In terms of representation, TRACE does not necessarily assume that coarticulatory effects introduce "noise" onto an idealized string of phonemes. Rather, variability due to phonetic context is considered "lawful" and can serve as a rich source of information for processing in TRACE (Elman & McClelland, 1986). Less clear, however, is how TRACE might cope with variability due to factors other than phonetic context. As we have argued, variation due to talker and

rate, for example, is lawful as well and as such, provides an additional rich source of information in speech analysis. Yet, within the framework of TRACE, there is no mechanism proposed for extracting these other types of information or for accounting for their effects on phonetic processing. Nevertheless, TRACE is an impressive model of speech perception and spoken word recognition. Its precise description of the speech perceptual process as well as its clearly testable claims make it a significant contribution to the study of speech perception.

Summary and Conclusions

This review has attempted to elucidate several of the fundamental problems confronting research and theory in speech perception. Theoretical and empirical issues such as acoustic-phonetic invariance, perceptual normalization, and perceptual organization have been identified and discussed with regard to the types of representation and analysis that underlie speech perception. In particular, we have attempted to point out the basic assumptions which have motivated and constrained the theoretical investigation of these issues.

One of the most important of these assumptions that we have highlighted in this chapter stems from the problem of variability in the signal. Variation and perceptual constancy have emerged as reoccurring themes both in empirical work generated by speech researchers and in the theoretical perspectives that have been endorsed. In general, representations in speech perception have been characterized as highly abstract idealized constructs. However, much of the empirical work outlined in this chapter points to a very different conclusion; namely, that variation in the speech signal is an important source of information that interacts flexibly with processing of the phonetic content of speech. If this revised notion of variability proves to be true, then important changes in the basic assumptions regarding representation and processing will have to occur. Perhaps the traditional canonical idealized phonetic segment may need to be abandoned for a more rich, highly-detailed representation. However, in any such account, the evidence for abstraction must of course be balanced with the evidence for detail in the representation. In short, we believe that a more careful assessment of the role of variability in the processing of speech is warranted.

In addition, we believe that speech analysis is a highly flexible adaptive process in which information from a variety of knowledge sources may interact to meet particular processing situations and task demands. The body of research presented in this chapter suggests that listeners encode and utilize any information available to them from the physical signal itself as well as from their own linguistic experience to extract meaning from a talker's utterance. We have also seen that information above and beyond a strict segmental analysis is derived from the communicative setting and may interact with the on-line processing of the phonemes, words, and phrases that carry linguistic information in speech.

In summary, we believe the theoretical and empirical work presented in this chapter lead to a reassessment of the nature of linguistic units and analysis. In the future, emphasizing informationally-rich, detailed representations in speech as well as flexible, adaptive analysis may well provide the key that will unlock the long-standing problems confronting speech perception research.

References

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

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 & Psychophysics, 29,* 191-211.

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

Bever, T. G., Lackner, J., & Kirk, R. (1981). The underlying structures of sentences are the primary units of immediate speech processing. *Perception & Psychophysics, 5,* 191-211.

Blumstein, S. E. (this volume). Neurobiology of speech and language. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition. Volume 11: Speech, language, and communication.* New York: Academic Press.

Blumstein, S. E., & Steven, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66,* 1001-1017.

Blumstein, S. E., & Steven, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67,* 648-662.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge: MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89,* 244-249.

Bregman, A. S., & Doehring, P. (1984). Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception & Psychophysics, 36,* 251-256.

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

Burton, M., Baum, S., & Blumstein, S. (1989). Lexical effects on the phonetic categorization of speech: The role of acoustic structure. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 567-575.

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.* Unpublished doctoral dissertation. Indiana University.

Cherry, C. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America, 25,* 975-979.

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

Ciocca, V., & Bregman, A. S. (1989). The effects of auditory streaming on duplex perception. *Perception & Psychophysics, 46,* 39-48.

Cohen, A., & Nooteboom, S. G. (Eds.) (1975). *Structure and process in speech perception.* Heidelberg: Springer-Verlag.

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

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

Connine, C. M. (1990). Effects of sentence context and lexical knowledge in speech processing. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 281-294). Cambridge, MA: MIT Press.

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 & Psychophysics, 20,* 55-60.

Cutler, A. (in press). Spoken word recognition and production. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition. Volume 11: Speech, language, and communication.* New York: Academic Press.

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

Cutler, A., Mehler, J., Norris, D., & Sequi, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25,* 385-400.

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review, 83,* 114-140.

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* (pp. 38-72). Cambridge, MA: MIT press.

Dannenbring, G. L., & Bregman, A. S. (1976). Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 544-555.

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 178-194). Heidelberg: Springer-Verlag.

Darwin, C. J. (1976). The perception of speech. In E. C. Carterette & Friedman, M. P. (Eds.), *Handbook of perception* (pp. 175-216). New York: Academic Press.

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

Darwin, C. J., & Gardner, R. B. (1986). Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America, 79,* 838-845.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology, 36A,* 193-208.

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 & Psychophysics, 27,* 51-59.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1,* 121-144.

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

Eimas, P. D., Hornstein, S. B. M., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language, 29*, 160-180.

Eimas, P. D., & Miller, J. L. (1992). Organization in the perception of speech by young infants. *Psychological Science, 3*, 340-345.

Eimas, P. D., & Nygaard, L. C. (1992). Contextual coherence and attention in phoneme monitoring. *Journal of memory and language, 31*, 375-395.

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

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

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*, 143-165.

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

Fant, G. (1967). Auditory patterns of speech. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 111-125). Cambridge: MIT Press.

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 & Psychophysics, 27*, 343-350.

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

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

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

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

Fujimura, O., & Lovins, J. B. (1978). Syllables as concatenative phonetic units. In A. Bell & J. B. Hooper (Eds.), *Syllables and segments*. Amsterdam: North-Holland.

Fujisaki, H., & Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual report of the engineering research institute, volume 28* (pp. 67-73). Tokyo, Japan: University of Tokyo.

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* (pp. 207-214). Tokyo, Japan: University of Tokyo.

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* (pp. 59-68). Tokyo, Japan: University of Tokyo.

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

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.

Garner, W. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

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

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

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

Goldinger, S. D., Pisoni, D. B., & Luce, P. A. (in press). Speech perception and spoken work recognition: Research and theory. In N. J. Lass (Ed.), *Principles of experimental phonetics.* Toronto, B. C.: Decker, Inc.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50,* 524-536.

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.

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

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

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.

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

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

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

Julesz, B., & Hirsch, I. J. (1972). Visual and auditory perception: An essay of comparison. In E. E. David and P. B. Denes (Eds.), *Human communication: A unified view* (pp. 283-340). New York: McGraw Hill.

Jusczyk, P. (1986). Speech perception. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Volume II. Cognitive processes and performance* (pp. 1-57). New York: John Wiley and Sons.

Jusczyk, P. (this volume). Language acquisition: The sounds of speech and phonology. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition. Volume 11: Speech, language, and communication.* New York: Academic Press.

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.

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 Acoustic 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.

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

Labov, W. (1966). *The social stratification of English in New York City*, Center for Applied Linguistics, Washington, D. C.

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

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

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

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

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

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society, 29*, 117-123.

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

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

Liberman, A. M., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. (1963). A motor theory of speech perception. In C. G. M. Fant (Ed.), *Proceedings of the Speech Communication Seminar, Stockholm, 1962*. Stockholm: Royal Institute of Technology, Speech Transmission Laboratory.

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

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.

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.

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.

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

Mack, M., & Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. *Journal of the Acoustical Society of America, 73*, 1739-1750.

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

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

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 676-681.

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, 79*, 124-445.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

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 & 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., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice (Volume 3)* (pp. 129-165). New York: Academic Press.

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

Mattingly, I. G., & Liberman, A. M. (1989). Speech and other auditory modules. *Science, 243*, 489-494.

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

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

Mehler, J. (1981). The role of syllables in speech processing. *Philosophical Transactions of the Royal Society, B295*, 333-352.

Miller, G. A. (1962). Decision units in the perception of speech. *IRE transactions on information theory, IT-8*, 81-83.

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

Miller, J. D., Wier, C. C., Pastore, R. E., Kelley, W. J., & Dooling, R. J. (1976). Discrimination 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. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the psychology of language* (pp. 119-157). Hillsdale, NJ: Erlbaum.

Miller, J. L. (1990). Speech perception. In D. N. Osherson & H. Lasmik (Eds.), *An invitation of cognitive science (Volume 1)* (pp. 69-93). Cambridge, MA: MIT Press.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America, 73*, 1751-1755.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and perception for the voicing contrast. *Phonetica, 43*, 106-115.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41*, 215-225.

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

Miller, J. L., & Volaitis, L. E. (1989). Effects of speaking rate on the perceived internal structure of phonetic categories. *Perception & Psychophysics, 46,* 505-512.

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.

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

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

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

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

Norris, D. G., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics, 43,* 541-550.

Nygaard, L. C. (in press). Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. *Journal of Experimental Psychology: Human Perception and Performance.*

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.

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

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.

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

Peters, R. W. (1955). The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise. *Joint Project Report, 56,* 1-9. U.S. Naval School of Aviation Medicine, Pensacola, Florida.

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

Pisoni, D. B. (in press). Long-term memory in speech perception: Some new findings on talker variability, speaking rate, and perceptual learning. In K. Hirose, s. Kiritani, & G. Fant (Eds.), *Festschrift in honor of Hiroya Fujisaki.* Amsterdam: Elsevier North-Holland.

Pisoni, D. B. (1971). On the nature of categorical perception of speech sounds. *Supplement to status report on speech research, SR-27.* New Haven, CT: Haskins Laboratories.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & 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. (1978). Speech perception. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes, volume 6* (pp. 167-233). Hillsdale, NJ: Erlbaum.

Pisoni, D. B., & Luce, P. A. (1986). Speech perception: Research, theory, and the principal issues. In E. C. Schwab & H. C. Nusbaum (Eds.), *Perception of speech and visual form: Theoretical issues, models, and research* (pp. 1-50). New ;York: Academic Press.

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

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. (1987). Units of organization and analysis in the perception of speech. In M. e. H. Schouten (Ed.), *Psychophysics of speech perception* (pp. 419-432). Dordrecht: Martinus Nijhoff.

Remez, R. E., & Rubin, P. E. (in press). Acoustic shards, perceptual glue. In P. A. Luce and J. Charles-Luce (Eds.), *Proceedings of a workshop on spoken language.* Hillsdale, NJ: Ablex Press.

Remez, R. E., Rubin, P. E., Berns, S. M., Nutter, J. S., & Lang, J. M. (under review). On the perceptual organization of speech.

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 N. J. Lass (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,* 129-132.

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

Rubin, P. E., Turvey, M. T., & van Gelder, P. (1975). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics, 19,* 394-3948.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31,* 307-314.

Samuel, A. G. (1990). Using perceptual-restoration effects to explore the architecture of perception. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 295-314). Cambridge, MA: MIT Press.

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

Sekiyama, K., & Tohkura, Y. (1991). Japanese listeners are less influenced by vision than Americans in speech perception [Summary]. *Journal of the Acoustical Society of America, 90,* 2361.

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

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

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1992). Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing, volume 1* (pp. 217-220). Edmonton, Canada: Priority Printing.

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

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 P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Erlbaum.

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

Studdert-Kennedy, M. (1976). Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 213-293). New York: Academic Press.

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

Summerfield, Q. (1975). Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. *Speech perception: Series 2, Number 4, Spring*. Department of Psychology, The Queen's University of Belfast.

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

Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1074-1095.

Summerfield, Q. (1983). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 117-137). Hillsdale, NJ: Erlbaum.

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

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

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics, 13*, 19-28.

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

Verbrugge, R. R., & Rakerd, B. (1986). Evidence of talker-independent information for vowel. *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*, 1948-212.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influences of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America, 92*, 723-735.

Walley, A. C., & Carrell, T. D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 1011-1022.

Watkins, A. J. (1988, July). *Effects of room reverberation on the fricative/affricate distinction.* Paper presented at the Second Franco-British Speech Meeting, University of Sussex.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. *Psycholigische Forschung, 41,* 301-350. [Reprinted in translation as "Laws of organization in perceptual forms," in W. D. Ellis (Ed.), *A sourcebook of Gestalt psychology* (pp. 71-88). London: Routledge & kegan Paul, 1938.]

Whalen, D. H., & 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 E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception (Volume 7)* (pp. 227-264). New York: Academic Press.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Variability and Invariance in Speech Perception: A New Look at Some Old Problems in Perceptual Learning[1]

David B. Pisoni and Scott E. Lively

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

[1]This research was supported, in part, by NIH Research Grant DC00111-16 and, in part by NIDCD Research Training Grant DC00012-14 to Indiana University in Bloomington, IN. To appear in W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research.* Timonium, MD: York Press.

Abstract

Traditionally, researchers who were interested in theoretical issues in speech perception devoted a great deal of attention to looking for invariance in the acoustic signal. In general, this search has been unsuccessful and has led to a style of research which attempts to map perceptual cues onto phonetic categories. One important characteristic of this research strategy is that variability is removed from the speech signal. Recently, however, the focus of attention has shifted toward examining variability as a source of information. This change of focus has been guided by findings from research on categorization and memory which suggest that perceivers encode highly detailed information about individual events. In the present chapter, we review findings from a number of studies examining the role of talker variability in speech perception, word recognition and perceptual learning. We conclude that listeners encode highly detailed information about a talker's voice and that this information is used in a variety of ways in speech perception and spoken language processing.

Variability and Invariance in Speech Perception: A New Look at Some Old Problems in Perceptual Learning

From the earliest days of modern cognitive psychology, theorists have devoted much of their research to abstractionist accounts of perception, learning and memory. In keeping with the Zeitgeist of the times, most researchers assumed that the stimulus environment was impoverished and that the perceiver engaged in a great deal of constructivist processing in order to make sense of the chaotic world (Neisser 1967). Perhaps one of the best examples of this approach in cognitive psychology can be found in the field of speech perception, which has always relied very heavily on the abstractionist views derived from formal linguistic theory to define the units of perceptual analysis. By viewing language as an idealized symbolic system consisting of discrete context-free elements, linguists could focus their research efforts on more abstract theoretical issues such as phonology and syntax without having to worry about how speech is perceived or how it is represented in the mind of the listener. The following statements from Chomsky about the role of idealized forms of language in linguistic theory and his remarks about the competence-performance distinction are well-known to linguists and psycholinguists and epitomize the abstractionist viewpoint:

> "Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance." (Chomsky 1965 p. 3)

> "We thus make a fundamental distinction between competence (the speaker-hearer's knowledge of his language) and performance (the actual use of language in concrete situations). Only under the idealization set forth in the preceding paragraph is performance a direct reflection of competence." (Chomsky 1965 p. 4)

One consequence of this formal approach to language has been the almost total disregard for a variety of important problems that deal with stimulus variability and acoustic-phonetic invariance in speech perception. Even when attention has been directed to acoustic and perceptual analyses of the speech signal, most researchers have been content in occupying themselves with the search for acoustic invariance that corresponded in a one-to-one manner with a set of classically defined static perceptual categories such as phonemes or phonetic segments (Stevens & Blumstein 1979, 1980). However, the concept of the phoneme as an abstract idealized linguistic unit has always remained problematical to engineers, perceptual psychologists and speech scientists, who have spent many years trying to find first-order acoustic invariants for phonemes in the speech waveform. The continued search for acoustic-phonetic invariance is surprising given the fact that there has always been a great deal of disagreement, even in linguistics, about precisely what phonemes are and how they should be defined within a particular linguistic theory.

From the very earliest days of modern speech research going back to the invention of the sound spectrograph, it became apparent that linguistic units such as phonemes were not discrete elements in the speech waveform (Fant 1972). Instead, they turned out to be highly context-dependent

units that could be affected by a wide variety of factors that modulated their physical realization in the speech waveform. Numerous perceptual experiments in the early 1950's revealed the existence of multiple acoustic cues to almost every phonetic contrast. In many cases, these cues were acoustically quite diverse overlapping in time and often highly redundant so that the listener could reliably perceive a particular phonetic distinction despite noise or degradation in the signal.

The traditional abstractionist view of speech as an idealized sequence of discrete symbolic units has had a profound and long-lasting influence in the field of speech perception. Much of the early work on speech cues carried out at Haskins Laboratories in the 1950's was initially concerned with identifying acoustic invariants in the speech signal that corresponded uniquely to linguistic units such as phonemes of the linguistic message. However, within a short period of time, researchers discovered that the acoustic cues to many speech sounds were influenced in systematic ways by the surrounding phonetic context (Lisker 1978). These findings suggested that a search for a first-order set of acoustic-phonetic invariances, that is, simple one-to-one correspondences between speech cues and successive phonemes of the perceived linguistic message, was unlikely to be very successful. The conclusions that Cooper et al. arrived at in 1952 is a good example of the thinking about the problem at the time:

> "...the important point, however phrased, is a caution that one may not always be able to find the phoneme in the speech wave, because it may not exist there in free form; in other words, one should not expect always to be able to find acoustic invariants for the individual phonemes." (Cooper 1952 p. 605).

Despite these conclusions which were made over 40 years ago, the abstractionist assumptions about the role of phonemes and discrete segmental representations in speech perception have been maintained over the years and the search for acoustic-phonetic invariants has continued even up to the present time although these views are currently framed within the context of spoken word recognition and lexical access (Stevens et al. 1992) or neurobiological accounts that employ neurally-inspired recognition algorithms (Sussman et al. 1991). The consequences of these views about speech have been quite substantial and wide reaching in terms of both theory and research as well as experimental methodology in a number of different areas such as speech recognition, speech synthesis, infant perceptual development, clinical audiology and cross-language studies of speech perception.

Considered against this historical background, a number of recent findings have raised important questions about the traditional metatheory and formalization of language that has been assumed by most speech researchers since the late 1940's. In particular, the issue of stimulus variability has come to the forefront in recent discussions of perception, learning and memory(Brooks 1978; Elman & McClelland 1986). The heart of the problem deals with the mapping of highly variable context-sensitive speech signals on to sequences of discrete context-free perceptual categories that the listener is assumed to construct as the end product of the perceptual process.

There are good reasons for believing that the major problems of variability in speech perception can be accounted by several recent proposals that have been made concerning categorization, classification and concept learning. The literature in this area is quite extensive and we do not plan on discussing many of the recent findings, except in a general way as these results bear directly on problems of categorization in speech perception (Medin & Barsalou 1989).

Along with recent developments in the field of categorization, there is also a growing body of research that provides evidence for the encoding of specific episodic information in memory along with the details of perceptual analysis (Schacter & Church 1992; Goldinger 1992; Kolers 1973). These findings from studies of "nonanalytic cognition" have raised a number of additional questions about the traditional views surrounding the nature of perception and memory and the more general claims for the primacy of abstractionist symbolic representations in cognition (Jacoby & Brooks 1984). If we consider the problems of variability in speech perception to be a special case of the more general problems of categorization and classification, then it seems appropriate to examine how recent models of categorization might contribute to the solution of several long-standing problems in speech perception (Medin & Barsalou 1989). We attempt to do that here.

This chapter is divided into three major sections. In the first section, we consider whether the properties of speech are compatible with the criteria proposed in recent studies of nonanalytic cognition. Despite the long history of abstractionist or symbolic accounts of speech perception, there is evidence that the details of specific instances and episodes are also encoded in memory and affect subsequent perception processing and retention. In the second section, we summarize several recent studies on variability in speech perception and spoken word recognition. These studies demonstrate that stimulus variability is not lost as a consequence of perceptual processing and may be useful and informative to listeners in a variety of perceptual and memory tasks. Finally, in the third section, we describe the results of several recent laboratory training studies on the acquisition of English /r/ and /l/ by Japanese listeners. Our findings on perceptual learning of novel linguistic contrasts demonstrate that under certain experimental conditions, where there is high stimulus variability, Japanese listeners can learn to perceive novel linguistic contrasts in a robust manner. We also show that this knowledge generalizes to new words containing /r/ and /l/ and to novel tokens produced by new talkers. In addition, we have found that the perceptual learning and knowledge that is acquired under these particular high-variability training conditions appears to be retained over time even without additional exposure to these contrasts in the linguistic environment.

Abstractionist vs. Episodic Approaches to Speech Perception

Over the last few years, a number of studies on categorization and memory have provided evidence for the encoding and retention of episodic information and the details of perceptual analysis (Jacoby & Brooks 1984; Brooks 1978; Tulving & Schacter 1990; Schacter 1990). According to this approach, stimulus variability is considered to be "lawful" and informative to perceptual analysis (Elman & McClelland 1986). Memory involves encoding specific instances, as well as the processing operations used during recognition (Kolers 1973; Kolers 1976). The major emphasis of this view of cognition is on particulars, rather than abstract generalizations or symbolic coding of the stimulus input into idealized categories. Thus, the problems of variability and invariance found in speech perception can be approached in a fundamentally different way by non-analytic or instance-based accounts of perception and memory.

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

High Stimulus Variability.

Speech signals display a great deal of physical variability primarily because of factors associated with the production of spoken language. Among these factors are within- and between-talker variability, changes in speaking rate and dialect, differences in social contexts, syntactic, semantic and pragmatic effects and emotional state, as well as a wide variety of effects due to the ambient environment such as background noise, reverberation and microphone characteristics (Klatt 1986). These diverse sources of variability produce large changes in the acoustic-phonetic properties of speech and they need to be accommodated in theoretical accounts of the categorization process in speech perception.

Complex Category Relations.

The use of phonemes as perceptual units in speech perception entails a set of complex assumptions about category membership. These assumptions are based on linguistic criteria involving principles such as complementary distribution, free variation and phonetic similarity. In traditional taxonomic linguistics, for example, the concept of a phoneme is used in a number of different ways, as shown by the following definitions from Gleason (1961):

> "The phoneme is the minimum feature of the expression system of a spoken language by which one thing that may be said is distinguished from any other thing which might have been said."
>
> "A phoneme is a class of sounds...There is no English phoneme which is the same in all environments, though in many phonemes the variation can easily be overlooked, particularly by a native speaker."
>
> "A phoneme is a class of sounds which: (1) are phonetically similar and (2) show certain characteristic patterns of distribution in the language or dialect under consideration."
>
> "A phoneme is one element in the sound system of a language having a characteristic set of interrelationships with each of the other elements in that system."
>
> "The phoneme cannot, therefore, be acoustically defined. The phoneme is instead a feature of language structure. That is, it is an abstraction from the psychological and acoustical patterns which enables a linguist to describe the observed repetitions of things that seem to function within the system as identical in spite of obvious differences...The phonemes of a language are a set of abstractions..."

Thus, speech sounds display complex category relations that place a number of strong constraints on the class of models that can account for these operating principles.

Incomplete Information.

Spoken language is a highly redundant symbolic system which has evolved to maximize transmission of linguistic information. In the case of speech perception, research has demonstrated the existence of multiple speech cues for almost every phonetic contrast. While these speech cues are, for

the most part, highly context-dependent, they also provide information that can facilitate comprehension of the intended message when the signal is presented under degraded conditions. This feature of speech perception permits very high rates of information transmission even under poor listening conditions.

High Analytic Difficulty.

Speech is inherently multidimensional in nature. As a consequence, many quasi-independent articulatory attributes can be mapped on to the phonological categories of a specific language. Because of the complexity of speech and the high acoustic-phonetic variability, the category structure of speech is not amenable to simple hypothesis testing. As a result, it has been extremely difficult to formalize a set of explicit rules that can successfully map speech cues onto discrete phoneme categories. The perceptual units of speech are also highly automatized. The underlying category structure of a language is learned in a tacit and incidental way by young children.

Relations Among Perception, Production and Acoustics.

Among category systems, speech appears to be unique because of the close relations between production and perception. Speech exists simultaneously in three very different domains: the acoustic domain, the articulatory domain and the perceptual domain. While the relations among these three domains are complex, they are not arbitrary. The sound contrasts used in a language function within a common linguistic system that is assumed to encompass both production and perception. Thus, the phonetic contrasts generated in speech production by the vocal tract are precisely the same acoustic differences that are distinctive in perceptual analysis (Stevens 1972). As a result, any theoretical account of speech perception must also take into consideration aspects of speech production and acoustics. The perceptual spaces mapped out in speech production have to be very closely correlated with the same ones used in speech perception.

In learning the sound system of a language, children must not only develop abilities to discriminate and identify sounds, but they must also be able to control the motor mechanisms used in articulation to generate precisely the same phonetic contrasts in speech production that he/she has become attuned to in perception. One reason that the developing perceptual system might preserve very fine phonetic details, as well as the specific characteristics of the talker's voice, would be to allow young children to accurately imitate and reproduce speech patterns heard in their surrounding language learning environment (Studdert-Kennedy 1983). This skill would provide children with an enormous benefit in acquiring the phonology of the local dialect from speakers they are exposed to early in life.

In summary, when properties of speech are examined closely, it becomes plausible to assume that very detailed information about specific instances in speech perception might be stored in memory. In contrast to a symbolic rule-based approach, listeners may store a very large number of instances and then use them in an analogical rather than analytic way to categorize novel stimuli (Brooks 1978; Whittlesea 1989). Recent findings from studies on talker variability in speech perception support this conclusion.

Talker Variability in Speech Perception

We have carried out a number of experiments to study the effects of different sources of variability on speech perception and spoken word recognition (Pisoni 1990). Instead of reducing or eliminating variability in the stimulus materials, as most speech researchers have routinely done over the years, we specifically introduced variability from different talkers to study the effects of these

variables on perception (Pisoni 1992). Our research on this problem began with the observations of Mullennix et al. (Mullennix, Pisoni & Martin 1989) who found that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers that were used to generate the test words in the stimulus ensemble. In one condition, all the words in a test list were produced by a single talker; in another condition, the words were produced by 15 different talkers, including male and female voices. The results were very clear. Across three different signal-to-noise ratios, identification performance was always better for words that were produced by a single talker than words produced by multiple talkers. Trial-to-trial variability in the speaker's voice apparently affected recognition performance. These findings replicated results originally reported by Peters (1955) and Creelman (1957) and suggested to us that the perceptual system must engage in some form of "recalibration" each time a novel voice is encountered during the set of test trials using multiple voices.

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

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

The results from these perceptual experiments were surprising given our assumption that the indexical and linguistic properties of speech are perceived independently. To study this problem further, we carried out a series of memory experiments to assess the mental representation of speech in long-term memory. Experiments on serial recall of lists of spoken words by Martin et al. (1989) and Goldinger et al. (1991) demonstrated that specific details of a talker's voice are also encoded into long-term memory. Using a continuous recognition memory procedure, Palmeri et al. (1993) found that detailed episodic information about a talker's voice is also encoded in memory and is available for

explicit judgments even when a great deal of competition from other voices is present in the test sequence.

Finally, in another set of experiments, Goldinger (1992) found very strong evidence of implicit memory for attributes of a talker's voice which persists for a relatively long period of time after perceptual analysis has been completed. He also showed that the degree of perceptual similarity between voices affects the magnitude of the repetition effect, suggesting that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations.

Taken together, our findings on the effects of talker variability in perception and memory tasks provide support for the proposal that detailed perceptual information about a talker's voice may be preserved in some type of perceptual representation system (PRS) (Schacter 1990) and that these attributes are encoded implicitly into long-term memory. At the present time, it is not clear whether there is one composite representation in memory or whether these different attributes are encoded in parallel in separate representations (Eich 1982; Hintzman 1986). It is also not clear whether spoken words are encoded and represented in memory as a sequence of abstract symbolic phoneme-like units along with much more detailed episodic information about specific instances and the processing operations used in perceptual analysis. These are important questions for future research on the internal representation of speech in memory.

These recent findings on talker variability have encouraged us to examine more carefully the tuning or adaptation that occurs when a listener becomes familiar with the voice of a specific talker (Nygaard, Sommers & Pisoni submitted). This particular problem has not received very much attention despite the obvious relevance to problems of speaker normalization, acoustic-phonetic invariance and the potential application to automatic speech recognition and speaker identification (Kakehi 1992; Fowler in press). Our search of the research literature on talker adaptation revealed only a small number of behavioral studies on this topic and all of them appeared in obscure technical reports from the mid 1950's.

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

The results of the ineligibility experiment are shown in Figure 1 for the two groups of subjects. We found that identification performance for the trained group was reliably better than the control group at each of the signal-to-noise ratios tested. The subjects who had heard novel words produced by familiar voices were able to recognize words more accurately than subjects who received the same novel words produced by unfamiliar voices. Two other groups of subjects were also run in the intelligibility experiment as controls; however, these subjects did not receive any training in recognizing the voices and were therefore not exposed to any of the stimuli prior to listening to the same set of words in noise. One control group received the set of words presented to the trained experimental group; the other control group received the words that were presented to the trained

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

Insert Figure 1 about here

The findings from this perceptual learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by a familiar talker. Thus, speech perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that was obtained in an entirely different experimental task-- explicit voice recognition as compared to a speech intelligibility test.

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

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

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

Developments in nonanalytic approaches to cognition and exemplar-based approaches to categorization have led us to reconsider a number of issues related to the acquisition of new phonetic categories. In particular, we have become interested in examining the contribution of stimulus variability to the formation of robust perceptual categories. In this section we summarize the results of two experiments that were designed to investigate several issues in perceptual learning of /r/ and /l/ by Japanese listeners. First, we examined the role of talker variability in training. To study this, we compared the performance of listeners who were trained with only a single voice to the performance of

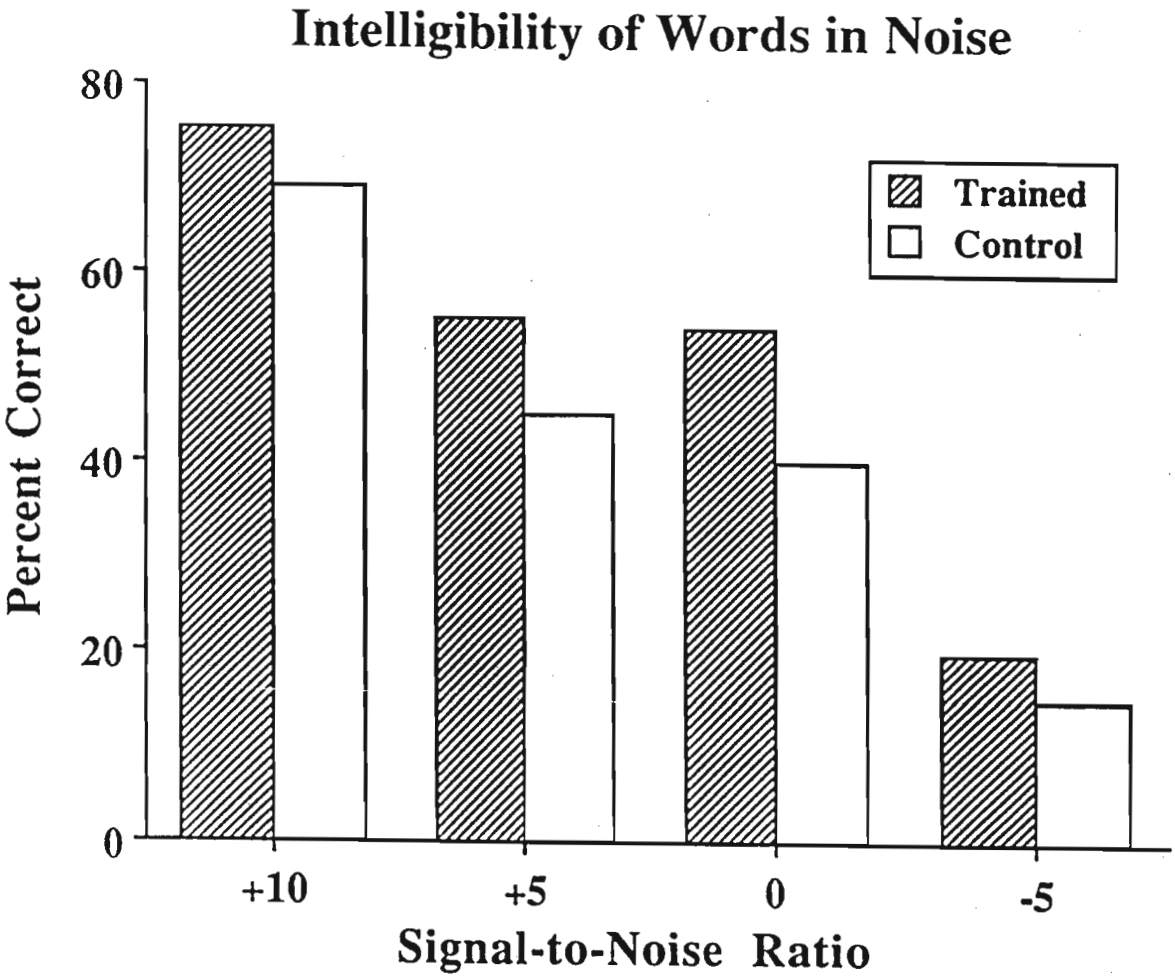


Figure 1. Percent correct word recognition (intelligibility) as a function f signal-to-noise ration for the trained and control subjects on the transfer task administered after voice recognition training was completed. (From Nygaard, Sommers & Pisoni, 1993).

listeners who were trained with several talkers. Second, we examined the retention of new phonetic categories over time. The issue of retention without additional training is important because it allowed us to assess the ultimate success of our training procedure by examining how new information is represented and retained in memory.

Several aspects of our general training strategy are important to emphasize before discussing any results. One of the goals of cross-language training experiments is to facilitate acquisition of robust new phonetic categories. Two criteria are important in defining robust new perceptual categories: First, the categories must be applied across a wide variety of new talkers and new phonetic environments. This means that listeners must demonstrate generalization both to new talkers and to novel words. Second, the new categories must be stable over time. In other words, if listeners form robust categories during training, then their performance should be above baseline levels after extended intervals without any further training.

Another important aspect of our approach to perceptual learning concerns how new categories are formed. The traditional approach to cross-language speech perception training has been to use synthetic stimuli that vary only in the critical cues used by native speakers of the language (see, however, Yamada & Tohkura 1992). Although, this approach has been successful in some cases in modifying speech perception (see Pisoni et al., 1982; McClaskey et al. 1983), it overlooks the rich diversity of cues that are present in natural speech. In both of the experiments we describe below, listeners were trained with natural speech tokens. We hypothesized that the richness and diversity of cues found in natural speech would aid listeners is forming robust new perceptual categories that could be generalized to new phonetic environments and new talkers (see also Kuhl 1983). We also assumed that by training some listeners with tokens from multiple talkers, we would maximize the number of cues that listeners had available to them in recognizing new words.

Task variables also play an important role in training nonnative listeners to perceive new phonetic contrasts (Jenkins 1979). Two types of tasks, discrimination and identification, have traditionally been used during training. In discrimination training, listeners are presented with sequences of stimuli and are asked to decide if the stimuli are the same or different or if a member of the stimulus ensemble is unique. The assumption is that listeners will focus their attention on cues that contrast new perceptual categories. The drawback of this approach is that listeners' attention may be focused too sharply on any stimulus differences they can detect. Instead of responding to higher more abstract category-level differences, subjects may respond to subtle, within-category changes that differentiate *stimuli* rather than *categories*. Thus, discrimination training encourages listeners to attend to small, within-category differences and does not promote the formation of new perceptual categories that are robust to the variability in the natural environment (Carney, Widin, & Viemeister 1977; Liberman, Harris, Kinney, & Lane 1961; Pisoni 1973; Strange & Dittmann 1984; Werker & Logan 1985; Werker & Tees 1984).

In identification training, on the other hand, subjects are asked to uniquely identify a single stimulus on each trial. Uncertainty in the task is controlled by restricting the number of possible response alternatives. Whereas discrimination tasks require listeners to attend to small within-category differences, identification training encourages subjects to group perceptually similar objects into the same category (Lane 1965 1969). Thus, discrimination training promotes "acquired distinctiveness," while identification training promotes "acquired equivalence" (Lawrence 1949, 1950; Gibson 1955; Liberman et al. 1961).

In both of the experiments described below, we employed a pretest-posttest design that was identical to the tests used by Strange and Dittmann (1984). Training was conducted over a 15 day period using a two-alternative forced-choice identification training procedure. Immediate feedback was given only during training. All training and testing was conducted using natural speech. In the first experiment, subjects were trained with only single talker. In the second experiment, subjects were trained with five different talkers. After the completion of training, subjects were given two tests of generalization. One test consisted of new words produced by a familiar talker; the other test consisted of novel words produced by an unfamiliar talker.

Training with a single talker.

Logan, Lively and Pisoni (1991) recently published the results of a perceptual learning experiment in which they trained Japanese listeners to perceive English /r/ and /l/ using a two-alternative forced-choice identification training procedure. Subjects were trained using five different talkers who produced English words containing /r/ and /l/ in five different phonetic environments. The authors found that listeners' accuracy improved by 5%-7% from the pretest to the posttest as well as during training. In addition, they found a marginal difference in accuracy between talkers during the tests of generalization: The familiar talker was responded to slightly more accurately than the unfamiliar talker although the generalization results were obtained with only three subjects. Based on these preliminary results, Logan et al. concluded that the high-variability identification paradigm was effective in training Japanese listeners to acquire the /r/-/l/ contrast.

Logan et al.'s training study included two sources of variability. First, /r/ and /l/ appeared in several phonetic environments. We assumed that variability within a single phonetic environment might not be sufficient to foster generalization to /r/'s and /l/'s in other phonetic environments (Jamieson & Morosan 1986, 1989). Second, we presented listeners with tokens from multiple talkers during training. We assumed that this would provide listeners with a rich set of cues to the new contrast. Moreover, this procedure would prevent listeners from becoming attuned too closely to a particular voice (Goto 1971). Training with multiple talkers was also thought to encourage generalization to new voices.

Because we varied several sources of variability in the training stimuli, it is not clear what the relative contributions of each source of variability was to the observed pattern of results (Logan, Lively, & Pisoni, in press; Pruitt, in press). Recently, we conducted an experiment to determine more precisely how talker variability affected performance during training and generalization to new words and new talkers (Lively, Logan, & Pisoni, in press). We trained a group of six Japanese listeners in a pretest-posttest design with the same two-alternative forced-choice identification paradigm used in the earlier study by Logan et al. The only difference was that listeners were trained with only a single talker, rather than five different talkers. Subjects were trained for 15 days on the same set of 136 words that contained /r/ or /l/ in five phonetic environments (initial singleton, initial consonant clusters, intervocalic, final consonant clusters, and final singleton positions). Generalization to new words and to a new talker were also assessed at the conclusion of training.

We predicted that the reduction in talker variability should have several consequences on identification performance. First, increases in accuracy and decreases in response latency should be observed during training. This result would not be surprising, given that we had already observed changes in performance with a much more variable stimulus set. Second, generalization should be adversely affected by the reduction in talker variability. If listeners become attuned to the specific characteristics of a particular voice during perceptual learning (Goto 1971), then they would not be

expected to generalize very well to a new talker used in the generalization tests. Moreover, if subjects are learning about specific stimuli rather than general cues or rules to the new contrast, then they would not be expected to generalize very well to novel stimuli produced by a familiar talker.

The results of the single-talker experiment confirmed our predictions. Not surprisingly, listeners' accuracy increased significantly from the pretest to the posttest for contrasts in some phonetic environments. Response times also decreased significantly from the pretest to the posttest for phonemes in all phonetic environments. During training, subjects' accuracy increased and response latencies decreased from Week 1 of training to Week 2 of training. No significant changes in performance were observed between Week 2 and Week 3 of training. Accuracy improved the most for /r/'s and /l/'s in initial consonant clusters and intervocalic position.

Insert Figure 2 about here

The results of the tests of generalization which are shown in Figure 2 revealed the limitations of the single-talker training procedure. Listeners responded more accurately to words produced by the familiar talker when the /r/'s or /l/'s occurred in initial singleton or intervocalic positions. A trend was also observed for better performance with the familiar talker when /r/'s or /l/'s were in initial consonant clusters. Responses also tended to be faster to the familiar talker. However, absolute level of performance on both tests of generalization was relatively low. Mean accuracy with the familiar talker was equivalent to the level of performance observed during the first week of training. Similarly, mean accuracy with the unfamiliar talker was worse than performance during the first week of training for contrasts in initial singleton, initial consonant clusters, and intervocalic environments. These findings demonstrate that when listeners are trained with only a single talker they do not generalize very well to new words produced by a new talker or to new words produced by the old talker used in training.

Taken together, the results of this experiment indicate that the single-talker training paradigm was generally ineffective in facilitating robust acquisition of /r/ and /l/. Although subjects encoded some stimulus-specific knowledge, they did not appear to be able to apply this knowledge in the generalization tests. The results support Goto's (1971) observations that Japanese listeners become attuned to a small set of voices when they acquire English and that more extensive training with different voices is required for robust generalization to new words produced by novel talkers.

Long-term retention of new phonetic categories.

Our initial training study demonstrated that Japanese listeners could be trained in the laboratory to perceive the English /r/-/l/ contrast. Moreover, the results demonstrated the importance of talker variability to generalization to new words and new voices. In the experiment using only a single talker, we found that an absence of talker variability in the stimulus set used during training was detrimental to generalization performance. The outcome of these two experiments jointly satisfy one criterion for a successful training paradigm: Trained listeners should demonstrate generalization to both new voices and new words. Logan et al.'s results suggest that training with large amounts of talker variability encourages generalization, whereas training with only a single talker does not (see also Lively et al., in press, Exp. 1).

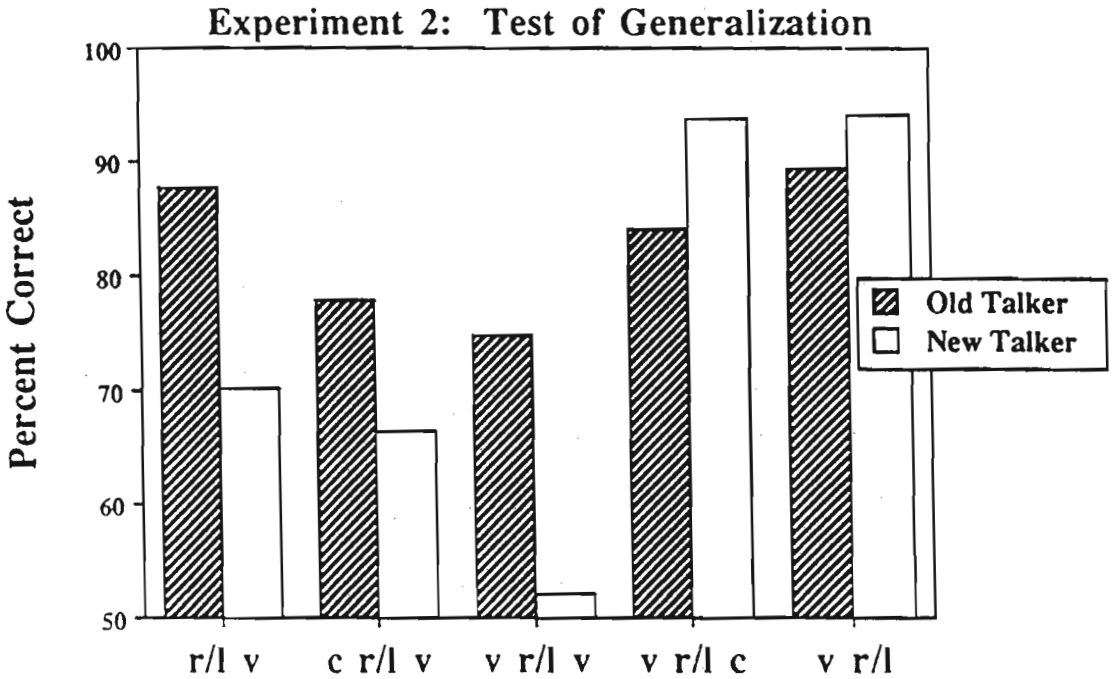


Figure 2. Percent correct identification from the single-talker training condition on the test of generalization is shown as a function of phonetic environment. The filled bars show the results for novel words produced by a familiar talker; the open bars show results for a new talker. The data also reveal a significant interaction between talker and phonetic environment. (From Lively, Logan and Pisoni, in press).

Neither Logan et al.'s (1991) investigation, nor the single-talker experiment described above addressed the second criterion for a successful training procedure. Any robust training paradigm should also encourage the long-term retention of new phonetic contrasts. It is important to determine whether short-term laboratory-based procedures produce only temporary reorganization of a listener's perceptual capabilities or whether these changes are more permanent. Two predictions concerning the retention of new phonetic categories can be made. First, it is possible that changes observed in the laboratory may be short-lived. If listeners are living in a monolingual speaking environment in which the new linguistic contrast is rarely encountered, then subjects might be expected to return to baseline levels of performance without any further training or exposure. On the other hand, if our training procedures encourage the development of long-term changes in perceptual organization, then listeners might be expected to retain much of what they learned during training without additional training or feedback.

We recently assessed these predictions in collaboration with researchers at the ATR laboratories in Kyoto, Japan (Lively, Pisoni, Yamada, Tohkura, & Yamada 1993). Nineteen monolingual speakers of Japanese were trained using a slightly modified version of Logan et al.'s high-variability training procedure. Subjects were trained with exactly the same stimuli used by Logan et al. Five talkers produced the /r/-/l/ contrast in five different phonetic environments. Subjects were tested in a pretest-posttest design. Training lasted for 15 days and listeners heard one training talker each day. By the end of training, listeners had heard each talker three times. Following the conclusion of training, listeners were given the same tests of generalization described earlier. Three months after the conclusion of training, subjects returned to the laboratory and were given a follow-up posttest and the two tests of generalization again. As in our previous experiments, a two-alternative forced-choice identification task was used throughout the entire experiment and feedback was given only during training.

For the most part, the overall pattern of results replicated those obtained in our original study (Logan et al. 1991). Subjects improved from the pretest to the posttest by 12%. During training, listeners' accuracy increased by an average of approximately 11% and response times decreased by approximately 600 ms. The increases in accuracy were almost twice as large as those obtained by Logan et al. The difference in the size of the training effects between the two studies may be due to the fact that in our earlier study we tested subjects who were living in the United States at the time of the experiment and were also enrolled in English classes. Thus, our subjects may have received extensive exposure to the /r/ and /l/ contrast outside of the laboratory before the training procedures began and this may have affected their ability to show additional improvements in the laboratory environment. In contrast, the subjects in the present study were living in a monolingual Japanese speaking environment and it is unlikely that they would receive any exposure to this contrast in their immediate surrounding.

The results of the tests of generalization showed that familiarity with the voice producing the stimuli facilitated identification performance. Subjects were significantly more accurate when words were produced by a talker used in training than when the words were produced by an unfamiliar talker. In terms of absolute level of performance, generalization accuracy was quite good. Average performance with the familiar talker was equivalent to mean accuracy during the third week of training. Similarly, accuracy with the unfamiliar talker was equivalent to mean performance during the second week of training.

Because we observed large differences in performance among talkers used during training and we had selected the most intelligible talker to use during the tests of generalization, we wanted to assess the differences in base-line intelligibility as the source of the results obtained during generalization. We tested this hypothesis by having an additional 14 naive Japanese listeners perform the tests of generalization without any prior training. No significant differences in intelligibility were observed: Mean accuracy with the familiar talker was 71%, while mean accuracy with the unfamiliar talker was 70%. These results rule out simple differences in intelligibility between the familiar and unfamiliar talkers as the source of our generalization results.

The most interesting data from the retention experiment come from the follow-up tests given three months after the conclusion of the original training. In these tests, 16 of the original 19 subjects returned and were given the posttest and the two tests of generalization again. The posttest results are shown in Figure 3. Surprisingly, mean accuracy decreased only 2% from the posttest given at the end of training to the follow-up posttest given three months later. This decrease was not statistically significant. A similar pattern was obtained in the follow-up tests of generalization shown in Figure 4. Mean accuracy decreased only 1.5% from the original generalization tests to the follow-up tests. Interestingly, the effect of talker was still significant, even after a three month interval. Words produced by a familiar talker, the talker used during training, were identified more accurately than words produced by an unfamiliar talker.

Insert Figures 3 and 4 about here

The results of the retention experiment are theoretically important for several reasons. First, the present findings demonstrate that the high-variability identification training paradigm meets our second criterion for successful training procedures: Listeners show long-term retention of new perceptual categories without any further training. To our knowledge, these results are the first demonstration of long-term retention of new phonetic categories acquired in a laboratory training experiment.

Second, our findings provide support for the nonanalytic approach to cognition outlined above. We suggest that listeners encode detailed representations of spoken words into long-term memory and that these representations are used to facilitate the later recognition of new items. These representation are assumed to include attributes of a talker's voice. During the tests of generalization, subjects were more accurate in responding to words produced by a talker used in training than to words produced by an unfamiliar talker. Moreover, these differences were still evident three months after the conclusion of training. The pattern of results cannot be accounted for by differences in base-line intelligibility between the two talkers. Precisely what characteristics of the stimulus materials are retained or how long this information is preserved remains an important question for future research.

Taken together, the results of the training experiments described above suggest several important methodological and theoretical conclusions. First, the present findings indicate that adult Japanese listeners can be trained to identify the English /r/-/l/ contrast. Second, stimulus variability, particularly talker variability, appears to be an important factor in promoting robust generalization. When listeners were trained with a single talker, generalization to new words and a new talker was poor. In contrast, when listeners were trained with a more variable stimulus set that included several talkers, generalization accuracy improved substantially. Third, training with a high-variability, two-

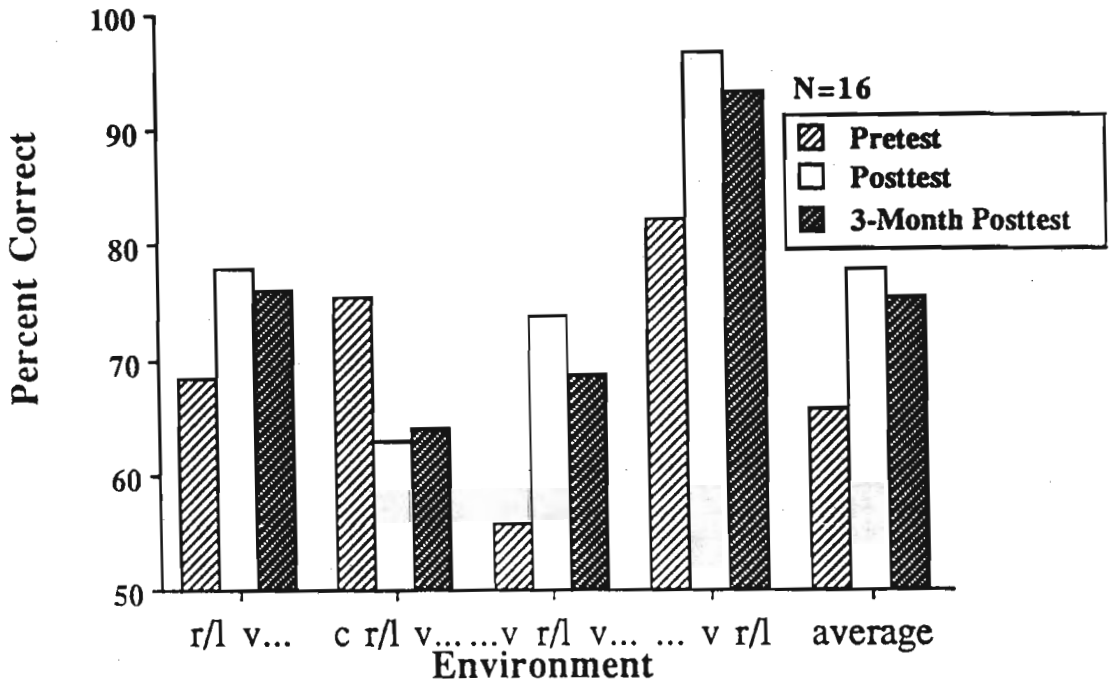


Figure 3. Accuracy scores for monolingual Japanese listeners for the pretest, posttest and three-month follow-up as a function of phonetic environment. (From Lively et al., 1993).

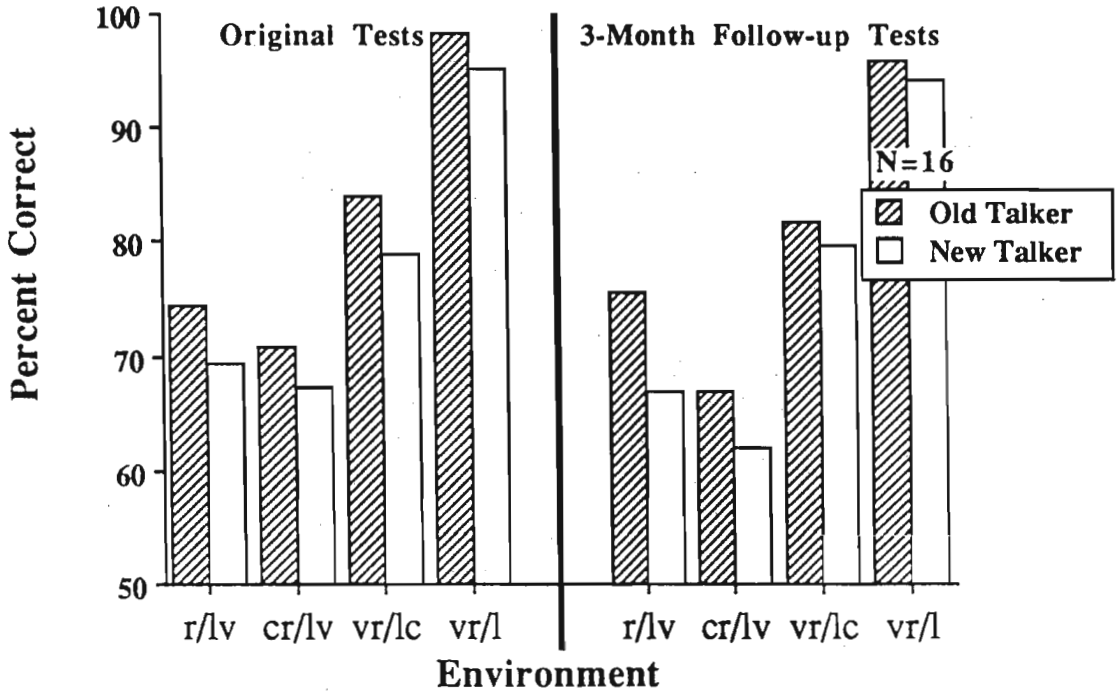


Figure 4. Percent correct identification on the tests of generalization. The left panel shows the results immediately after training; the right panel shows the results after a three-month retention interval from the follow-up tests. (From Lively et al., 1993).

alternative forced choice training paradigm meets both of the criteria we outlined for successful training procedures. Subjects show generalization to new words and new talkers and they demonstrate retention of the new phonetic contrast over time. It should be noted, however, that generalization was not complete. We still observed differences among talkers and some loss over time. It is possible that other training procedures may more effective in promoting long-term retention and generalization of new phonetic contrasts. Finally, we suggest that the results of our high-variability laboratory-based training procedures provide evidence for the encoding of talker-specific information in speech perception and that this information is used to facilitate the recognition of new words. Thus, for our listeners variability was highly informative in helping them to develop robust perceptual categories (Elman & McClelland 1986).

Summary and Conclusions

We have covered a lot of ground in this chapter in our attempts to bring together a number of different but closely related areas of research that bear on several long-standing theoretical issues in speech perception. Here we summarize the major findings and draw several conclusions.

One of the most salient findings to emerge from this research is the importance of stimulus variability in perceptual learning of novel linguistic contrasts. In contrast to the findings reported by Strange and Dittmann (1984), who failed to observe any change in the performance of their Japanese subjects, we found moderate but highly consistent changes in performance. Moreover, our subjects displayed generalization to new words and to tokens produced by a new talker. We suggest that these findings are due to the high-variability training procedure which promotes the development of robust perceptual categories.

Another important finding was the observation that listeners acquire information about these perceptual categories by encoding specific instances or exemplars from the stimulus ensemble, including details of individual voices. The subjects in our experiments did not appear to acquire abstract context-independent categories for /r/ and /l/ that were invariant across different talkers. The perceptual learning that took place using these laboratory-training procedures was stimulus-specific, although evidence for generalization was found when the stimulus set contained a great deal of variability.

These findings raise a number of important theoretical issues in speech perception that go well beyond the specific demonstration that Japanese listeners can be trained to perceive English /r/ and /l/ reliably. Our findings from this training study and our other experiments on talker variability show that several long-held views about the speech perception may be incorrect. The emphasis on mapping of speech onto discrete symbolic units has drawn many researchers away from the fundamental questions of how the perceptual categories of speech and their mental representations should be conceptualized theoretically, given the enormous variability in the physical signal. The traditional idealized view of speech has also encouraged an approach and methodology for conducting research that is designed to deliberately reduce or eliminate as many sources of variability as possible in the stimulus materials in order to obtain reliable perceptual data from listeners in a wide variety of experimental paradigms. The underlying assumption of this approach to research in speech is that stimulus variability is a source of noise-- something to be eliminated from the signal so that the perceptual invariants for the idealized linguistic categories would somehow emerge.

The present results on perceptual learning demonstrate that variability is lawful and informative. Listeners encode and retain detailed stimulus information from the signal. When viewed within the context of nonanalytic approaches to cognition, variability in speech is simply a natural consequence of the complex category structure of spoken language. Attempts to reduce or eliminate stimulus variability in perceptual and memory experiments on spoken language over the last 40 years, may have provided a misleading or distorted picture of the underlying perceptual process which appears to be able to cope quite well with these diverse sources of variability in the speech signal.

Our findings also raise several theoretical issues about the mental representation of speech and the types of information in the signal these representations encode and preserve. Several results show that detailed information about a talker's voice is encoded into long-term memory and is used in speech perception and spoken word recognition. In past accounts, indexical information about the speaker's voice was clearly dissociated from properties of the linguistic message. The present findings demonstrate that some sources of indexical information are encoded into memory and do become part of the representation of spoken words.

The present findings are also relevant to several long-standing assumptions about the perceptual normalization for speech, particularly claims about the loss of stimulus-specific information. The evidence we have obtained in our variability and perceptual learning experiments suggests that the process of talker normalization may not be carried out automatically without cost and that information may not be lost as a consequence of perceptual analysis and categorization. The locus of the talker normalization process, if one actually exists, may not be in the auditory periphery as many researchers have assumed in the past, but may be more centrally located in the recognition process itself which draws heavily on specific knowledge in long-term memory for categorization.

Finally, the present results bear on a number of recent claims about the development and use of prototypes in speech perception. Prototypes for speech sounds are appealing from the abstractionist point of view because they provide cognitive economy (Neisser 1967). According to prototype theorists, a single abstract representation serves as a summary for the entire category (Rosch 1975). Whether this summary is based on a measure of central tendency across category members (Smith & Medin 1981) or idealized forms (Oden & Massaro 1978) is open to debate. Psychological evidence for prototypical speech sounds comes from a variety of experimental procedures. For example, Samuel (1983) showed that some stop consonants make better adapters in a selective adaptation task than others. Miller and Volaitis (1989) and Volaitis and Miller (1992) have shown that subjects rate some stop consonants as more prototypical than others. They have also shown that the internal category structure is altered by changes in perceived speaking rate and place of articulation. Finally, Kuhl (1991a,b, 1992) has shown that some vowel tokens are rated as better category members than others and that prototypical vowels act like "perceptual magnets" because they appear to make other vowels more similar to themselves. Other, less typical members, that are purported to be members of the same category do not alter perceived similarity relations.

We see two problems with the proposal that phonetic categories should be represented by prototypes. First, researchers never make explicit exactly what form of representation these prototypes take. For example, Kuhl states:

> "...I do not intend to make a claim about the specific form that these speech representations take. Speech representations could consist of

average or modal values abstracted from the stimuli heard by listeners, or category 'ideals,' or specific individual instances" (Kuhl 1991a).

Current models of categorization have become very sophisticated about the nature of the mental representations for perceptual categories (Kruschke 1991; Nosofsky 1986, 1987, 1992). The use of the term "prototype" as a stable context-free unit carries with it a theoretical commitment to a particular type of category membership that differs from the classical view of categories, as well as from recent exemplar or instance-based accounts. Thus, it is important to specify precisely which version of prototype theory a researcher has assumed.

Second, the findings, in some cases, are not consistent with the views that the authors attempt to support. For example, several recent studies have demonstrated that ratings of prototypicality change as a function of the context in which stimuli are presented (e.g. Miller & Volaitis 1989; Volaitis & Miller 1992). These findings demonstrate that phonetic prototypes are not stable or invariant across contexts. However, the original appeal of the prototype approach was that a single representation could serve as a summary for the entire category. One could posit different prototypes for different conditions or that prototypes are only used in categorization after a normalization process. However, the first assumption violates cognitive economy, one of the major reasons for proposing prototypes and the second is inconsistent with the findings we reported concerning the retention of talker-specific details.

The nonanalytic approach we have outlined in this chapter suggests an alternative perspective for dealing with problems such as the stimulus-specificity and nature of category representations and context-sensitive changes in category structure. The specificity issue is one in which the exemplar approach is quite explicit. It assumes that specific instances or perceptual episodes are stored in long-term memory. No effort is made to provide an abstraction or summary representation of perceptual categories based on measures of central tendency or idealized form. Admittedly, however, the approach has been vague about which attributes of a talker's voice, for example, are retained in memory. This is an important issue for future investigation.

With regard to the context sensitivity of category structures, we suggest that it is not the *prototypicality* of items that changes across contexts, but rather it is the *salience* of particular exemplars that is changing. In other words, the *stimulus* bias for particular items changes as a function of the context in which they are presented (Nosofsky 1991). This is consistent with recent exemplar-based models of categorization which suggest that changes in attention to stimulus dimensions alter the internal structure of perceptual categories (Kruschke 1991; Nosofsky, 1986, 1987). When stimulus dimensions are reweighted, the salience of category exemplars also changes to reflect the new distribution of attention.

In summary, we suggest that the traditional approach to speech perception has been somewhat misleading with regard to the nature of the perceptual operations that occur when listeners process spoken language. Variability may not be noise. Rather, it appears to be informative to perception. We have briefly sketched the results of several studies which have demonstrated the encoding and retention of talker-specific details in speech perception. We believe that these studies point to important new directions in speech perception research in which variability, rather than invariance, is regarded as an important problem for study. This approach to speech perception leads to the view that the perceptual categories in speech must be adaptive, dynamic and extremely flexible in order to

accommodate the changing stimulus environment that is one of the most distinctive characteristics of speech.

References

Barsalou, L.W. (1993). Flexibility structure and Linguistic Vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A.C. Collins, S.E. Gathercole, M.A. Conway, & P.E.M. Morris (Eds.) *Theories of Memories*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Brooks, L. (1978). Nonanalytic Concept Formation and Memory for Instances. In E. Rosch and B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Carney, A., Widin, G. & Viemeister, N. (1977). Noncategorical perception of stop consonants varying in VOT. *Journal of the Acoustical Society of America*, **62**, 961-970.

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

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M. & Gerstman, L.J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, **24**, 597-606.

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

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

Elman, J.L. & McClelland, J.L. (1986).. Exploiting Lawful Variability in the Speech Wave. *Invariance and Variability in Speech Processes*, Hillsdale, NJ: Erlbaum, pp. 360-380.

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

Fowler, C.A. (In Press). Listener-talker Attunements in Speech. In T. Tighe, B. Moore, and J. Santroch (Eds.), *Human Development and Communication Sciences*. Hillsdale, NJ: Erlbaum.

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

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

Goldinger, S.D. (1992). Words and Voices: Implicit and Explicit Memory for Spoken Words. *Research on Speech Perception Technical Report No. 7*, Indiana University, Bloomington, IN.

Goldinger, S.D., Pisoni, D.B. & Logan, J.S. (1991). On the Locus of Talker Variability Effects in Recall of Spoken Word Lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **17**, 152-162.

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

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

Jacoby, L.L. & Brooks, L.R. (1984). Nonanalytic Cognition: Memory, Perception, and Concept Learning. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, New York: Academic Press, pp. 1-47.

Jamieson, D. & Morosan, D. (1986). Training non-native speech contrast in adults: Acquisition of English /θ / - /∂ / contrasts by francophones. *Perception & 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.

Kakehi, K. (1992). Adaptability to Differences Between Talkers in Japanese Monosyllabic Perception. In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*, Tokyo, Japan: IOS Press, Inc.

Klatt, D.H. (1986). The Problem of Variability in Speech Recognition and in Models of Speech Perception. In J.S. Perkell and D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Erlbaum.

Klatt, D.H. (1979). Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access. *Journal of Phonetics*, **7**, 279-312.

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

Kolers, P.A. (1976). Pattern Analyzing Memory. *Science*, **191**, 1280-1281.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavioral Development*, **6**, 263-285.

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

Kuhl, P. K. (1991b). Speech prototypes: Studies on the nature, function, ontogeny and phylogeny of the "centers" of speech categories. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure.* (pp. 239-264). Tokyo: OHM Publishing Co., Limited.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B, (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, **255**, 606-608.

Kruschke, J. (1992). ALCOVE: An Exemplar-based connectionist model of category learning. *Psychological Review*, **90**, 22-44.

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 K. Salzinger & S. Salzinger (Eds.), *Research in verbal behavior and some neurological implications*. (pp. 79-98). New York: Academic Press.

Laver, J. & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K.R. Scherer and H. Giles (Eds.) *Social Markers in Speech*. Cambridge: Cambridge University Press, pp. 1-31.

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

Lawrence, D.H. (1950). Acquired distinctiveness in cues: II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology*, **40**, 175-188.

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

Lisker, L. (1978). Rapid vs. rabid: A catalogue of acoustic features that may cue distinction. *Status Report on Speech Research*, **SR-54**, 127-132.

Lively, S.E., Pisoni, D.B. & Logan, J.S. (1992). Some effects of training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura (Ed.) *Speech Perception, Production and Linguistic Structure, Tokyo: Ohmsha Publishing Co. Ltd*, pp. 175-196.

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

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1993). Training Japanese listeners to identify English /r/ & /l/: III. Listeners show long-term retention of new phonetic contrasts. *Research on Speech Perception Progress Report No. 18.* Bloomington, IN: Speech Research Laboratory, Indiana University.

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

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.

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

Medin, D.L. & Barsalou, L.W. (1987). Categorization processes and categorical perception. In S. Harnad (Ed.) *Categorical Perception: The Groundwork of Cognition*, New York: Cambridge University Press.

Miller, J. L. & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.

Mullennix, J.W. & Pisoni, D.B. (1990). Stimulus Variability and Processing Dependencies in Speech Perception. *Perception & Psychophysics*, **47**, 379-390.

Mullennix, J.W., Pisoni, D.B. & Martin, C.S. (1989). Some Effects of Talker Variability on Spoken Word Recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

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

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*, **15**, 700-708.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity and classification. *Cognitive Psychology,* **23**, 94-140.

Nosofsky, R.M. & Kruschke, J.K. (1992). Investigations of an exemplar-based connectionist model of category learning. *The Psychology of Learning and Motivation*, **28**, 207-250.

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

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

Palmeri, T.J., Goldinger, S.D. & Pisoni, D.B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning, Memory and Cognition, **19***, 1-20.

Peters, R.W. (1955). The Relative Intelligibility of Single-voice and Multiple-voice Messages Under Various Conditions of Noise. *Joint Project Report No. 56, U.S. Naval School of Aviation Medicine*, pp. 1-9. Pensacola, FL.

Pisoni, D.B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, 12-17 October 1992, Pp. 587-590.

Pisoni, D.B. (1992). Some Comments on Talker Normalization in Speech Perception. In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*, Tokyo, Japan: IOS Press, Inc.

Pisoni, D.B. (1990). Effects of Talker Variability on Speech Perception: Implications for Current Research and Theory. *Proceedings of 1990 International Conference on Spoken Language Processing*, Kobe, Japan, pp. 1399-1407.

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

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

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.

Roediger, H.L. (1990). Implicit Memory: Retention Without Remembering. *American Psychologist*, **45**, 1043-1056.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, **7**, 532-547.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

Schacter, D.L. (1992). Understanding Implicit Memory: A Cognitive Neuroscience Approach. *American Psychologist*, **47**, 559-569.

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

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

Stevens, K.N. (1971). Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds. *Proceedings of the Seventh International Congress of Phonetic Sciences*. The Hague: Mouton.

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

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. (1980). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas & J.L. Miller (Eds.), *Perspectives on the Study of Speech.* Hillsdale, NJ: Erlbaum.

Stevens, K.N., Manuel, S.Y., Shattuck-Hufnagel, S. & Liu, S. (1992). Implementation of a model for lexical access based on features. *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, 12-17 October 1992, Pp. 499-502.

Strange, W. & Dittmann, S. (1984). The effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, **32**, 131-145.

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

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

Sussman, H.M., McCaffrey, H.A. & Matthews, S.A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, **90**, 1309-1325.

Tulving, E. & Schacter, D.L. (1990). Priming and Human Memory Systems. *Science*, **247**, 301-306.

Volaitis, L. E. & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate in the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

Werker, J. & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, **37**, 35-44.

Werker, J. & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, **75**, 1866-1878.

Yamada, R. A. & Tohkura, Y. (1991). Perception of American English /r/ and /l/ by native speakers of Japanese. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure*, (pp. 155-174). Tokyo: OHM Publishing Co., Limited.

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Effects of Stimulus Variability on the Representation of Spoken Words in Memory[1]

Lynne C. Nygaard, Mitchell S. Sommers, and David B. Pisoni

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

[1]This research was supported by NIH Training Grant #DC-00012-13 and NIH Research Grant #DC-0111-16 to Indiana University. A portion of this research was presented at the 123rd meeting of the Acoustical Society of America, Salt Lake City, Utah (May, 1992).

Abstract

The present paper reports a series of experiments designed to investigate the effects of three sources of stimulus variability on the memory representations for spoken words. The aim was to determine if variability in speaking rate and overall amplitude have consequences for the encoding and processing of spoken words and if these consequences are comparable to those found for talker variability. A serial recall task was used to study the effects of changes in speaking rate, talker variability, and amplitude on the initial encoding, rehearsal, and recall of lists of spoken words. Presentation rate was manipulated to determine the time course and nature of processing. The results indicate that at fast presentation rates, variations in both speaking rate and talker characteristics incur a processing cost which influences the initial encoding and subsequent rehearsal of spoken words. At slower presentation rates, however, variation in talker results in improved recall in initial list positions while variation in speaking rate has no effect on recall performance. Amplitude variability had no effect on serial recall at any presentation rate. These results suggest that the encoding of stimulus variability due to changes in speaking rate, talker differences, and amplitude may be the result of distinct perceptual operations.

Effects of Stimulus Variability on the Representation of Spoken Words in Memory

One of the fundamental problems confronting theories of speech perception is how to characterize the listener's ability to extract consistent phonetic percepts from a highly variable acoustic signal. Factors such as phonetic context (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), linguistic stress (Klatt, 1976), utterance length (Klatt, 1976; Oller, 1973), vocal tract size and shape (Fant, 1973; Joos, 1948; Peterson & Barney, 1952), and speaking rate (Miller, 1981, 1987), for example, can all have profound effects on the acoustic realization of linguistic units. The consequences of these many sources of variability for speech perception is that phonetic segments do not necessarily have any invariant acoustic form (but see, Stevens & Blumstein, 1978; Kewley-Port, 1983). Rather, stable linguistic units that listeners report stem from an acoustic signal that may be anything but stable. The purpose of the present series of experiments was to examine in detail three of these sources of variability--changes in talker characteristics, changes in speaking rate, and changes in overall amplitude. Our aim was to assess the consequences of changes along each of these dimensions for the perceptual processing and subsequent representation of spoken words in human memory.

Traditionally, accounts of speech perception have characterized variation in the acoustic speech signal as a perceptual problem that perceivers must solve (Shankweiler, Strange, & Verbrugge, 1976). Listeners are thought to achieve consistent phonetic percepts given variation in talker and rate at least, through some kind of perceptual normalization process in which linguistic units are evaluated relative to the prevailing rate of speech (Miller, 1987; Miller & Liberman, 1979; Summerfield, 1981) or relative to characteristics of a talker's vocal tract (Joos, 1948; Ladefoged & Broadbent, 1957; Summerfield & Haggard, 1973). Implicit in this view of normalization is the assumption that the end product of perception is an idealized canonical linguistic unit. Variation is assumed to be stripped away to arrive at the prototypical representations that are used for further linguistic analysis. An excellent example of this view can be seen in the following quote:

> ...when we learn a new word we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears; for example, we do not remember the voice quality, speed of utterance, and other properties directly linked to the unique circumstances surrounding every utterance are discarded in the course of learning a new word. (Halle, 1985, page 101).

Despite this view of speech which strongly emphasizes its idealized abstract form, there has been considerable research on the problem of perceptual compensation in speech perception although many investigations have assumed that the final product of perceptual analysis is equivalent to the linguists' description of speech as a sequence of discrete segments.

Talker Variability

Consider first the perceptual consequences of variation in talker characteristics. Summerfield and Haggard (1973) as well as Mullennix, Pisoni, and Martin (1988) have shown that phoneme and word recognition performance is poorer when listeners are presented with words produced by multiple talkers compared to words produced by only a single talker. Likewise, using a Garner (1973) speeded classification task, Mullennix and Pisoni (1990) reported that subjects had difficulty ignoring irrelevant variation in a talker's voice when asked to classify words by initial phoneme. Taken together, these

findings suggest that variations due to changes in talker characteristics are time and resource demanding. Further, the processing of talker information appears to be inseparable from the processing of the phonetic content of a talker's utterance.

Additional research has suggested that talker variability can affect memory processes as well. At relatively fast presentation rates, Martin, Mullennix, Pisoni, and Summers (1989) and Goldinger, Pisoni, and Logan (1991) found that serial recall of spoken words was better in initial list positions when all the words in the list were produced by a single speaker compared to a condition in which each word was produced by a different speaker. Interestingly, at longer presentation rates, Goldinger et al. (1991) found that recall of multiple-talker lists was superior to recall of words from single-talker lists. These results suggest that at fast presentation rates, variation due to changes in the talker affects the initial encoding and subsequent rehearsal of items in the to-be-remembered lists. At slower presentation rates, on the other hand, listeners are able to fully process and encode each word along with the concomitant talker information. Consequently, listeners are able to use the additional redundant talker information as an aid in recall.

Further evidence that talker information is encoded and retained in a long-term memory store comes from a recent series of experiments conducted by Palmeri, Goldinger, and Pisoni (1992). Using a continuous recognition memory procedure, specific voice information was shown to be retained along with word information and these attributes were found to aid later recognition memory. The finding that subjects are able to use talker-specific information suggests that this source of variation may not be discarded or normalized in the process of speech perception, as widely assumed in the literature. Rather, variation in a talker's voice may become part of a rich and highly detailed representation of the speaker's utterance in long-term memory. The decrements in performance due to talker variability would then be due to the additional attention and resources necessary to encode information conveyed by a talker's voice.

The research investigating effects of talker variability on perceptual analysis and memory suggests that changes in talker characteristics may not be resolved by the listener by a normalization process per se. That is, it appears that listeners may not strip away information about a talker's voice from the phonetic content of an utterance as is traditionally assumed. Rather, the increased time, attention, and resources needed to compensate for changes in talker characteristics may be the result of the analysis and representation of talker characteristics as a rich source of information. Given this view of the processing of talker information, the purpose of the present set of experiments was two-fold. First, we wanted to replicate the earlier findings on the effects of talker variability on perceptual and memory processes involved in serial recall. Second, we wanted to extend the investigation to two additional sources of variability -- variability due to speaking rate and variability due to changes in overall amplitude.

Variation in Speaking Rate

An extensive body of research suggests that listeners are affected by changes in speaking rate. For example, Miller and Liberman (1979) found that changes in speaking rate can affect phonetic identification. In their experiment, listeners were presented with a synthetic /ba/-/wa/ continuum in which the stop-glide distinction was cued primarily by the duration of the formant transitions. By varying the duration of the steady-state portion of the syllable, Miller and Liberman observed a shift in identification boundaries toward longer values of transition duration as the overall duration of the syllable became longer. Thus, according to the authors, as the steady-state portion of the syllable became longer specifying a slower speaking rate, listeners adjusted their perceptual judgments accordingly.

Further research has shown that information about speaking rate preceding a phonetic contrast can affect perceptual judgments as well. Summerfield (1981) conducted a series of experiments to evaluate the effect of a precursor phrase varying in rate of articulation on the identification of voiced versus voiceless stop consonants. His results showed that phoneme identification boundaries shifted to shorter values of voice onset time as the articulation rate of the precursor phrase increased. Thus, listeners were apparently basing their classification of the initial stop consonants on information about the prevailing rate of speech in the precursor phrase.

Recently, Miller and her colleagues (Miller & Volaitis, 1989; Volaitis & Miller, 1991) have shown that phonetic category boundaries and accompanying phonetic category structures that rely on temporal information are quite sensitive to relative changes in rate of articulation. Their results indicate that changes in speaking rate affect the mapping of acoustic information onto the organization of phonetic category structures. Thus, listeners seem to compensate for changes in speaking rate not only by shifting or re-adjusting category boundaries, but also by re-organizing the entire structure of their phonetic categories.

Although it appears clear that listeners are sensitive to changes in rate of articulation, less attention has been paid to the processing consequences, in terms of attention, processing resources, and memory of this type of variability. At issue here is whether the observed changes in perceptual judgments are due to compensatory processes that require time and attention or whether adjustments to differences in speaking rate are automatic and cost-free in terms of analysis. Recently, Sommers, Nygaard, and Pisoni (1992) investigated this issue by presenting listeners with lists of words mixed in noise under two conditions--one in which all the words in the list were presented at a single speaking rate and one in which words in the list were presented at multiple speaking rates. The results were quite similar to those found earlier for talker variability (Mullennix et al., 1989). Listeners were better able to identify words mixed in noise if all the words were produced at a single speaking rate. Changes in speaking rate from word to word in a list apparently incurred some kind of processing cost that made identifying words more difficult. The conclusion is that if a compensatory process exists, it must demand the attention and processing resources of the listener.

Given the effects of overall speaking rate on the perceptual processing of temporal contrasts, the question arises whether rate variability would have effects similar to those found for talker variability in terms of perceptual encoding and memory representation as revealed in the serial recall task. That is, given that changes in speaking rate appear to incur processing costs that draw upon a limited pool of resources, is speaking rate encoded into long-term memory representations in the same manner as talker variability? By comparing performance at different presentation rates (e.g., Goldinger et al., 1991) in the serial recall task with lists of words varying in talker differences and speaking rate, our aim was to assess these sources of variability which have typically been assumed to be normalized or compensated for in the process of speech perception and spoken word recognition.

Amplitude Variability

In contrast to the demonstrated effects of variability due to changes in speaking rate and talker differences, little attention has been devoted to variability in the overall amplitude level of an utterance. A compensatory or normalization process has not been proposed for variation in overall amplitude because changes in overall amplitude (as opposed to relative amplitude) have not been found to affect phonetic judgments or perceptual processing of speech. Take, for example, the study conducted by Sommers, Nygaard, and Pisoni (1992) discussed earlier. In addition to comparing identification of words in noise in a multiple speaking rate context versus a single speaking rate context, we compared word identification in noise for lists with items presented at multiple amplitude

levels over a 30 dB range versus lists with items presented at a single amplitude level. In the latter case, there was no difference in identification performance for lists with changing amplitude levels as compared to lists with a constant amplitude level. These results suggest that variability in amplitude may not incur a processing cost in the analysis of spoken words.

In the present experiment, changes in overall amplitude were studied to assess the effect of a source of variability that is assumed to be irrelevant for segmental analysis on the perceptual processing and memory representation of spoken words. Our aim was to specifically address the nature of perceptual compensation and linguistic representation of spoken words by assessing the effect of amplitude variability on the memory processes involved in the serial recall task. Is it the case that any source of variability in the acoustic speech signal has consequences for the perceptual encoding and rehearsal of spoken words or rather, is it the case that only sources of variation that are phonetically relevant in a given task affect perception and memory of spoken words? That is, would variation along a dimension such as overall amplitude affect memory for serial order in the same manner as talker variability? If so, it would suggest that all aspects of spoken words are attention and resource demanding regardless of the effect on the perception of phonetic distinctions. If not, it would suggest either that only aspects of spoken words that are linguistically relevant affect processing, encoding, and rehearsal of spoken words or that the perceptual system adjusts for changes in overall amplitude level early and easily in the analysis of speech and consequently, these changes would have little effect on subsequent stages of processing.

To assess and contrast the effects of talker, speaking rate, and amplitude variability on list rehearsal in a serial recall task, we examined recall for spoken word lists varying along each of these dimensions. In the first experiment, we sought to replicate and confirm the effects of talker variability on the serial recall of spoken words at three different presentation rates. Recall of lists of words produced by a single talker was compared to recall of lists produced by multiple talkers. Words were presented at 100, 1000, or 4000 ms inter-word intervals. In addition, we extended the investigation to speaking rate. Recall of lists produced at multiple speaking rates was compared to lists produced at a single speaking rate at the same three presentation rates used for the voice manipulation. In the second experiment, we examined the effects of variation in overall amplitude. Recall of lists presented with multiple amplitude levels was compared to lists presented at a single amplitude level. Again, the same three presentation rates were used. This experimental design allowed us to compare the effects of variability due to talker with the effects of variability in speaking rate and overall amplitude on the serial recall of spoken words. In addition, using the presentation rate manipulation, we were interested in determining if variability due to speaking rate and variability due to overall amplitude, like talker information, aids serial recall at slower presentation rates.

EXPERIMENT 1

The first experiment was designed to replicate the results reported by Goldinger et al. (1991). As was noted, Goldinger et al. (1991) found that at relatively fast presentation rates (250 and 500 ms ISI), serial recall of words in initial list positions was poorer for lists produced by multiple talkers compared to recall of words from lists produced by a single talker. At moderate presentation rates (1000 and 2000 ms ISI), however, Goldinger et al. found that the advantage of single-talker lists over multiple talker lists disappeared. Words in initial list positions were recalled just as well in multiple-talker lists as in single-talker lists. Finally, at a relatively slow presentation rate of items in the list

(4000 ms ISI), a reversal of the initial effect was found. Listeners were better able to recall words in initial list positions from multiple-talker lists than from single-talker lists. Goldinger et al. interpreted these results as evidence that variation in talker from word to word in the serial recall task incurs a processing cost for the listener. At fast presentation rates, talker variability may interfere with encoding and rehearsal processes resulting in poorer performance for multiple-talker lists. At slow presentation rates, however, sufficient time and attentional resources are available for the changing talker information to be fully processed and encoded into memory. In this case, the to-be-remembered items are actually more distinctive because additional information about item and temporal order is present in memory.

We conducted this experiment for two reasons. First, we wanted to replicate and confirm Goldinger et al.'s (1992) results with a new set of stimulus materials as well as obtain a set of results with talker variability that could be used for comparison with the effects of speaking rate and amplitude variability. Although our replication is quite similar to Goldinger et al.'s previous experiments, two important differences should be noted. First, our stimulus set consisted of a set of monosyllabic words with a wider phonetic inventory and more varied syllable structure than the set used by Goldinger et al. Second, our presentation rates were slightly different than those used by Goldinger et al. (1991). Although both studies varied presentation rate by changing the inter-stimulus interval of words in the lists, our ISI's were 100, 1000, and 4000 ms while Goldinger et al. used 250, 500, 1000, 2000, and 4000 ms ISI's. Despite these slight differences, we predicted that at the fastest presentation rate (100 ms ISI), listeners would recall more words in initial list positions from single-talker lists than from multiple-talker lists. At the medium presentation rate (1000 ms ISI), we predicted that there would be no difference between multiple and single talker lists. Finally, at the slowest presentation rate (4000 ms ISI), we predicted there would be superior recall for items from the multiple-talker lists than from the single-talker lists.

Second, we wanted to extend our investigation of the effects of variability on serial recall to changes in speaking rate. Our methodology was identical to that used in the talker variability condition except that variations in articulation rate rather than talker characteristics were assessed. Consequently, serial recall of lists of words produced at a single speaking rate was compared to serial recall of lists produced at multiple speaking rates. Our prediction was that because speaking rate appears to affect phonetic judgments and also appears to require processing resources and attention, variation in speaking rate would produce results similar to those found for talker variability in the serial recall task. Again, the three presentation rates were used to determine the role of processing time on the perceptual encoding and representation of rate information.[2] At fast presentation rates, we expected that recall for words in initial list positions would be poorer for multiple-rate lists than for single-rate lists. In the 1000 ms ISI condition, recall performance was expected to be comparable for multiple and single speaking rate lists. Finally, at the slowest presentation rate, we expected that as variations in speaking rate were fully processed and encoded, recall performance for multiple speaking rate lists would be superior to single speaking rate lists.

[2]It should be noted that the term *rate* is being used here in two different contexts. Speaking rate refers to the speed and time course of the articulation of the actual stimulus items used in the serial recall task. Presentation rate refers to the how quickly items within a list are presented.

Method

Subjects

One hundred and eighty undergraduate students enrolled in introductory psychology courses at Indiana University served as subjects. They were given partial course credit for their participation. All subjects were native speakers of American English and reported no history of speech or hearing disorders at the time of testing.

Stimuli

The stimuli consisted of a set of 100 monosyllabic words drawn from phonetically balanced (PB) word lists. To obtain the database of words produced by different talkers at different speaking rates, words was embedded in the carrier phrase "Please say the word _____" for presentation to speakers. Ten speakers (6 male and 4 female) were asked to pronounce each phrase at three different speaking rates -- slow, medium, and fast -- for a total of 3000 words (100 words x 10 speakers x 3 rates). The words were digitized on-line at a 10 kHz sampling rate and subsequently edited from the carrier phrase for presentation. The mean root-mean-square amplitude of all stimulus tokens was equated using a signal processing package.

To ensure that variations in articulation rate were perceptually salient, rate judgments were collected for the complete set of words from a separate group of listeners. For each speaker's utterance, five subjects were asked to judge whether the words were produced at a fast, medium, or slow rate. Percent correct as defined by the percentage of times subjects chose a rate that corresponded to the intended rate of the talker was 83, 81, and 75 for slow, medium, and fast words respectively. In addition, durations of words produced at each rate by each speaker were measured. The durations for slow, medium, and fast words averaged across speakers were 903, 564, and 383 ms, respectively. Thus, the rate judgments and measured durations confirm that the stimulus materials included a wide range of articulation rates and that this variation was perceptually salient to listeners.

From this original set of 3000 words, 8 ten-word lists were constructed. In the multiple-talker condition, each word in a list was produced by a different talker. Likewise, in the multiple-speaking rate condition, words were selected from slow, medium, and fast items such that each word in a list was produced at a different speaking rate. In the single-talker and single-rate condition, all words were produced by one of two talkers (one male or one female) at a normal or medium speaking rate.

Procedure

Subjects were tested in groups of 6 or fewer in a quiet testing room. Stimuli were presented over matched and calibrated TDH-39 headphones at approximately 80 dB (SPL). A PDP-11/34 computer was used to present stimuli and to control the experimental procedure in real time. The digitized stimuli were reproduced using a 12-bit digital-to-analog converter and were low-pass filtered at 4.8 kHz.

During the experiment, subjects first heard a 500 ms, 1000 Hz warning tone to alert them that a list was about to be presented. Then, a list of ten words was presented at one of three rates by varying the inter-item interval, either 100, 1000, or 4000 ms between words. After each list had been presented, another warning tone sounded indicating the beginning of the recall period. Subjects had 60 seconds in which to recall the words in the list. A third tone signaled the end of the recall period.

Subjects were instructed to recall the words in the exact order in which they were presented and to write down their responses on answer sheets provided by the experimenter.

Talker, speaking rate, and presentation rate were all between-subjects variables. Of the 180 subjects, sixty were tested at each presentation rate. Of those sixty, twenty were tested in the multiple-talker, twenty in the multiple-speaking rate, and twenty in the single-talker/single-speaking rate conditions. The same words were heard by all subjects. The variables that changed across condition were the number of talkers, the number of speaking rates, and the presentation rate. In addition, for the single-talker/single-speaking rate condition, half of the subjects (10) received lists produced by a male speaker and half received lists produced by a female speaker to ensure that differences between conditions were not due to the idiosyncratic characteristics of a particular talker. As this factor was never found to significantly influence recall performance, it will not be included in subsequent discussion or analyses.

Results and Discussion

Subjects' responses were scored as correct if and only if the target word or phonetically equivalent spelling of the target word was recalled in the same serial position as the word presented in the list. Figure 1 shows the effects of talker variability on serial recall at the three presentation rates. Percent correct recall is plotted as a function of serial position for single versus multiple talker conditions. The top panel shows serial recall performance for multiple-talker and single-talker lists presented at the 100 ms ISI. The middle panel shows recall performance from the 1000 ms ISI and the bottom panel shows the results from the 4000 ms ISI. A three-way analysis of variance (ANOVA) with talker condition (multiple versus single), serial position (1-10), and presentation rate (100, 1000, 4000 ms ISI) as factors was conducted on the number of correct responses. As expected, a significant main effect of serial position was found showing reliable primacy and recency effects in recall [$F(9,1026) = 153.20$, $p < .001$]. Serial recall performance was better overall for words in initial and final list positions. In addition, a significant main effect of presentation rate was found [$F(2,114) = 39.21$, $p < .001$]. Recall performance increased overall as presentation rate decreased. Thus, serial recall performance was found to be best at the slowest presentation rate or the longest ISI. The interaction of serial position and presentation rate was also significant [$F(18, 1026) = 3.72$, $p < .001$] indicating that the shape of the serial position curve changes as recall performance improved at the slower presentation rates. No other main effects or interactions were found to be significant.

Insert Figure 1 about here

In order to investigate the effect of talker variability on serial recall performance, three separate three-way ANOVAS for the factors of talker and serial position were conducted for primacy (list position 1-3), middle (list positions 4-7), and recency (list positions 8-10) portions of the serial recall curve. For early, middle, and late list positions, significant main effects of serial position [early, $F(2, 228) = 118.50$, $p < .001$; middle, $F(3,342) = 17.53$, $p < .001$; late, $F(2,228) = 130.56$, $p < .001$] and presentation rate [early, $F(2,114) = 42.52$, $p < .001$; middle, $F(2,114) = 27.98$, $p < .001$; late, $F(2,114) = 6.11$, $p < .001$] were found reflecting primacy and recency effects in each serial position curve and improved recall performance as presentation rates became slower. In addition, these analyses revealed a significant interaction in each list position between serial position and presentation

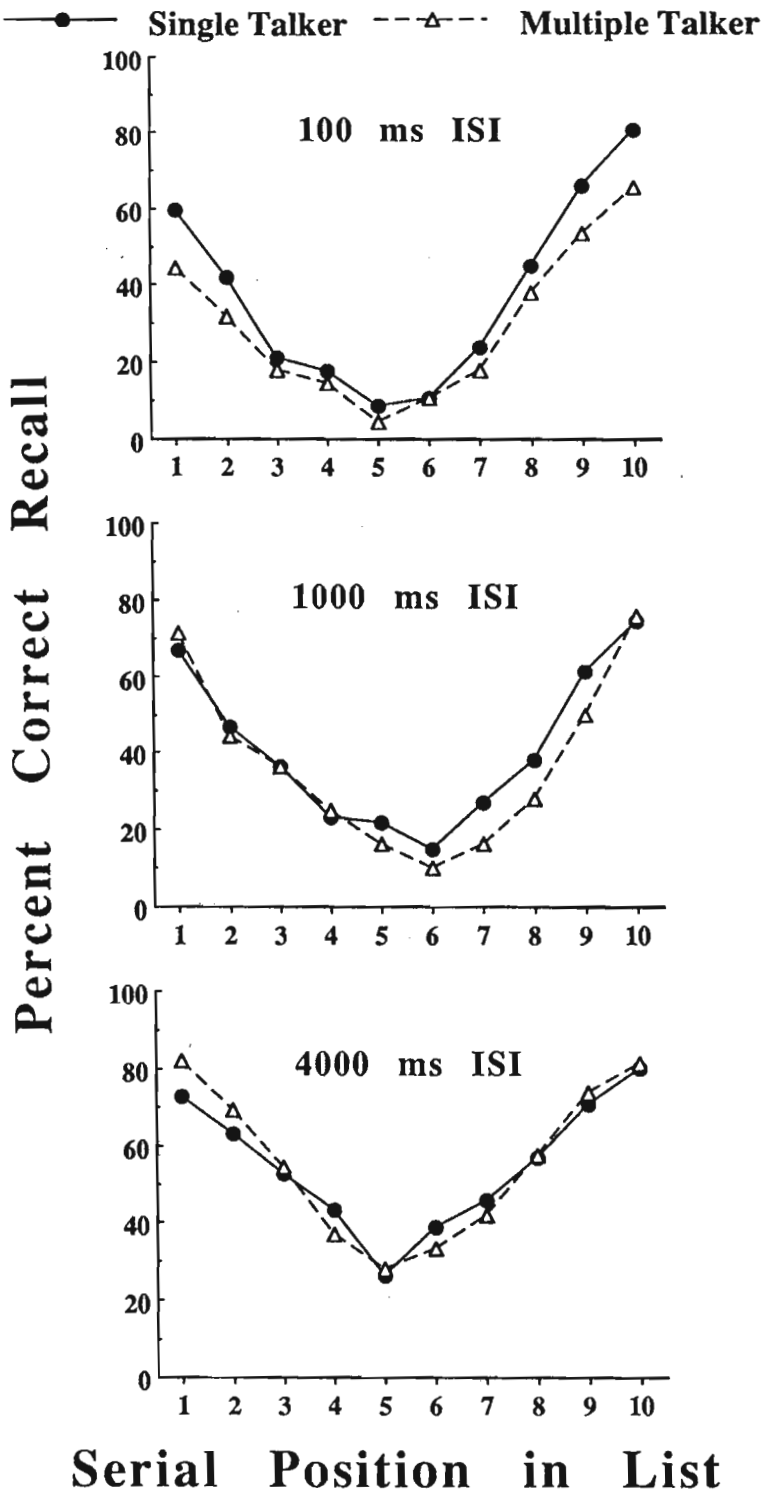


Figure 1. Mean percentage of correctly recalled words for both the single- and multiple-talker lists as a function of serial position and presentation rate.

rate [early, $F(4,228) = 2.45$, $p < .05$; $F(6, 342) = 2.84$, $p < .01$; late, $F(4.228) = 3.58$, $p < .05$] indicating that the serial position curve changed in shape as recall performance improved with slower presentation rates. Finally, a significant interaction was found between presentation rate and talker variability at early list positions [$F(2,114) = 2.92$, $p < .05$], but not for middle and late list positions. Recall of words from the early portions of multiple-talker lists was affected more by changes in presentation rate than recall of words from the single-talker lists.

Figure 2 shows the effects of variations in speaking rate on serial recall at the three presentation rates. Again, percent correct is plotted as a function of serial position. The top panel shows recall performance for multiple speaking rate and single speaking rate lists presented at the 100 ms ISI. The middle panel shows recall performance from the 1000 ms ISI condition and the bottom panel shows the results from the 4000 ms ISI. A three-way ANOVA with speaking rate (multiple versus single), serial position, and presentation rate as factors revealed significant main effects for presentation rate [$F(2,114) = 21.00$, $p < .001$)]; serial position [$F(9,1026) = 147.86$, $p < .001$]; and speaking rate [$F(1,114) = 3.99$, $p < .05$]. Recall performance was better overall at slow presentation rates than at fast presentation rates; recall was better at early and late list positions; and finally, recall was better overall for the single rate lists than the multiple rate lists. In addition, a significant three-way interaction [$F(18,1026) = 2.17$, $p < .005$] was found indicating that a larger difference was obtained between multiple- and single-rate lists at the fast presentation rates than at the slow presentation rate.

Insert Figure 2 about here

To further assess the effects of speaking rate on recall performance, three separate three-way ANOVAS were conducted for early (1-3), middle (4-7), and late (8-10) list positions. For early, middle, and late list positions, significant main effects of serial position [early, $F(2,228) = 95.05$, $p < .001$; middle, $F(3,342) = 16.76$, $p < .001$; late, $F(2,228) = 166.63$, $p < .001$] and presentation rate [early, $F(2,114) = 32.93$, $p < .001$; middle, $F(2,114) = 23.30$, $p < .001$; late, $F(2,114) = 3.06$, $p < .05$] were found reflecting primacy and recency effects in each serial position curve and improved recall performance as presentation rates became slower. In addition, these analyses revealed a significant interaction in each list position between serial position and presentation rate [early, $F(4,228) = 2.64$, $p < .04$; $F(6, 342) = 5.12$, $p < .001$; late, $F(4.228) = 5.07$, $p < .001$] indicating that the serial position curve became shallower as recall performance improved with slower presentation rates. The results revealed a significant two-way interaction [$F(2,114) = 3.08$, $p < .05$] between speaking rate and presentation rate, but only for early list positions.[3] This finding indicates that the difference between single- and multiple-rate lists was smaller at the slow presentation rate than at the fast presentation rates. A significant three-way interaction among speaking rate, serial position, and presentation rate was also found [$F(4,228) = 2.40$, $p < .05$].

The critical difference between the talker variability conditions and the rate variability conditions can be seen in the slowest presentation rate condition. Separate two-way ANOVAS for the talker condition and the rate condition conducted with responses just to items in early list positions (1-2) and just at the longer ISI revealed a significant main effect of talker [$F(1,38) = 3.98$, $p < .05$], but

[3]Although the effects of talker and rate do not result in significant main effects for items in the recency portion of the serial recall curve, it is clear that recency effects are present in our data. Because these recency effects appear to come and go in our data set (see also Goldinger et al., 1991), we will not consider them further in this paper. Further research will need to be done to precisely specify the nature and locus of these effects.

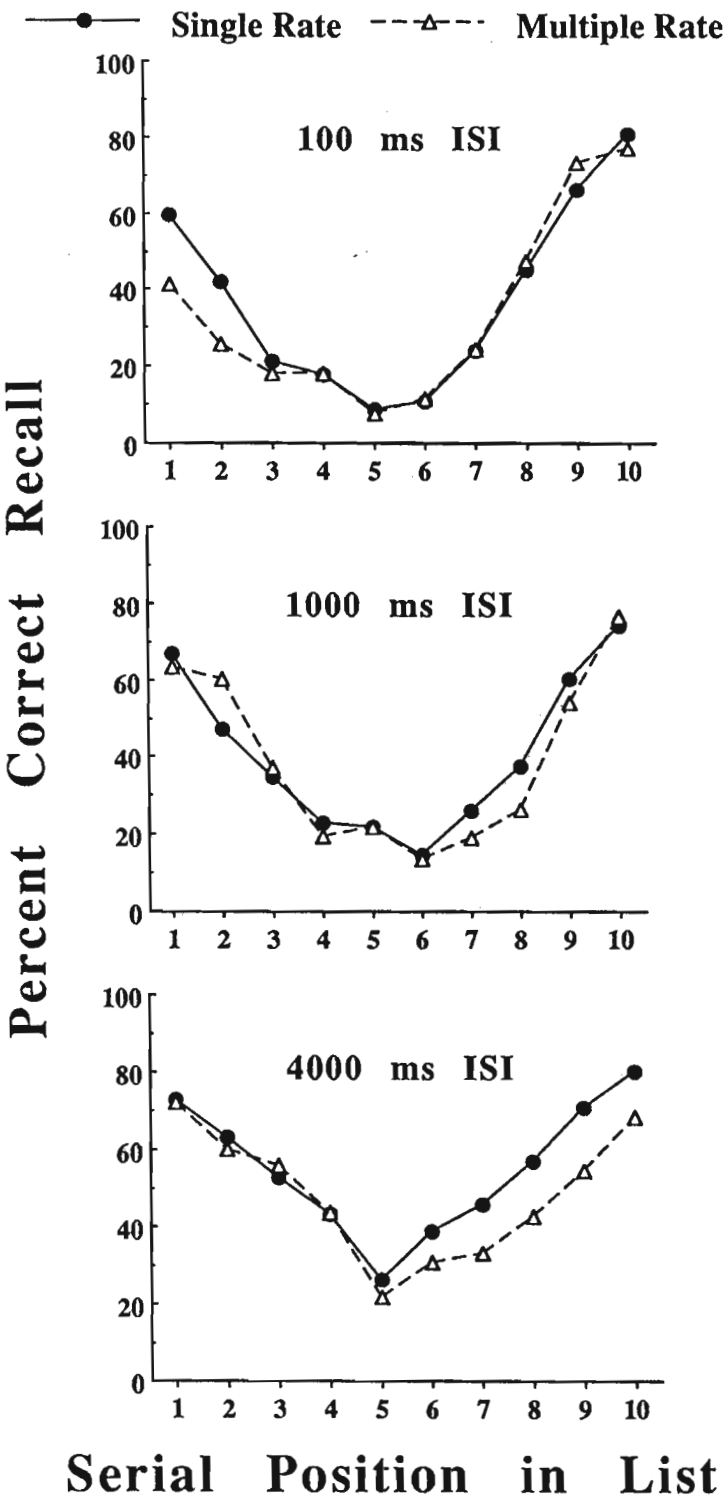


Figure 2. Mean percentage of correctly recalled words for both the single- and multiple-speaking rate lists as a function of serial position and presentation rate.

no significant main effect of rate. That is, these analyses revealed a difference between multiple-talker and single-talker conditions, but no difference between multiple-rate and single-rate conditions in initial list positions. Recall performance was better for multiple-talker lists than for single-talker lists at the slow presentation rate. However, a comparable benefit of variation in speaking rate was not found in the primacy portion of the curve. Thus, the two sources of variability do not have comparable effects on recall performance.

Turning first to the effects of talker variability, our findings replicate the results reported by Martin, Mullennix, Pisoni, and Summers (1989) and Goldinger et al. (1991). Further, given that the previous findings used words selected from the Modified Rhyme Test (MRT), the present results demonstrate reliable effects of talker variability with a new set of stimulus materials (PB words). At fast presentation rates, recall of items in early list positions is poorer for multiple talker lists than for single talker lists. Assuming that recall performance in the primacy portion of the serial recall function reflects the amount of processing and the efficiency of rehearsal processes needed to transfer items into long-term memory (Baddeley & Hitch, 1974), poorer recall of multiple-talker lists suggests that talker variability incurs a processing cost which somehow affects the efficient transfer of items into long-term memory. In the 1000 ms ISI condition, we found no difference between multiple-talker and single-talker lists. This result suggests that given additional time subjects are able to process and encode multiple-talker lists at least as well as single-talker lists. Interestingly, at the slowest presentation rate, multiple-talker lists displayed an advantage in recall performance. In this condition, words in early list positions were recalled better in multiple-talker lists than in single-talker lists. As Goldinger et al. (1991) have argued, this advantage in recall for multiple-talker lists at slow presentation rates suggests that talker information is retained in the long-term representation of items and appears to be used by subjects to aid in subsequent recall. This change from talker variability impairing recall performance at fast list presentation rates to aiding recall performance at slow presentation rates suggests that talker information may be integrated into subjects' representation of spoken words. It may be that it is this integration process that accounts for the increased time and resources needed to process lists of words produced by a variety of talkers.

The results from the multiple- and single-speaking rate conditions suggest a somewhat different picture. As for the talker variability conditions, at the fast presentation rate, serial recall of words in initial list positions was better for single-rate lists than for multiple-rate lists. Again, this finding suggests that variation in speaking rate incurs some kind of processing cost which affects the successful encoding and rehearsal of early list items, especially at fast presentation rates. At 1000 ms ISI, just as for talker variability, no difference was found in recall performance between multiple-speaking rate and single-speaking rate lists. With speaking rate as well, given sufficient time, subjects are able to process and encode multiple-rate lists as well as single-rate lists. At the 4000 ms ISI, however, a difference does emerge between the talker variability conditions and the speaking rate variability conditions. For variations in speaking rate, no benefit was found for multiple-speaking rate versus single-speaking rate lists at the slow presentation rate. This finding contrasts with recall performance at the slowest presentation rate with talker variability. In this case, talker information appeared to aid in serial recall because subjects presumably had sufficient time to make use of the distinctive attributes provided by each different voice. Information about speaking rate, in contrast, does not appear to provide any additional information at the slow presentation rate that subjects can use to aid in their serial recall.

Taken together, these findings suggest that both talker and rate variability require time and resources. However, even though both sources of variability impair performance at fast presentation

rates, and consequently, under more difficult encoding conditions, only talker information appears to be retained incidentally for later use as an additional redundant cue to temporal order in the serial recall task. The question remains whether the processing time and resources need to analyze spoken words with rate and talker variability have a common origin. We consider this issue in the general discussion.

EXPERIMENT 2

The results from the first experiment suggest that factors such as talker and speaking rate variability can affect the perceptual analysis and encoding of spoken words and these factors require processing time and resources. Talker and rate variability were both shown to impair recall performance under the more difficult encoding conditions imposed by the fast presentation rate in the serial recall task. The present experiment was designed to study variation in the overall amplitude level of spoken words. As mentioned previously, overall amplitude has not been shown to affect phonetic categorization and as such, is not assumed to require processing time and resources for perceptual analyses and encoding. By comparing the effects of amplitude variability with the effects of talker and rate variability, we sought to evaluate the role of phonetic relevance on the encoding, rehearsal, and transfer of spoken words into memory. If variations in overall amplitude do have an effect on recall performance in the serial recall task, then it would suggest that any variation in the speech signal can draw attention and therefore, processing resources away from the evaluation of the phonetic content of an utterance. If variability in overall amplitude does not affect recall performance, then it can be argued that only factors which directly affect the construction of the internal representations of spoken words compete for time and resources in the serial recall task.

Given the perceptual findings of Sommers et al. (1992) mentioned earlier, we hypothesized that variability in overall amplitude would have no affect on recall performance at any of the three presentation rates used. At the fast presentation rate, it was predicted that variation in amplitude would not require the additional perceptual analysis and encoding that would result in impaired recall of words in initial list positions. Likewise, at the medium and slow presentation rates, we expected no effect of amplitude variability. If amplitude variability does not affect perceptual encoding and rehearsal, then presumably it would not be preserved in memory and consequently, could not be used to aid in the serial recall task at the slow presentation rate.

Method

Subjects

One hundred and twenty undergraduate students enrolled in introductory psychology courses at Indiana University served as subjects. They were given partial course credit for their participation. All subjects were native speakers of American English and reported no history of speech or hearing disorders at the time of testing.

Stimuli

The stimuli consisted of the 100 medium rate words used in Experiment 1. To create different overall amplitude levels, words were digitally manipulated to create three different overall amplitudes using a signal processing software package. A maximum level was set to a specified value for each waveform and the remaining amplitude values in the digital file were then rescaled relative to that maximum. The three overall amplitude levels, 40, 50, and 60 dB, were chosen to be both perceptually salient and free of distortion. Thus, from the original 100 medium rate words used in Experiment 1,

300 new stimuli were created--100 words at each of the three amplitude levels. From this set of 300 words, 8 ten-word lists were constructed. In the multiple-amplitude condition, each consecutive word in the list was chosen from a different overall amplitude level. In the single-amplitude condition, each word in the list was chosen from the medium or 50 dB amplitude stimuli. Words varied only in amplitude; speaking rate and talker did not vary at all within or across lists. However, in the single-amplitude condition, all words in a given list were produced by one of two talkers (both female).

Procedure

The design and procedure was the same as in Experiment 1. Subjects were given a warning tone to signal the presentation of a list. Then, a list of ten words at one of three presentation rates (100, 1000, or 4000 ms ISI) was presented and subjects were asked to recall and write down the words from the list in the order in which they were presented.

Amplitude and presentation rate were between-subjects variables. Of the 120 subjects, forty were tested at each presentation rate. Of those forty, twenty were tested in the multiple-amplitude condition and twenty were tested in the single-amplitude condition. The same words were heard by all subjects, but the overall level of items varied depending on condition. In addition, for the single-amplitude condition, half of the subjects received lists produced by one female speaker and half of the subjects received lists produced by another female speaker. As for Experiment 1, talker was varied in this manner to control for idiosyncratic characteristics of individual talkers. Female speakers were used because the signal processing techniques used to create stimuli at different levels produced less distortion in the signal for female speakers than for male speakers and we did not want to confound variation in amplitude with variation in amount of distortion. Again, the talker's voice in this case did not influence recall performance, so it will not be included in subsequent discussion and analyses.

Results and Discussion

As in the first experiment, subjects' responses were scored as correct if and only if the target word or a phonetically equivalent spelling of the target word was recalled in the same serial position as the word presented in the list. Figure 3 shows the effects of variability in overall amplitude on serial recall at the three presentation rates. The top panel shows serial recall performance for multiple-amplitude versus single-amplitude lists at the 100 ms ISI. The middle panel shows recall performance from the 1000 ms ISI and the bottom panel shows the results from the 4000 ms ISI. A three-way ANOVA with amplitude condition (multiple versus single), serial position (1-10), and presentation rate (100, 1000, 4000 ms ISI) as factors was conducted on the number of correct responses. As expected, significant main effects of serial position [$F(9,1026) = 143.21$, $p < .001$] and presentation rate [$F(2,114) = 52.81$, $p < .001$] were found showing better recall performance in initial and final list positions and better recall performance overall as presentation rate decreased. Finally, no significant main effect of amplitude was found and none of the interactions were significant. Likewise, additional three-way ANOVAS conducted for items in early (1-3), middle (4-7), and late (8-10) list positions uncovered no significant main effects or interactions involving amplitude.

Insert Figure 3 about here

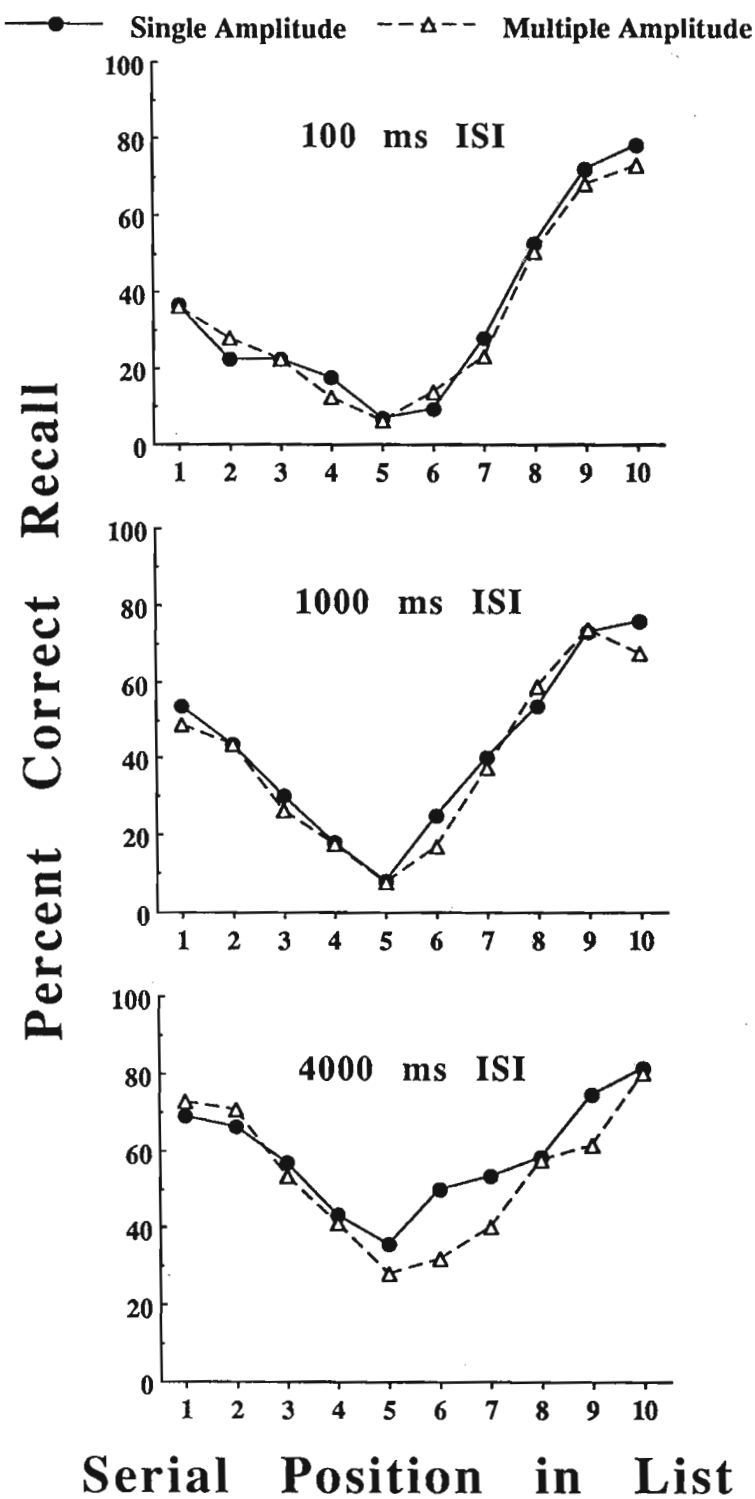


Figure 3. Mean percentage of correctly recalled words for both the single- and multiple-amplitude lists as a function of serial position and presentation rate.

These results suggest that variation in overall amplitude has no effect on the perceptual encoding and rehearsal processes involved in the serial recall task. Although null results such as these must be interpreted with caution, it appears that amplitude variability does not necessarily demand the processing time and resources that would have resulted in impaired recall of words from initial list positions. This finding replicates and extends the effects of amplitude variability on perceptual identification found by Sommers et al. (1992). Recall that in this study we found that variability in overall level did not impair subjects' ability to identify spoken words presented in noise. Our finding provides additional evidence that the introduction of amplitude variability into a set of spoken words does not engage time and resource demanding perceptual mechanisms.

The absence of an effect of amplitude variability has two implications for the nature of perceptual encoding and memory representation for spoken words. The first is that all variations in the speech signal are not created equal. That is, variability introduced by changes in overall amplitude level does not appear to require the same type of perceptual analyses as those involved in the resolution of variation in talker and speaking rate. The reason for this dissociation between amplitude variability and the other types of variability under consideration may be due to the phonetic relevance of each factor. As mentioned earlier, changes in talker characteristics and speaking rate can have profound effects on the spectral and temporal composition of a spoken words whereas changes in overall amplitude appear to preserve the acoustic underpinnings of essential phonetic contrasts. Thus, if important acoustic/phonetic information is preserved even with changes in overall amplitude, then there would be no need for increased perceptual resources to be marshaled for the resolution of amplitude variability.

The second implication is that even though the changes in amplitude from trial to trial in this serial recall task were readily apparent to listeners in the experiment, they did not use this source of variability to aid in subsequent serial recall. That is, amplitude variability, like rate variability, did not appear to be encoded incidentally into long-term memory as talker information routinely is. Although the reason for the differences in recall performance at the slow presentation rates between the talker variability condition and the rate and amplitude variability conditions is unclear, it does not appear to be directly related to the phonetic relevance of the source of variability in question. Both changes in speaking rate and in overall amplitude did not benefit listeners' recall performance in the slow presentation rate condition.

GENERAL DISCUSSION

The present set of experiments was designed to assess the effects of talker variability, speaking rate, and amplitude on the perceptual encoding, rehearsal, and memory processes involved in the recall of lists of spoken words. To that end, serial recall performance for lists with high stimulus variability was compared to serial recall of lists with low stimulus variability. In addition, presentation rate was manipulated to assess the time course of processing and encoding of talker, rate, and amplitude information.

In general, the results showed that each type of stimulus variability is analyzed and encoded in a distinct manner. The pattern of results across presentation rate conditions suggests that differences may exist in the way listeners analyze, rehearse and encode information about a talker's voice, speaking rate, and the overall amplitude of an utterance. At relatively fast presentation rates, variations in a talker's voice from trial to trial impaired the perceptual encoding and transfer of spoken words into long-term memory. Words from multiple-talker lists were not recalled as often as words

from single-talker lists. The same results were found for variability introduced by changes in speaking rate in the fast presentation rate condition. Words from multiple-speaking rate lists were not recalled as often as words from single-speaking rate lists suggesting that for speaking rate as well, variability along this dimension competed for limited pool processing resources. When processing demands were great at fast presentation rates, variability in talker characteristics and speaking rate interfered with the successful perceptual analysis and transfer of words into memory. These results may be contrasted with the third source of variability investigated in these experiments. Overall amplitude level did not impair serial recall performance in the fast presentation rate condition.

One explanation for the differential effects of talker and rate versus amplitude variability is that differences in the first two dimensions have profound ramifications for the resolution of the spectral and temporal properties of phonetic contrasts in speech while variation in amplitude does not. If this explanation is indeed true, then listeners may be quite sensitive only to variations in the speech signal that are phonetically relevant. That is, listeners may only attend to changes in the speech signal that affect the nature and structure of their phonetic categorization and subsequent word recognition. Another possibility is that the resolution of variations in overall level may occur very early and automatically in processing. Consequently, although listeners must compensate for changes in overall amplitude, this analysis may require very little time and very few processing resources. Thus, the absence of an effect of amplitude variability may simply reflect the ease of processing speech varying along this dimension.

Turning to the effects of stimulus variability on serial recall in the slow presentation rate condition, we see an entirely different pattern of results. In this case, for talker variability, our study replicates the findings of Goldinger et al. (1991) in which recall of words in multiple-talker lists was superior to recall of words in single-talker lists when listeners are given sufficient processing time to fully encode talker information. Apparently, the distinctive information provided by each of the voices associated with the words in the list allowed listeners to better remember the temporal order of the test items. Thus, it appears that subjects are not simply normalizing or "stripping away" variation in the speech signal due to talker-specific information as traditionally assumed, but rather listeners in both studies appear to be incidentally encoding distinctive talker information into a rich and highly-detailed memory representation of spoken words. These findings are consistent with recent research also showing that talker-specific information is retained in long-term memory and can be used not only to aid recognition memory (Palmeri et al., 1993), but also to facilitate the subsequent perceptual analysis of the phonetic content of a talker's novel utterance (Nygaard, Sommers, & Pisoni, 1992).

For variability in speaking rate, however, we did not replicate the effects found for talker variability. No advantage was found for multiple-speaking rate lists over single-speaking rate lists at the slow presentation rate. One reason that speaking rate may not be beneficial at slow presentation rates is simply that the two different sources of variability may be processed and encoded differently--at least in this type of task. For example, changes in speaking rate and talker characteristics have very different effects on the acoustic realization of spoken words. Further, they also have very different roles in terms of their properties and functions in speech perception. Talker characteristics are relatively permanent and can provide potentially important indexical information about a talker's gender, dialect, age, size, emotional state, and physical state (Laver & Trudgill, 1979). Speaking rate, in contrast, is often a more short-term phenomenon and may convey prosodic information, but generally at the level of the phrase or sentence and not at the level of the word. For these reasons, speaking rate may not necessarily be encoded in long-term memory representations in the same manner as talker-specific attributes for a task using isolated words. Rather, speaking rate may be used only to

assess the phonetic value of linguistic segments and then be discarded or ignored by the listener as irrelevant for this task. This interpretation implies that rate information in speech may at times be treated in a fundamentally different manner than talker information. Likewise, the lack of any effect of amplitude variability may simply be a consequence of the low information value of this variable in the encoding and rehearsal of spoken words in the serial recall task. Not only does overall amplitude provide little information to the listener, but it may not even be relevant for the perception of the phonetic content of a talker's utterance.

The present set of findings reveal that the representations of spoken words constructed during speech perception are much more robust and detailed than previously believed. Variation in the speech signal due to voice characteristics does not seem to be discarded or dissociated from the phonetic content of the signal when listeners develop long-term representations of spoken words. Information about a talker's voice affects memory for spoken words, and appears to be encoded in parallel with the more abstract phonetic information. This demonstrated relationship between the perception of spoken words and variability in the speech signal suggests a reconsideration of the traditional idealized, canonical representations that have been endorsed in speech perception research.

References

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The Psychology of learning and memory, vol. 8* (pp. 47-90). New York, Academic Press.

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

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.

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

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

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.

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.

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

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

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

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: Human Perception and Performance*, **15**, 676-684.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the psychology of language* (pp. 119-157). Hillsdale, NJ: Erlbaum.

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

Miller, J. L., & Volaitis, L. E. Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.

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

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1988). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1992). Speech perception as a talker-contingent process. Submitted to *Psychological Science*.

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.

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

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.

Shankweiler, D. P., Strange, W., & Verbrugge, R. R. (1976). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, knowing: Toward an ecological psychology* (pp. 315-346). New Jersey: Erlbaum.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1992). Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing, volume 1* (pp. 217-220). Edmonton, Canada: Priority Printing.

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.

Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1074-1095.

Summerfield, Q., & Haggard, M. P. (1973). Vocal tract normalisation as demonstrated by reaction times. *Report on Research in Progress in Speech Perception*, **2**, Belfast, Northern Ireland: Queen's University.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influences of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Training Japanese Listeners to Identify English /r/ and /l/: III. Long-term Retention of New Phonetic Categories[1]

Scott E. Lively, David B. Pisoni, Reiko A. Yamada[2], Yoh'ichi Tohkura[3] and Tsuneo Yamada[4]

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

[1]This work was supported, in part, by NIDCD Research Training Grant DC00012-14 and NIH Research Grant DC00111-16 to Indiana University in Bloomington, IN. Send correspondence to Scott E. Lively, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47401. The authors wish to thank John Logan for his helpful suggestions and discussions on this work.
[2]Also at ATR Human Information Processing Research Laboratories in Kyoto, Japan.
[3]Also at ATR Human Information Processing Research Laboratories in Kyoto, Japan.
[4]Also at Department of Behavioral Engineering Faculty of Human Sciences, Osaka University, Osaka, Japan.

Abstract

Monolingual speakers of Japanese were trained to identify English /r/ and /l/ using Logan, Lively and Pisoni's (1991) high variability, identification training procedure. Subjects improved from the pretest to the posttest and during training. Performance during training was found to vary as a function of talker and phonetic environment. Generalization accuracy to new words depended on the voice of the talker producing the /r/-/l/ contrast: Subjects were significantly more accurate when new words were produced by a familiar talker than when new words were produced by an unfamiliar talker. This difference could not be attributed to simple differences in intelligibility. Three months after the conclusion of training, subjects returned to the laboratory and were given the posttest and tests of generalization again. Performance was surprisingly good on each test: Accuracy decreased only 2% from the posttest given at the end of training to the posttest given three months later. Similarly, no significant decrease in accuracy was observed for the tests of generalization. The present results suggest that the high variability training paradigm encourages a long-term reorganization of listeners' phonetic perception. Changes in perception are brought about by changes in selective attention to cues that signal phonetic contrasts. Moreover, these modifications in attention appear to be retained over time, despite the fact that listeners are not exposed to this contrast in their native language environment.

Training Japanese Listeners to Identify English /r/ and /l/: III. Long-term Retention of New Phonetic Categories

Introduction

To what extent can short-term laboratory training experiments provide useful information to researchers working on theoretical issues in speech perception? This question has been the topic of a long-standing debate (see Lane, 1965; Pisoni et al., 1982; Strange, 1972; Strange & Jenkins, 1978; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). In the past, some researchers have argued quite strongly that it is difficult, if not impossible, to modify the speech perception abilities of adult listeners in a short period of time using simple laboratory training techniques. In a review chapter on the role of linguistic experience in speech perception, Strange and Jenkins concluded:

> "The research to date suggests that significant modification of phonetic perception is not easily obtained by simple laboratory training techniques. While changes did occur in some cases, the modification seemed to be rather narrowly restricted to just the dimensions on which the listeners were trained. There was little, if any, generalization to new stimulus materials and only limited generalization to new perceptual tasks (Strange & Jenkins, 1978, p. 153)."

More recently, Strange and Dittmann have maintained the same position with regard to the effects of laboratory training:

> "Thus, for adults learning a foreign language, modification of phonetic perception appears to be slow and effortful, and is characterized by considerable variability among individuals (Strange & Dittmann, 1984, p. 132)."

Other researchers, however, have argued that the failures of earlier training studies to show any changes in performance were primarily due to methodological problems (McClaskey, Pisoni, & Carrell, 1983; Pisoni, Aslin, Perey, & Hennessy, 1982). They suggest that the basic underlying perceptual mechanisms can, in fact, be modified selectively in the laboratory in a relatively short period of time, if the appropriate stimulus materials and experimental procedures are used. For example, Pisoni et al. (1982) and McClaskey et al. (1983) successfully trained listeners to label synthetic stop consonants that varied in voice onset time (VOT) into three perceptual categories in one training session lasting only about an hour. Furthermore, McClaskey et al. (1983) reported that training on one place of articulation ("ba - pa") transferred without additional training to a new place of articulation ("da - ta"). In contrast to the earlier work on VOT (Edman, Soli & Widin, 1978; Strange, 1972; Strange & Jenkins, 1978), these findings suggest that phonetic perception can be modified by simple laboratory training procedures and that robust generalization to new contrasts is possible. Moreover, these findings demonstrate that the perceptual mechanisms employed in processing VOT contrasts are flexible and can be reorganized with the appropriate stimulus input.

Recent studies on the perception of /r/ and /l/ by Japanese listeners provide an important perspective in this debate because they demonstrate with another phonetic contrast that simple and effective laboratory methods can be used to modify phonetic perception (Lively, Logan & Pisoni, 1991, in press; Logan, Lively & Pisoni, 1991; Pisoni, Lively, & Logan, 1993). Studying the

acquisition of the /r/-/l/ contrast is of special interest because Strange and Dittmann (1984) have stated explicitly that contrasts based on spectral cues, such as /r/-/l/, may be much more difficult to acquire than contrasts based on temporal cues, such as VOT. Thus, these experiments directly address one of the more difficult test cases for modifying phonetic perception using laboratory-based training procedures.

The present study was carried out to examine the effectiveness of Logan et al.'s high variability identification training procedure with a large number of monolingual speakers of Japanese. In Logan et al.'s original experiment, only subjects who had received some intensive English training were tested. Subjects in the present experiment, in contrast, represent a relatively large and homogeneous sample. We were also interested in assessing the long-term effects of training on listeners' phonetic perception. In addition to assessing changes in perception immediately after training, listeners were also retested three months later to assess retention. These results provide an important test of the robustness of our training procedure and shed some light on the nature of the perceptual changes that occur when nonnative speakers acquire a new phonetic contrast.

Because Logan et al.'s (1991) findings serve as the primary motivation for the present experiment, we review the assumptions, methodology, results and limitations of their earlier study in detail. Logan et al. used a pretest-posttest design and tested generalization to new words and a new talker. Subjects responded in a two-alternative forced-choice identification task throughout the experiment. Immediate feedback was given to subjects only during the training phase. The minimal uncertainty of the two-alternative forced choice procedure, combined with the use of immediate feedback, was assumed to promote the formation of robust new phonetic categories (Jamieson & Morosan, 1986, 1989). The identification training procedure avoided the attentional demands of discrimination training by encouraging subjects to group similar objects into the same category and different objects into different categories (Lane, 1965, 1969). Discrimination training, in contrast, requires listeners' to attend to small within-category differences and does not encourage robust category formation (Carney, Widin, & Viemeister, 1977; Liberman, Harris, Kinney, & Lane, 1961; Pisoni, 1973; Werker & Logan, 1985; Werker & Tees, 1984). During training, listeners heard tokens from five talkers who produced English words with /r/ and /l/ in five phonetic environments. The high variability of the stimulus set was assumed to be important for developing perceptual constancy (Kuhl, 1983) and was thought to provide a broad base for generalization to new items and new talkers (Posner & Keele, 1968, 1970).

Several aspects of Logan et al.'s results motivated the present investigation. First, significant increases in identification accuracy were observed from the pretest to the posttest. It should be noted that Logan et al. used the same words in the pretest and posttest that Strange and Dittmann (1984) used in their earlier study. In Strange and Dittmann's study, Japanese subjects were trained in an AX fixed-standard discrimination paradigm. Although listeners improved during training in their ability to *discriminate* synthetically produced tokens that contrasted /r/ and /l/, they failed to find any change when generalization was tested with naturally produced speech. Logan et al.'s success in training subjects to perceive /r/ and /l/ in English words and Strange and Dittmann's failure to find any changes in pretest-posttest performance suggest that the high variability identification paradigm using natural speech was more effective than the low variability discrimination paradigm using synthetic speech. If this methodological difference is the major factor responsible for the differences between the two studies, then the outcome needs to be replicated in order to firmly establish the important role that high stimulus variability plays in studies of perceptual learning with speech stimuli (Jenkins, 1979). There

is some precedence for this in studies of visual perception, but the role of variability in speech perception has not been a topic of great interest until recently.

Second, significant increases in accuracy and decreases in response time were obtained during training. Listeners' accuracy increased about 5% and their response latencies decreased approximately 125 ms for /r/s and /l/s in final singleton position and final consonant clusters over the course of the three week training period. Thus, the high variability of the training set combined with the identification procedure appears to have been effective in modifying nonnative listeners' phonetic perception in a short period of time. However, given that the improvements were modest and that subjects were relatively experienced with English and were living in an English-speaking environment at the time of testing, we were interested in testing a monolingual group of subjects whose initial performance was poor in order to generalize the original findings to a different population of listeners.

Third, performance varied significantly as a function of phonetic environment. Tokens that contained /r/ and /l/ in final singleton position were identified most accurately, while /r/'s and /l/'s in initial consonant clusters were identified least accurately. Several other studies have reported a similar pattern of findings (Gillette, 1980; Goto, 1971; Lively et al., in press; Mochizuki, 1981; Sheldon & Strange, 1982; Strange & Dittmann, 1984). Durational and coarticulatory cues have been proposed as possible factors responsible for listeners' differential sensitivity to /r/ and /l/ in different phonetic environments (see Dissosway-Huff, Port, & Pisoni, 1982; Henly & Sheldon, 1986). The durations of /r/ and /l/ are longest in final singleton position and shortest in initial consonant clusters (Lehiste, 1960), paralleling the perceptual findings. Furthermore, /r/ and /l/ in final position tend to color the preceding vowel, while /r/ and /l/ in initial consonant clusters may be coarticulated with preceding consonants and may not reach their steady-state targets (Sheldon & Strange, 1982).[5]

Fourth, Logan et al. found that performance varied as a function of talker. Tokens produced by one of the female talkers used during training were identified significantly more accurately than tokens produced by the other talkers. This finding was surprising because all of the stimulus materials were pretested with native speakers of English and all talkers were found to be equally intelligible. One of the goals of the present study was to replicate the finding of talker-specific effects in perceptual learning in order to determine if reliable differences do exist among our training talkers for monolingual Japanese listeners learning English /r/ and /l/.

Finally, Logan et al. found only a marginal difference in identification accuracy in the test of generalization in which subjects had to respond to novel tokens produced by an old and new talker (old talker: 83.7%, new talker: 79.5%). Although the results of the test of generalization were somewhat ambiguous, the methodology of this test is important in establishing the robustness of the training paradigm. In order for a training paradigm to be considered robust, we believe that it is important to demonstrate that the effects of training extend well beyond the tokens listeners were trained on:

---

[5] The similarity of the nonnative contrast to phonetic distinctions found in the native language as well as the phonotactic constraints in the native language may also contribute to the difficulty of different phonetic environments (Best, 1993; Best & Strange, 1992; Flege, 1989, 1990; Flege & Wang, 1989; Mann, 1986; Polka, 1991, 1992). Japanese has an /r/ that is similar to English /d/ or /t/. It is produced either as a stop consonant or as a flap, depending on the vowel environment (Price, 1981; Yamada & Tohkura, 1992). In terms of phonotactic constraints, Japanese does not generally allow consonant clusters. Thus, in learning to perceive English /r/ and /l/ Japanese listeners have two disadvantages: First, they are being asked to discriminate between sounds that are not contrastive in their native language and are not similar to any known contrast in their language. Second, the /r/'s and /l/'s in initial consonant clusters occur in an unfamiliar phonotactic construction.

Listeners should demonstrate accurate performance with new words produced by new talkers. In the present investigation, we unconfounded possible differences in baseline intelligibility among our talkers from true generalization by pretesting the tokens used in the generalization tests with untrained Japanese listeners. If we can demonstrate that the talkers used in generalization are equally intelligible, *a priori*, then we can rule out intelligibility as the source of the differences observed during generalization (see Logan et al., in press; Pruitt, in press).

On the basis of the findings reviewed above, Logan et al. argued that the minimal uncertainty of the closed-set identification task, combined with the high variability of the stimulus items, promoted the acquisition of the English phonemes /r/ and /l/ by Japanese listeners. Their results demonstrated robust and generalizable category acquisition: Subjects responded accurately to new words produced by familiar and unfamiliar talkers. These findings suggest that simple, laboratory-based training procedures can be effective in modifying listeners' phonetic perception in a short period of time. However, these changes appear to be contingent upon the voice of the talker producing the contrast and the phonetic environment in which it occurs.

Several limitations of Logan et al.'s study can be identified which attenuate its impact and the overall significance of their conclusions (Logan et al., in press; Pruitt, in press). First, Logan et al. tested and trained only six subjects. Thus, some of the effects that they observed may have been due to the small sample size used in the original study. Second, only three listeners participated in the tests of generalization to new tokens and to a new talker. This severely limited the possibility of obtaining significant differences between the two talkers used in the generalization test (see however, Lively et al., in press). Finally, Logan et al. used subjects who had been living in the United States for several months and presumably had received exposure to English outside of the laboratory.

Because of the theoretical importance of Logan et al's results to issues in perceptual learning and development and the limitations of the original study noted above, we decided to carry out a complete replication and extension of their training study with larger number of monolingual Japanese listeners. We predicted that subjects would improve in their identification accuracy from the pretest to the posttest and during training. These findings would confirm the effectiveness of the training procedure. But more importantly, a replication would also establish once again the importance of methodological factors in speech perception studies and demonstrate that adult listeners have very flexible perceptual capabilities that may often be obscured by specific experimental procedures (Jenkins, 1979).

In addition to these considerations, we were also interested in testing subjects several months after the completion of training. If the changes observed during training are short-term, then listeners' performance should return to pretest levels in the absence of any further training or feedback in their environment. This outcome might be expected because subjects in the present study were living in a monolingual Japanese speaking environment and would not have much exposure to spoken English outside of the laboratory. However, if our training procedure encourages long-term changes in phonetic perception and reorganization of the listeners' phonetic categories, then we would predict very little change in performance from the posttest at the end of training to a test given three months later. Thus, the results of the tests given after three months without any further training provide a way to measure the retention of the changes that we observed during training and generalization.

Method

Subjects

The subjects were 19 native speakers of Japanese living in Kyoto, Japan . All of the subjects reported that they were monolingual speakers of Japanese. No subjects reported any history of a speech or hearing disorder at the time of testing.

Stimuli

The stimulus materials were identical to those employed by Logan et al. (1991). A computerized database containing approximately 20,000 words (Webster's Seventh Collegiate Dictionary, 1967) was searched to locate all minimal pairs contrasting /r/ and /l/. A total of 207 minimal pairs were found. These words contrasted /r/ and /l/ in word-initial and final positions, in singleton and cluster environments, and in intervocalic position. Six talkers, four male and two female, recorded the words in an IAC sound-attenuated booth using an Electro-Voice D054 microphone. Talkers were given no special instructions concerning pronunciation of the words, which were presented individually in random order on a CRT monitor located inside the recording booth. The words were low-pass filtered at 4.8 kHz and digitized at 10 kHz using a 12-bit analog-to-digital converter at Indiana University. The digitized waveform files were edited and equated for RMS amplitude using a specialized signal processing package. Files were then digitally transferred to ATR laboratories, where they were upsampled at a sampling rate of 44.1 kHz and rescaled to a resolution of 16 bits.

The stimuli were originally pretested at Indiana University with a separate group of native speakers of English to assess their intelligibility. An identification task was used in which listeners typed their response on a computer terminal after hearing each stimulus. The criteria for including a word in the experiment was that it have no more than a 15% error rate across all talkers and that no errors were due to misperception of /r/ or /l/.[6] After pretesting, a set of 136 words (68 minimal pairs -- 12 initial singleton pairs, 25 initial cluster pairs, 5 intervocalic pairs, 15 final singleton pairs, and 11 final cluster pairs) from five talkers was selected for use in the training phase of the experiment.

The stimuli used in the tests of generalization were also identical to those used by Logan et al. (1991). Two sets of tokens were recorded. The first set consisted of 95 novel words produced by a new male native speaker of English. In this set of stimuli, 38 words had /r/ or /l/ in initial singleton position, 29 had /r/ or /l/ in initial consonant clusters, 17 had /r/ or /l/ in final singleton position, and 11 had /r/ or /l/ in final consonant clusters. The second set of 97 new items was produced by one of the female training talkers. In this set of stimuli, 37 words had /r/ or /l/ in initial position, 32 had /r/ or /l/ in initial consonant clusters, 15 had /r/ or /l/ in final singleton position, and 13 had /r/ or /l/ in final consonant clusters.

Finally, the 24 minimal pairs used by Strange and Dittmann (1984) in the pretest-posttest phase of their experiment were recorded by a new male talker. Sixteen minimal pairs contrasted /r/ and /l/ in one of four phonetic environments (initial singleton, initial consonant cluster, intervocalic, final singleton). The remaining eight pairs contrasted phonemes other than /r/ and /l/. These items were processed in the same way as the other stimuli used in the present experiment.

---

[6] These data were reported earlier by Logan et al. (1991).

Procedure

The experimental design employed a pretest-posttest procedure closely modeled after the methods used by Strange and Dittmann (1984) and Logan et al. (1991). In this design, the effects of training were assessed by comparing performance on a pretest and a posttest administered before and after a three-week training period. In addition, in this study, we also assessed performance three months after the conclusion of training. All testing and training were carried out at ATR Auditory and Visual Perception Laboratories in Kyoto, Japan and were administered individually in a quiet sound-treated room. Subjects sat at a cubicle that was equipped with a desk, a keyboard, and a CRT monitor. Stimuli were binaurally presented over headphones (STAX-SR-Lambda Signature) at a comfortable listening level. Presentation of stimuli, feedback, and collection of responses were under control of a laboratory computer (NeXT cube). During training and tests of generalization, both identification responses and latencies were collected. Latencies were measured from the onset of the stimulus presentation to the subject's response. Feedback was given only during the training phase.

Before training began, subjects were given a pretest. This test consisted of 24 minimal pairs of words that were recorded onto digital audio tape using a DAT recorder (SONY DTC-1500ES). Two randomizations of the 48 words were recorded for a total of 96 trials. On each trial of the pretest, subjects were presented with an isolated word and were required to identify the stimulus by circling their response in an answer booklet. The same test items were presented after training (Posttest Phase) and again three months after the conclusion of training.[7] The pretest-posttest required approximately 20 minutes to complete.

The training phase also used a two-alternative identification task. On each trial, the two members of a minimal pair contrasting /r/ and /l/ were displayed in the lower left and right corners of the CRT for 500 ms prior to stimulus presentation. Subjects were then auditorally presented with a word over their headphones. Responses were made by pressing a button on the keyboard: Subjects identified stimuli corresponding to words on the left side of the CRT by pressing "1" and words on the right side of the screen by pressing "2." The position of the word containing /r/ varied randomly from trial to trial. On half of the trials, the word containing /r/ appeared on the left side of the CRT; on the remaining trials, the word containing /r/ was on the right side. Listeners had a maximum of 10 seconds to respond. If no response was made, the trial was considered an error.

Feedback was given on each trial during the training phase. If the listener responded correctly, a chime sounded and the next trial was presented two seconds later. If the subject made an error, a buzzer sounded and the stimulus word was repeated. Responses were made during the repetition and repetitions continued until the listener made a correct response. Correct responses made during repetitions were not included in the analysis of the subject's overall accuracy score. Subjects were also shown a graphical representation of a coin on the CRT every time they made three correct responses. At the end of a training session, listeners were paid an additional bonus based on the level of their performance.

Stimuli from a set of 68 minimal pairs were each presented twice during a training session, yielding a total of 272 trials in each session. During each training session, stimuli from only one talker were presented. Subjects cycled through the set of five talkers used during training three times

[7]The pretest-posttest procedure was under computer control for the three-month follow-up test. On each trial of the follow-up test, subjects saw each member of a minimal pair presented on a CRT monitor and responded by pressing a button that corresponded to the word that they heard over their headphones.

for a total of fifteen training sessions. Subjects were tested individually during training. Each session lasted approximately 40 minutes.

After the posttest, subjects were tested again to assess the degree to which training generalized to novel stimuli. The first test of generalization consisted of 95 novel words from minimal pairs contrasting /r/ and /l/ produced by a new talker (i.e., a talker not used in either the pretest/posttest phase or the training phase). A second test of generalization consisted of 97 novel words from minimal pairs contrasting /r/ and /l/ produced by Talker 4, who subjects had heard during training. The test stimuli consisted of new words that the subjects had not heard before. In both tests of generalization, the task was identical to the procedure used during training except that subjects did not receive any feedback. The tests of generalization were also administered individually. Both tests of generalization were also repeated three months after the conclusion of training.

Results

Pretest-Posttest

Mean accuracy scores for each subject from each test were submitted to an analysis of variance (ANOVA). Phonetic environment and pretest-versus-posttest were within-subjects variables. Post-hoc tests were conducted using Tukey's HSD procedure.

Insert Figure 1 about here.

The upper panel of Figure 1 displays accuracy in the pretest and posttest as a function of phonetic environment. Subjects improved significantly in their ability to identify /r/ and /l/ from the pretest to the posttest [pretest: 65%, posttest: 77%, $F(1,18)=92.99, p<.01$]. Accuracy varied widely as a function of phonetic environment [$F(3,54)=37.15, p<.01$]: /r/'s and /l/'s in final singleton position were identified significantly more accurately than /r/'s and /l/'s in initial singleton position and initial consonant clusters ($p<.05$). Target phonemes in initial singleton position were identified more accurately than phonemes in initial consonant clusters ($p<.05$).

Training

Separate ANOVAs were conducted on subjects' mean accuracy and response latency scores from each day of training. Week of training, talker, and phonetic environment were treated as within-subjects variables.

Insert Figure 2 about here.

The main effects for week of training on response accuracy and latency reveal the effects of the training procedures. As shown in Figure 2, subjects' responses became significantly more accurate and faster during training [$F_{pc}(2, 36)=31.67, p<.01$; $F_{rt}(2,36)=23.57, p<.01$] . Increases in accuracy were localized between Weeks 1 and 2 of training, although response latencies continued to decrease significantly during all three weeks of training.

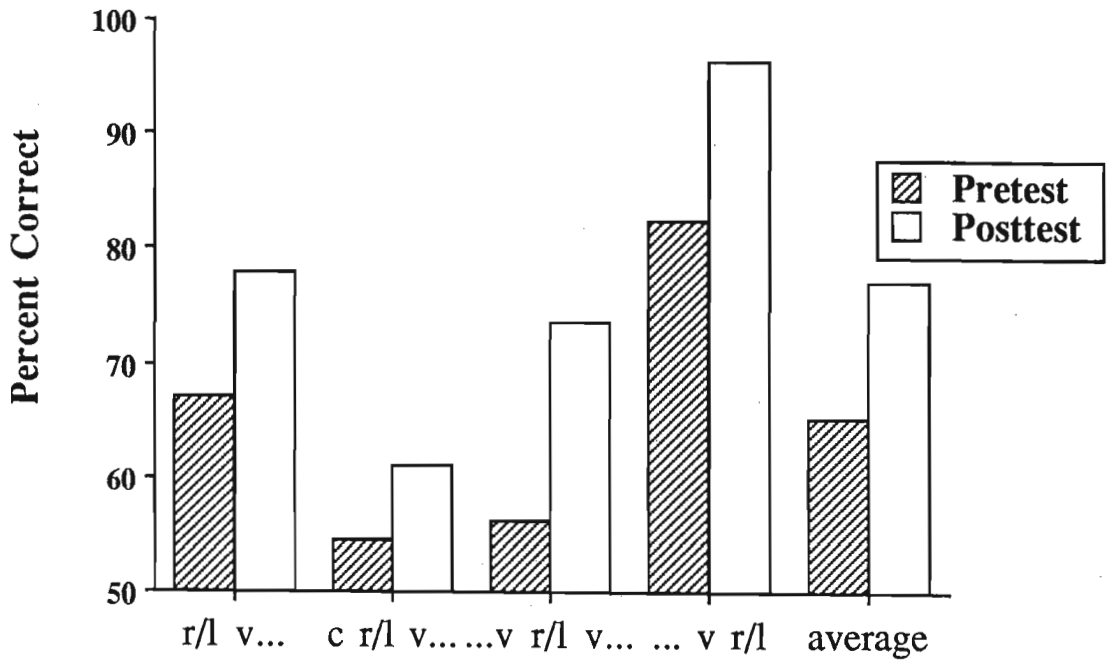


Figure 1. This figure shows subjects' accuracy in the pretest and posttest as a function of phonetic environment.

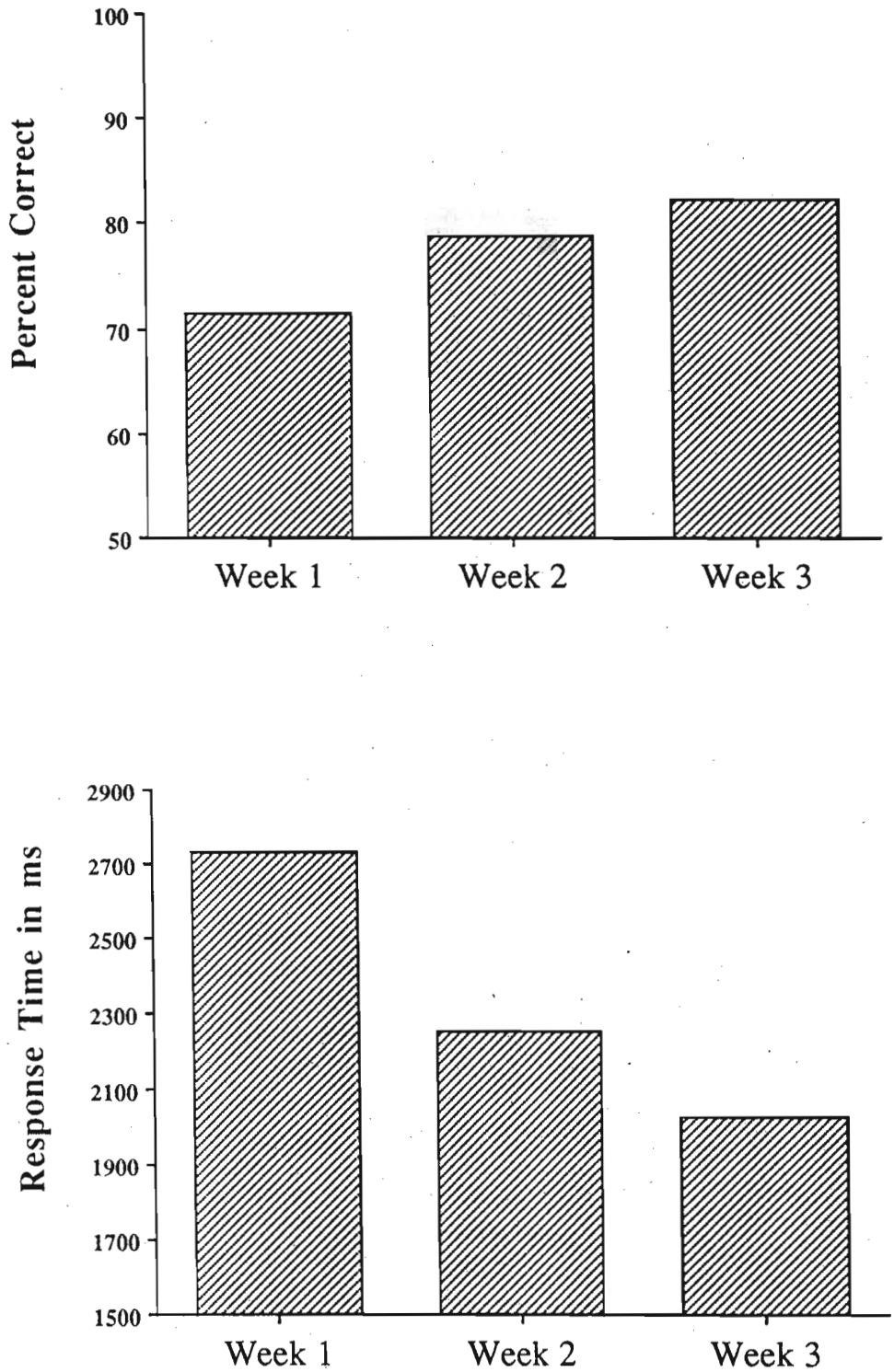


Figure 2. The top panel shows subjects accuracy as a function of week of training. The lower panel shows response latencies.

Insert Figure 3 about here.

Figure 3 shows the effects of talker in response accuracy and latency. The effects of talker variability obtained earlier by Logan et al. (1991) and Lively et al. (in press) were also replicated in the present experiment [$F_{pc}(4,72)=32.41$, $p<.01$, $F_{rt}(4,72)=3.77$, $p<.01$]. Tukey's HSD tests revealed that subjects responded more accurately to Talker 4 than to any other talker used in training, except Talker 5. Subjects also responded more accurately to Talker 5 than to Talker 1 and Talker 2. Responses to stimulus tokens produced by Talker 4 and Talker 5 were significantly faster than responses to items produced by Talker 1.

Insert Figure 4 about here.

A main effect for phonetic environment was observed in the accuracy and latency analyses [$F_{pc}(4,72)=153.60$, $p<.01$, $F_{rt}(4,72)=22.22$, $p<.01$]. As shown in Figure 4, subjects' responses to targets in final singleton position were faster and more accurate than to targets in all other positions. /r/'s and /l/'s in final consonant clusters were identified faster and more accurately than /r/'s and /l/'s in initial consonant clusters and intervocalic position. Finally, targets in initial singleton position were responded to more accurately than targets in initial consonant clusters. However, responses were faster to targets in initial consonant clusters than to phonemes in initial singleton position.

Insert Figure 5 about here.

In addition to the main effects for week, talker, and phonetic environment, several interactions were also significant in the analyses of the accuracy and latency data. First, an interaction between talker and week was observed in the accuracy scores [$F(8,144)=4.46$, $p<.01$]. These findings are displayed in Figure 5. Performance was significantly higher during Week 3 of training than during Week 1 for all talkers. However, significant increases were also obtained between Weeks 1 and 2 for Talkers 1, 2 and 3.

Insert Figure 6 about here.

Second, a significant interaction between week of training and phonetic environment was also observed in the accuracy scores [$F_{pc}(16,288)=8.22$, $p<.01$]. These data are displayed in Figure 6. Accuracy increased in all phonetic environments from Week 1 to Week 2. However, significant increases were obtained only for /r/'s and /l/'s in initial singleton, initial consonant clusters, and intervocalic positions from Week 2 to Week 3.

Insert Figure 7 about here.

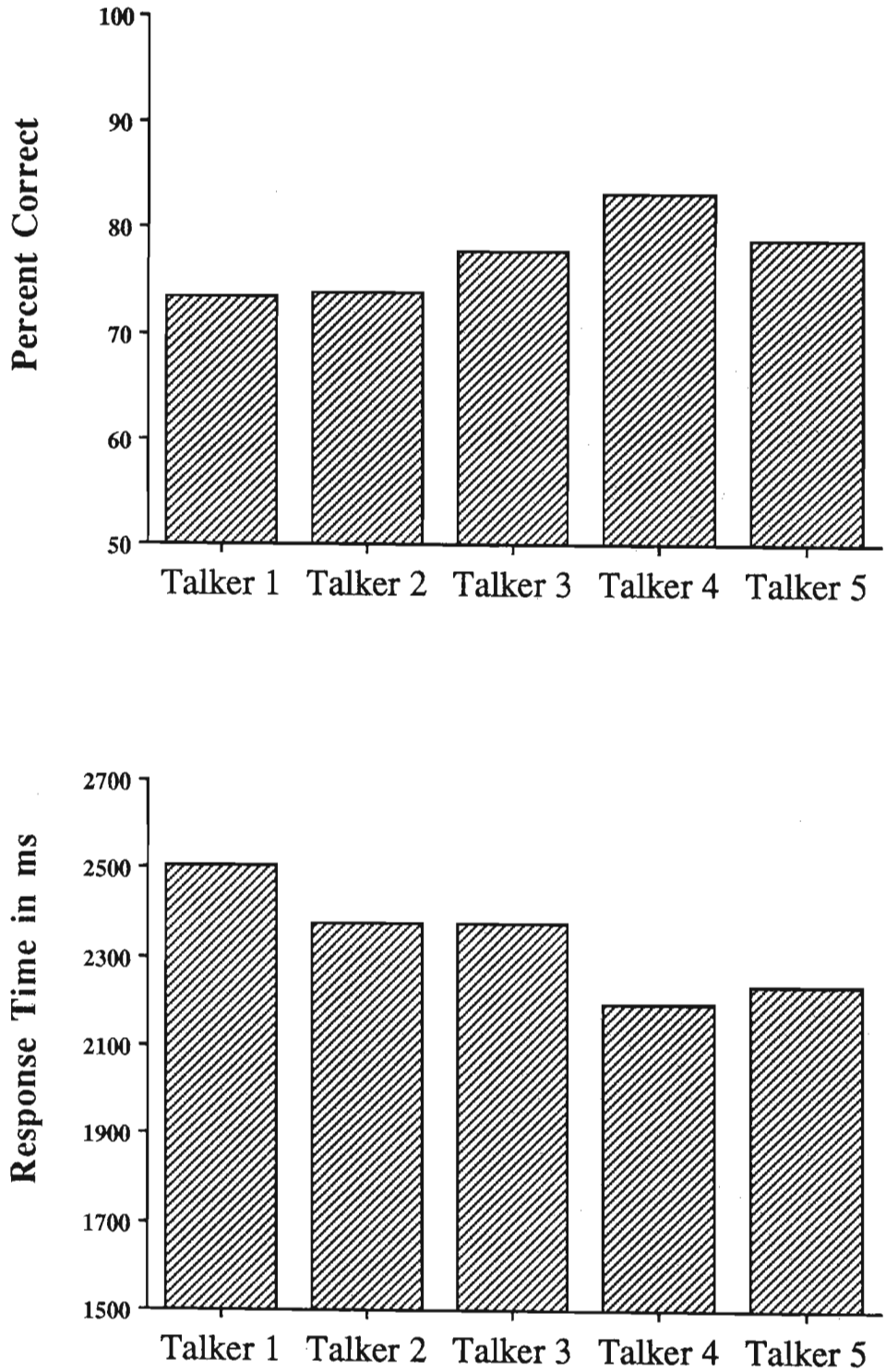


Figure 3. The top panel shows accuracy during training as a function of talker. The lower panel shows response latencies.

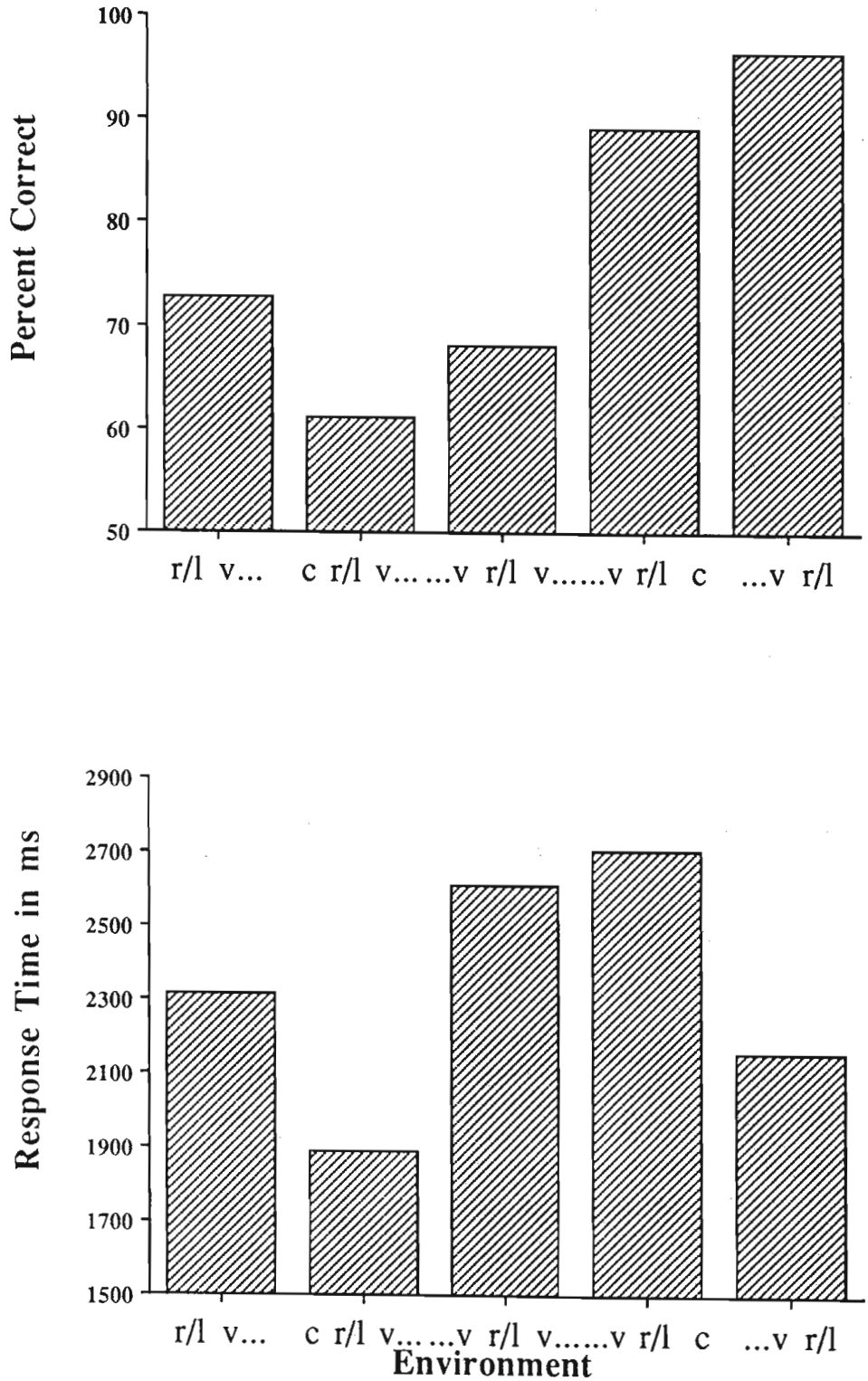


Figure 4. The top panel shows accuracy during training as a function of phonetic environment. The lower panel shows response latencies.

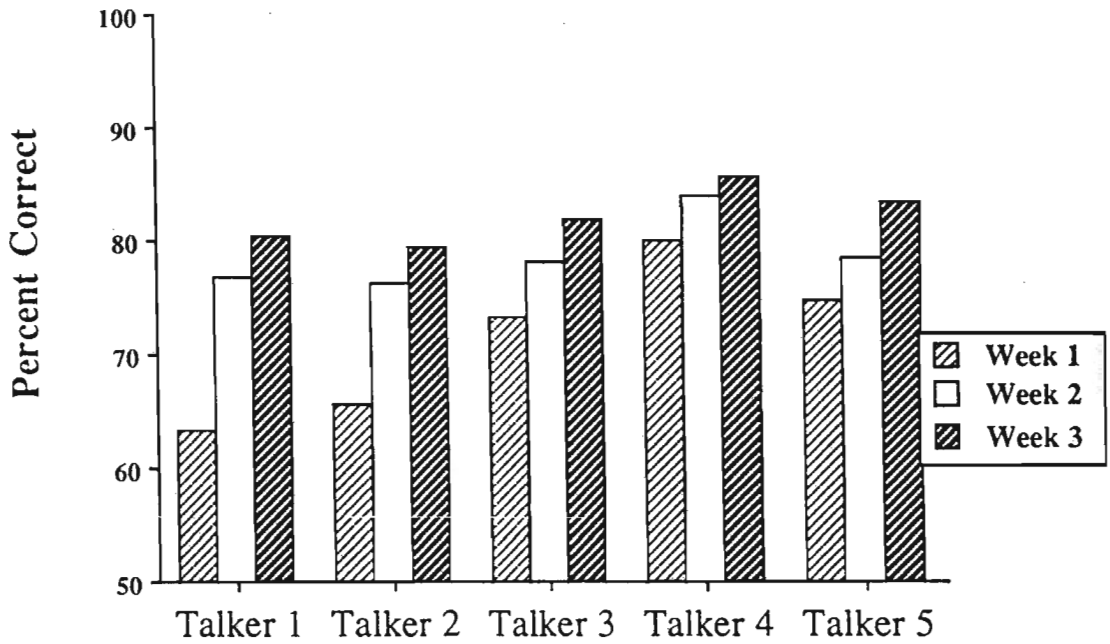


Figure 5. This figure shows the interaction of talker and week in the accuracy data during the training phase.

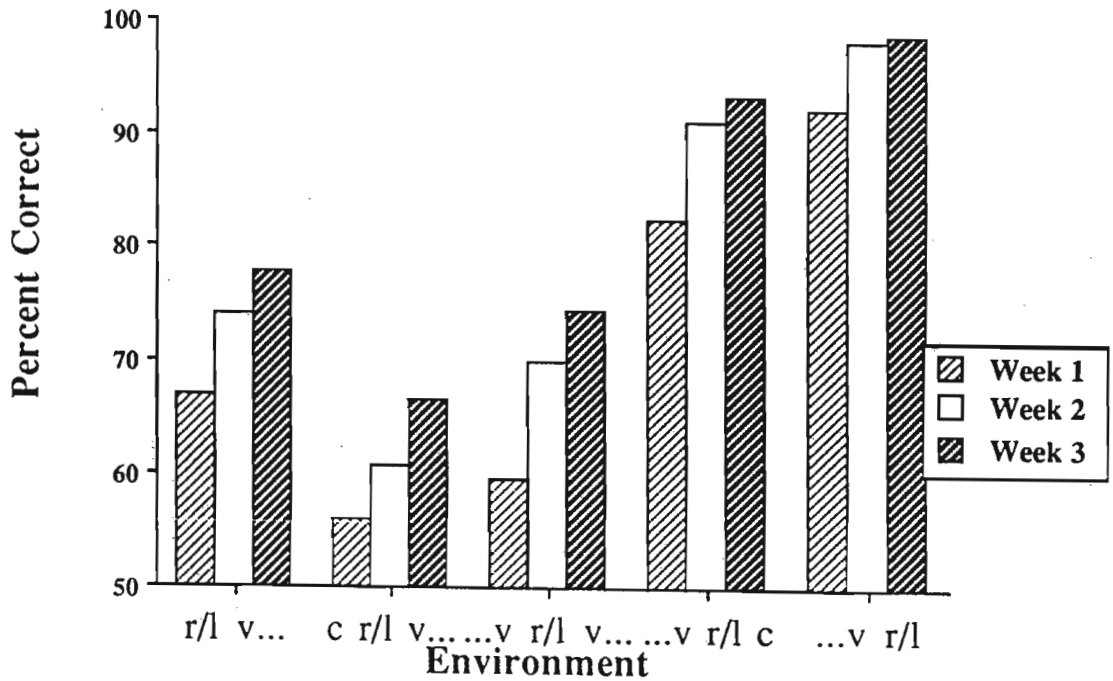


Figure 6. This figure shows the interaction of phonetic environment and week in the accuracy data during the training phase.

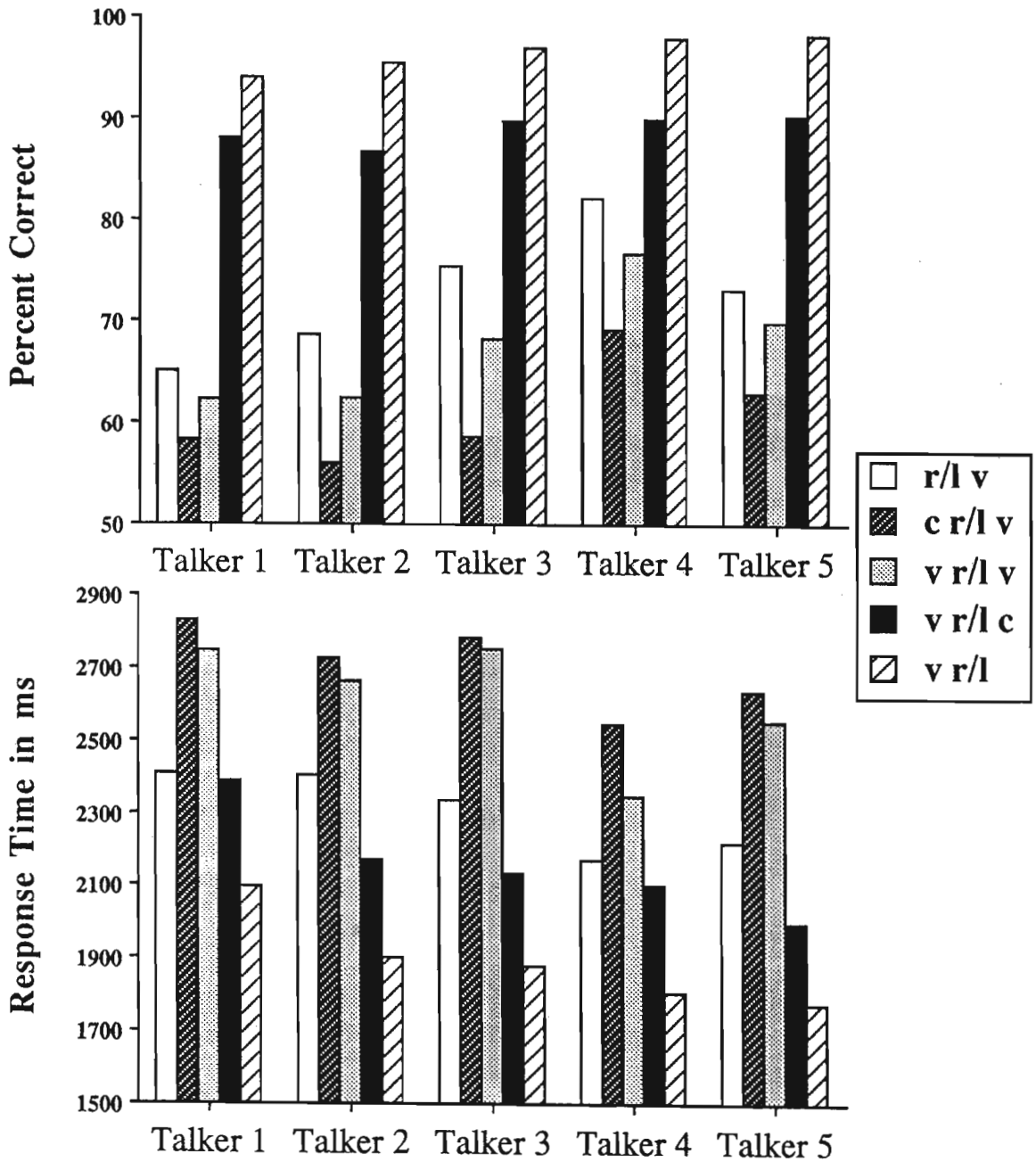


Figure 7. The top panel shows the interaction of talker with phonetic environment during training in the accuracy data. The lower panel shows a similar interaction in the response time data.

Third, interactions between talker and phonetic environment were observed in both the accuracy and latency analyses [$F_{pc}(16,288)=8.22, p<.01$; $F_{rt}(16,288)=2.64, p<.01$]. The top panel of Figure 7 shows accuracy as a function of phonetic environment and talker, while the bottom panel displays the corresponding response times. In general, Talker 4 was responded to more accurately than all other talkers when /r/ and /l/ occurred in initial singleton, initial consonant clusters and intervocalic positions. No significant differences were observed among any of the talkers for targets in final consonant clusters and final singleton position. Talker 4 was also responded to faster than Talker 1 when /r/ and /l/ occurred in initial singleton, initial consonant clusters, intervocalic, and final consonant cluster positions. Responses to Talker 5 were also faster than to Talker 1 when targets occurred in final consonant clusters and final singleton position.

Finally, a three-way interaction among week, talker and phonetic environment was obtained in the response time data [$F(32,576)=1.76, p<.01$]. Responses to Talker 1 tended to be slowest during Week 1 of training for /r/'s and /l/'s in final consonant clusters and final singleton position. By Week 3, however, no significant differences were observed among any talkers for targets in final consonant clusters and final singleton position.

Tests of Generalization

Separate ANOVAs were conducted on the mean accuracy and response latencies collected during the tests of generalization. Talker and phonetic environment were within-subjects variables in each analysis. A main effect for talker was obtained in the accuracy analyses [$F(1,18)=12.51, p<.01$]. These results are shown in the top panel of Figure 8. Responses to items produced by the old talker were significantly more accurate than responses to tokens produced by the new talker (82% vs. 77%, respectively). The main effect for talker did not approach significance in the analysis of the response time data [$F(1,18)<1$]. Mean response time to the old talker was 2067 ms, while mean response time to the new talker was 2053 ms.

Insert Figure 8 about here.

Main effects for phonetic environment were also observed in both the accuracy and latency analyses [$F_{pc}(3,54)=48.03, p<.01$; $F_{rt}(3,54)=9.05, p<.01$]. Targets in final consonant clusters and final singleton position were identified more accurately than targets in initial singleton position and initial consonant clusters. In addition, /r/'s and /l/'s in final singleton position were identified more accurately than /r/'s and /l/'s in final consonant clusters. Targets in initial singleton and final singleton positions were identified faster than targets in initial consonant clusters and final consonant clusters.

Because we observed significant differences in accuracy as a function of talker during the tests of generalization, it is important to determine if these results are due to simple differences in intelligibility or whether they are due to listeners encoding talker-specific information about the stimulus items during training. We tested this hypothesis by having 14 additional untrained Japanese listeners perform the tests of generalization. The same two alternative forced-choice identification procedure was used and the order of presentation was counterbalanced across subjects. Mean accuracy and latency scores from each subjects were submitted to separate ANOVAs. Talker and phonetic environment were within-subjects variables in each analysis. No main effect for talker was obtained in either the analysis of the accuracy data or the latency data [old talker: 71%, new talker: 70%,

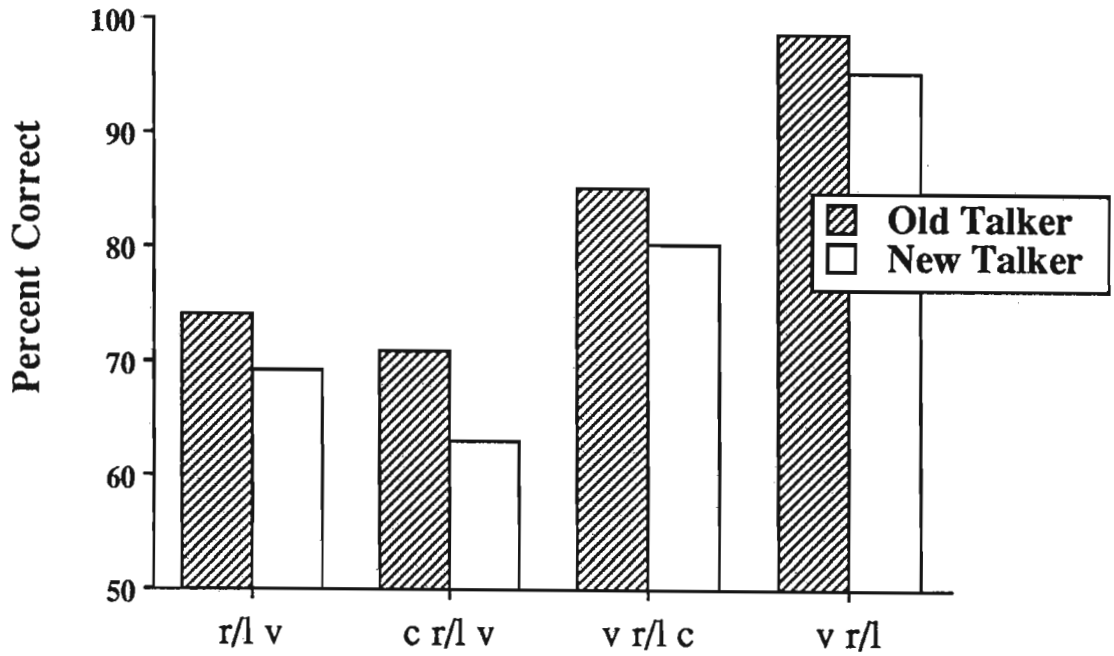


Figure 8. This figure shows subjects' accuracy in the tests generalization as a function of talker.

$F_{pc}(1,13)<1$; old talker: 2971 ms, new talker: 2746 ms, $F_{rt}(1,13)<1$]. The interaction of talker with phonetic environment was not significant in either analysis [$F_{pc}(3,39)=1.32, p<.3$; $F_{rt}(3,39)<1$]. Thus, these results, obtained with a control group of untrained listeners, rule out baseline differences in intelligibility as the source of the differences in performance between talkers in the tests of generalization.

Three-month tests

Three months after the conclusion of training, 16 of the original 19 subjects returned and performed the posttest and two tests of generalization again. A separate ANOVA was conducted on the accuracy data from these subjects to compare performance on the pretest and posttest and to assess retention after three months. Test and phonetic environment were within-subjects variables. Mean accuracy scores are presented as a function of test and phonetic environment for these subjects in Figure 9. A significant main effect for test was obtained [$F(2,30)=53.93, p<.01$]. Subjects accuracy increased significantly from the pretest to the posttest, but did not *decrease* significantly from the posttest to the three-month follow-up. The main effect for phonetic environment was also significant [$F(3,45)=38.14, p<.01$]. /r/'s and /l/'s in final singleton position were identified more accurately than targets in any other environment. In addition, phonemes in initial position were identified more accurately than targets in initial consonant clusters.

Insert Figure 9 about here.

Several additional ANOVAs were conducted on the mean accuracy and latency scores collected from the 16 subjects who participated in the original tests of generalization and the follow-up tests of generalization given three months after the conclusion of training. Time of test, talker and phonetic environment were treated as within-subjects variables. Mean accuracy scores are plotted as a function of each of these variables in Figure 10. Original test scores for these 16 subjects are displayed in the left panel, while three-month follow-up scores are displayed in the right panel.

Insert Figure 10 about here.

Several results are noteworthy: First, the main effect for time of test was not significant [$F(1,15)=2.60, p<.2$]. Mean accuracy for the original tests of generalization was 79.4%, while mean accuracy for the generalization tests given three months later was 77.9%. Second, a main effect for talker was obtained [$F(1,15)=11.80, p<.01$]. Subjects were more accurate in identifying /r/'s and /l/'s produced by a familiar talker than by an unfamiliar talker (81.0% vs. 76.3%, respectively). Separate ANOVAs conducted on the data from the original tests of generalization and the follow-up tests given after three months replicated this pattern [$F_{orig}(1,15)=9.08, p<.01; F_{3\text{-}month}(1,15)=10.04, p<.01$]. Finally, a main effect for phonetic environment was also obtained [$F(3,45)=37.83, p<.01$]. Targets in final singleton position were identified more accurately than phonemes in any other environment. Words containing /r/'s and /l/'s in final consonant clusters were also identified more accurately than words with /r/ and /l/ in initial singleton position or initial consonant clusters.

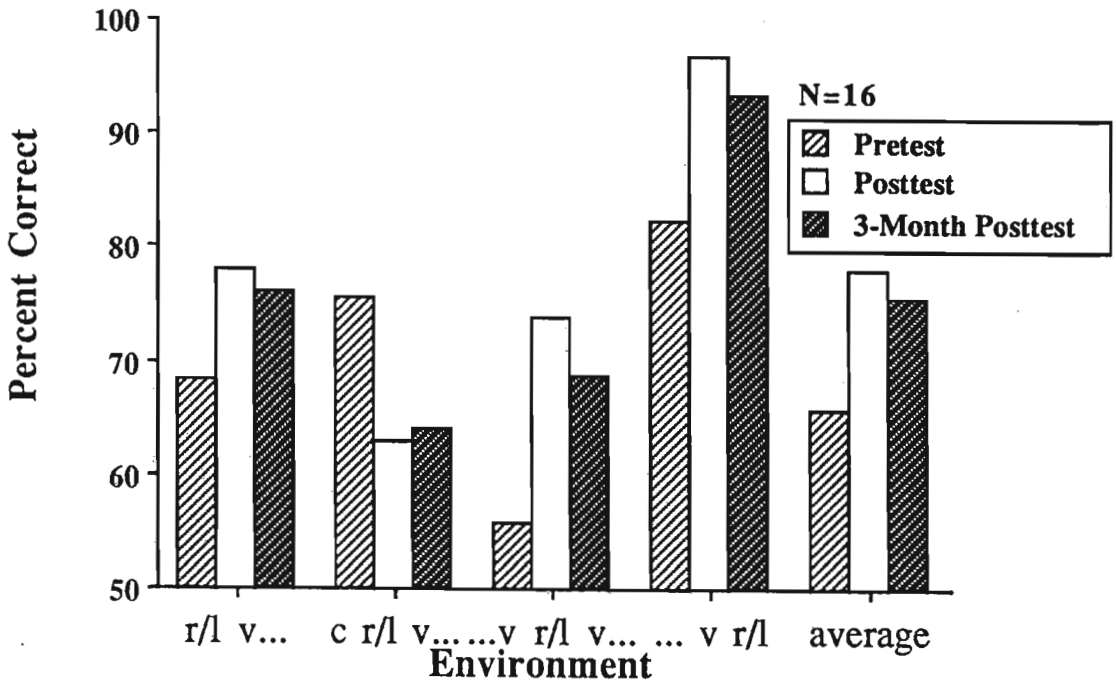


Figure 9. This figure shows accuracy scores for the pretest, posttest, and three-month follow-up tests of 16 subjects.

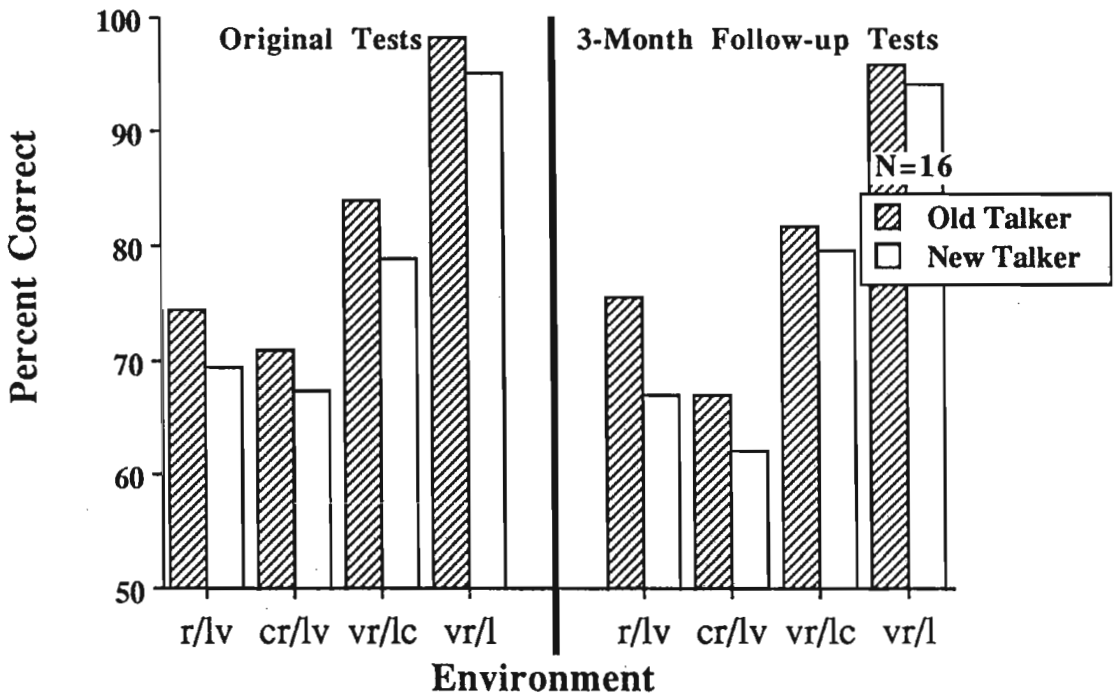


Figure 10. The left panel of the figure shows accuracy scores during the first tests of generalization for 16 subjects. The right panel shows accuracy scores from the same subjects during the three-month follow-up tests of generalization.

Discussion

One of the long-term goals of cross-language training experiments is to develop new perceptual categories for nonnative phonemic contrasts that are robust and permanent. These new perceptual categories must be applicable to highly variable stimuli, as listeners routinely encounter the nonnative phonetic contrast in many different phonetic environments. The categories must also be applicable to many different talkers. Procedures that encourage this goal may be considered effective, whereas procedures that are not robust to the variability of the linguistic environment may be considered ineffective. The methodological and theoretical significance of the present study and Logan et al.'s investigation should be viewed with these broad guidelines.

Several results from the present experiment and the earlier Logan et al.'s (1991) study demonstrate the effectiveness of the high-variability identification training paradigm. First, listeners' accuracy increased by 12% from the pretest to the posttest and by 11% during the three weeks of training. Logan et al. observed a slightly more modest gain of 5% to 7% from the pretest to the posttest and during training. In the present investigation, subjects' accuracy increased significantly during each week of training for phonemes that occurred in the three most difficult phonetic environments (initial singleton, initial consonant clusters, and intervocalic positions).[8] Although Logan et al. did not observe an interaction between phonetic environment and week of training, mean levels of performance in the earlier study were comparable to those reported here. These findings demonstrate that the training procedure is effective in modifying listeners' phonetic perception and the methods can be used with monolingual subjects. The magnitude of the training effect appears to be modulated by the amount of exposure subjects have to the contrast in the natural language environment (MacKain, Best, & Strange, 1982).

Second, listeners' identification accuracy increased significantly during the training phase for each of the talkers used in training. Initially, large differences in accuracy were observed among the five training talkers. By the end of training, however, performance was almost equivalent among all talkers. Only Talker 4 was responded to more accurately than Talker 1 and Talker 2 ($p < .05$). This finding suggests that listeners apparently become attuned to the specific characteristics of the training talkers voices in learning the new phonetic contrasts (Goto, 1971). Thus, listeners not only acquire information about acoustic-phonetic cues for phonetic categorization of the novel items, but they also acquire information about the indexical properties of the talker's voice as well. This information appears to become part of the mental representation of speech in long-term memory.

Third, the results from the test of generalization showed that the training effects extended to an unfamiliar talker. Mean response accuracy was 77% for tokens from the unfamiliar talker (old talker: 82%). This level of performance is approximately equal to mean performance obtained during the second week of training and is greater than mean performance during the first week of training for four of the five training talkers. It should be noted, however, that generalization to the new talker was not complete because a significant effect for talker was still obtained in the analysis of the accuracy data. These findings suggest that the training paradigm was moderately effective in promoting generalization

[8]Lively et al. (in press) reported a similar interaction in their second experiment. In that experiment listeners were trained using the high-variability identification paradigm with tokens from only a single talker. Accuracy increased from the week 1 to week 2 for /r/'s and /l/'s in initial singleton and intervocalic positions and initial consonant clusters. Significant increases were also observed from week 2 to week 3 for /r/'s and /l/'s in initial consonant clusters.

to new talkers. Moreover, the results indicate that generalization is not an "all or none" phenomenon; rather, it is graded (Logan et al., in press) and appears to be dependent on previous experience and the amount and type of variability available in the stimulus set (Goggin, Thompson, Strube, & Simental, 1991; Lively et al., in press).

Fourth, our results indicate that subjects retained very specific knowledge about the new phonetic contrast for a period of at least three months. Performance did not decrease significantly over the three month interval during which subjects received no further training or exposure to /r/ and /l/. This result was somewhat surprising, given that subjects were living in a monolingual Japanese-speaking environment and were not exposed to much spoken English during this time. This finding indicates that our high-variability training procedure was effective in changing the long-term representations of these perceptual categories. To our knowledge, this study represents the first attempt to assess the retention of new phonetic categories by adult listeners under laboratory conditions.

Although our results indicate that listeners retained knowledge about the new phonetic category over a period of three months, it is necessary to qualify this claim. When subjects were retested three months after the conclusion of training, they were given the same tests that they had previously received. Thus, they listened to the pretest-posttest stimuli for the third time during the three month posttest and heard the stimuli from the tests of generalization for the second time. It is possible that subjects learned something about the specific tests themselves and were therefore responding on the basis of stimulus-specific knowledge, rather than from knowledge about the new phonetic contrast. The methodology of the tests tends to discount this possibility. Listeners were given feedback only during the training phase of the experiment and would have no way to judge the accuracy of their responses from trial to trial during the pretest, posttest or tests of generalization. In future studies, new test items should be used whenever possible. However, it is worth pointing out here that the words used in the pretest and posttest, training and tests of generalization virtually exhaust all of the minimal pairs of words in English containing /r/ and /l/.

Taken together, the results of the present experiment suggest that the high variability identification procedure is effective in modifying nonnative listeners' phonetic perception. The use of natural speech tokens from several talkers promoted robust generalization to new words and new talkers. More importantly, however, these changes were retained over time for at least a three month period. Two questions are raised by these results. First, what is the nature of the mechanism that causes this reorganization of phonetic perception? Second, what factors contribute to listeners' long-term retention of the new categories, given that exposure to the contrast outside of the laboratory is minimal?

Several issues concerning the acquisition and retention of new phonetic categories can be addressed by considering recent proposals about the role of selective attention in categorization and perceptual learning (Jusczyk, 1989, 1993, in press; Nosofsky, 1986, 1987). According to Nosofsky, selective attention "stretches" and "shrinks" memory representations during category acquisition. Representations are stretched along contrastive dimensions so that items from different categories are made less similar to each other. Similarly, representations are shrunk along noncontrastive dimensions so that items from the same category are made more similar to each other. Jusczyk (1993, in press) has extended this basic framework to address specific issues of phonetic category acquisition in infants. He has proposed that acoustic dimensions are extracted from the incoming signal by a bank of auditory processing filters (Sawusch, 1986) and that these perceptually relevant dimensions are weighted

automatically in terms of their importance in determining linguistic contrasts. These attention weights are then used to generate a precompiled interpretative scheme that can be applied automatically to incoming fluent speech (Jusczyk, 1989, 1993; Klatt, 1979).

A similar process may occur when Japanese listeners are trained to identify English /r/ and /l/. Over the course of training, selective attention weights are changed to favor the new phonetic categories which cause a reorganization of listeners' phonological spaces or filters (Flege, 1989; Terbeek, 1977). Alternatively, the attention weights for contrasts in the native language may not change during training, but an unused portion of a general acoustic-phonetic space may be used for the new phonetic categories (Best, 1993). The extent to which either alternative applies will depend on the nature of the contrast in the listeners' native language.

In addition to accounting for the acquisition of new phonetic contrasts, an explanation must also be given for the results of the retention findings of the present investigation. In order to accommodate these results, it is important to consider the nature of adult listeners' phonological systems. Within the framework of Jusczyk's model, adult listeners have well-developed selective attention weights that serve to maximize contrasts in their native language. As training progresses, we assume selective attention weights are reorganized to incorporate the new phonetic contrast. Simultaneously, however, the weights remain stable enough so that contrasts in the native language are not disrupted. When training ends, the newly established attention weights may become fixed.

Two aspects of the linguistic environment suggest this possibility. First, the listeners in the present experiment were adults who already have well-established phonological systems that are capable of perceiving the contrasts of the native language. Infants who are in the process of acquiring their native language, in contrast, would not be expected to have a fixed set of attention weights. Rather, their phonological systems are dynamically adapting to sound patterns in their native language environment. Only after approximately 12 months of age would infants be expected to adopt a fixed set of attention weights (Werker, 1989; see however, Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Second, because subjects were living in a monolingual Japanese speaking environment, little exposure to the English /r/-/l/ contrast would be expected. Thus, subjects would have little experience or feedback (either explicit or implicit) in perceiving the /r/-/l/ contrast. Because feedback is unavailable after the end of training, selective attention weights would have little impetus to change. When subjects are tested again after three months without training, they may employ the same interpretative scheme they had used during testing at the end of training. Thus, little loss would be expected in the tests given three months after the conclusion of training.[9]

Two basic assumptions underlie this selective attention-based account of category acquisition and retention. First, we assume that selective attention weights were changed during training to favor the acquisition of a new phonetic contrast. Second, we assume that the new weights remained stable over time because little input or feedback from the environment was available to change them. This account makes an important prediction for future research. If subjects were trained during the three month interval on a new contrast that required a different weighting of the same stimulus dimensions, interference would be anticipated at the time of retest. The magnitude of the interference should be modulated by the degree to which the selective attention weights were modified. These changes would be determined by the similarity of the two contrasts on which subjects were trained (Best, 1993).

[9]We would not want to suggest that listeners' performance will never return to pretest levels without further exposure. Rather, we suggest that attention weights, at least in these listeners, do not appear to change over a three month period. Performance changes over longer intervals remain to be examined in future studies.

It is interesting to note that the present set of findings also provide important new information about the nature of the mental representation for the new phonetic contrast that is developed during training. In each of the training studies we have conducted, we have found that accuracy varies as a function of the specific talker used in training (Lively et al., 1991, in press; Logan et al., 1991). Moreover, we have found that generalization accuracy is also dependent on familiarity with the voice of the talker producing the stimuli. The differences we observed in generalization performance cannot be accounted for by simple differences in intelligibility because we found that tokens from the familiar and unfamiliar talkers were equally intelligible to untrained Japanese control listeners. The results suggest that aspects of a talkers voice are episodically encoded into long-term memory and that these representations are used to facilitate the later recognition of spoken words (Goldinger, 1992). This conclusion is contrary to the traditional assumption that indexical properties of speech, such as gender, dialect, age, and affect, are filtered out of the linguistic representation of speech by some kind of perceptual normalization mechanism (Laver & Trudgill, 1979; Laver, 1980). Instead, we argue that listeners may bring highly-detailed information to bear on the recognition and retention of new items (Hintzman, 1986).

Support for this proposal has come from several recent studies that have shown that information about a talker's voice is episodically encoded into long-term memory (Goldinger, 1992; Goldinger, Pisoni, & Logan, 1991; Nygaard, Sommers & Pisoni, 1992, 1993; Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, 1992). Mullennix and Pisoni (1990) showed that voice information cannot be filtered out or selectively ignored by listeners in a Garner selective attention task (Garner, 1974). In a recognition memory study, Palmeri et al. (1993) found that listeners recognized previously presented words more accurately when they were repeated in the same voice. The effect was robust even when 64 intervening items occurred prior to the repetition and when the stimulus set included 20 different voices. These findings suggest that listeners encode detailed information about a talker's voice into memory and use this information to facilitate the recognition of spoken words.

In another experiment, Nygaard et al. (1993) examined the effects of familiarity with a voice on spoken word recognition. The authors trained subjects to identify 10 different voices. After training, half of the subjects were presented with new words produced by the talkers used in training. The remaining subjects served as a control group and were presented with new words produced by new talkers. In each condition, subjects were asked to identify words presented against a background of white noise. Listeners who heard words produced by the talkers used in training responded more accurately than the subjects who heard new talkers. These results indicate that familiarity with a talker's voice facilitates the recognition of spoken words. Taken together with the present results, these findings concerning the effects of talker variability suggest that voice information and the phonetic forms of spoken words are integrally perceived, stored in memory, and used to facilitate later recognition of novel words.

Conclusions and Future Directions

In many fields of science, researchers routinely carry out replications of earlier published findings so that new studies can be designed to extend and elaborate on basic phenomena, fundamental knowledge and general principles. Replications serve an important function in establishing the robustness of phenomena and the reliability of the experimental methodology. In areas of controversy, replication is considered to be absolutely essential to the scientific process.

The present investigation was originally designed with these broad goals in mind. Given that we have replicated our previous findings using a larger group of monolingual subjects, we believe are in a position to make several conclusions. First, the high-variability training procedure we have developed provides an effective means for quickly modifying adult nonnative listeners' phonetic perception. In the present report, monolingual native speakers of Japanese learned to identify English /r/ and /l/. Identification accuracy increased 11% during three weeks of training. Second, the training procedure encourages the development of robust new phonetic categories. Listeners in the present experiment generalized to new words produced by a new talker. Generalization was not complete, however, as subjects performed more accurately when tokens were produced by a familiar talker. Third, listeners retained the new categories over a three-month interval, without any further training. Performance on the posttest and the tests of generalization did not decrease significantly over this retention interval.

Finally, to account for these results, we suggest that a selective attention mechanism differentiates items along contrastive dimensions. Over the course of training, this mechanism stretches and shrinks the underlying psychological space in order to make the contrastive phonemes more dissimilar from each other. Furthermore, we suggest that subjects episodically encode detailed stimulus information and that this information is brought to bear on the identification and recognition of new stimulus items.

The results of the present experiment suggest several interesting directions for future research. One important issue will be to localize the source of the talker-specific encoding effects observed in training. Careful acoustic analysis of the training tokens may reveal some important differences among the training talkers that may help us to explain the observed differences. Another important issue concerns the relationship between perception and production. In each of the experiments we have conducted (Logan et al., 1991; Lively et al., in press), we have only examined the effects of training on perceptual identification. An important new direction for future research will be to assess the effects of perceptual training on aspects of speech production and the effects of production training on speech perception (Sheldon & Strange, 1982). Results from these studies may give us valuable insights into the nature of the representation that listeners develop for new phonetic categories and how these representations mediate other aspects of perception and production.

In summary, Japanese listeners were trained to identify English /r/ and /l/ using a high variability identification procedure. Subjects improved from the pretest to the posttest and during training and generalized to new words produced by a new talker. Performance varied widely as a function of talker during training and generalization. Subjects also demonstrated retention of the new perceptual categories over a three month interval in which they were not exposed to the contrast in their native language environment. The results of the present experiment replicate Logan et al.'s laboratory training procedures and extend their findings on the perception of /r/ and /l/ by Japanese listeners in an important way by demonstrating robust category retention over time. The results have several implications for current theoretical conceptions of perceptual learning of nonnative phonetic contrasts.

References

Best, C. T. (1992). The emergence of language-specific influences in infant speech perception. In J. Goodman and H. C. Nusbaum (Eds.) *Development of speech perception: The transition from recognizing speech sounds to spoken words*. Cambridge, MA: MIT Press.

Best, C. T. & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, **20**, 305-320.

Carney, A., Widin, G., & Viemeister, N. (1977). Noncategorical perception of stop consonants varying in VOT. *Journal of the Acoustical Society of America*, **62**, w961-970.

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

Edman, T., Soli, S., & Widin, G. (1978). Learning and generalization of intra-phonemic VOT discrimination. *Journal of the Acoustical Society of America*, **63**, 19.

Flege, J. (1990). Perception and production: The relevance of phonetic input to L2 phonological learning. In C. Ferguson and T. Huebner (Eds.) *Crosscurrents in second language acquisition and linguistic theories*. Philadelphia, PA: John Benjamins.

Flege, J. (1989b). 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.

Flege, J. & Wang, C. (1989). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast. *Journal of Phonetics*, **17**, 299-315.

Garner, W. (1974). *The processing of information and structure*. Potomac, MD: LEA.

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**, 59-72 (University of Minnesota, Minneapolis).

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, **19**, 448-458.

Goldinger, S. D. (1992). Words and voices: Implicit and explicit memory for spoken words. Research on speech perception: Technical report no. 7. Bloomington, IN: Indiana University Press.

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.

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

Henly, E., & Sheldon, E. (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**, 411-428.

Jamieson, D. & Morosan, D. (1986). Training non-native speech contrasts in adults: Acquisition of the English /_/-/q/ contrast by francophones. *Perception & 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.

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429-466). Hillsdale, NJ: LEA.

Jusczyk, P. (1989). Developing phonological categories from the speech signal. Paper presented at The International Conference on Phonological Development, Stanford University.

Jusczyk, P. (1993). Infant speech perception and the development of the mental lexicon. In J. Goodman & H. C. Nusbaum (Eds.) *The transition from speech sounds to spoken words: The development of speech perception*. Cambridge, MA: MIT Press.

Jusczyk, P. (in press). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*.

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

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavioral Development*, **6**, 263-285.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B, (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, **255**, 606-608.

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 K. Salzinger & S. Salzinger (Eds.), *Research in verbal behavior and some neurological implications*. (pp. 79-98). New York: Academic Press.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

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

Lehiste, I. (1960). Acoustic characteristics of selected English consonants. *International Journal of American Linguistics*, **30**, 10-115.

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

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

Lively, S. E., Pisoni, D. B., & Logan, J. S. (1991). Some effects of training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.) *Speech perception, production and linguistic structure.* (pp. 175-196). Tokyo: OHM Publishing Co., Limited.

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

Logan, J. S., Lively, S. E., & Pisoni, D. B. (in press). Training listeners to perceive novel phonetic categories: How do we know what is learned? *Journal of the Acoustical Society of America.*

Mann, V. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "R" and "L". *Cognition*, **24**, 169-196.

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

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

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

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

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,* **15**, 700-708.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1992). Effects of speaking rate and talker variability on the representation of spoken words in memory. In *ICSLP-92, Banff*, 209-212.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1992). Speech perception as a talker- contingent process. *Research on speech perception, Progress report No. 18,* Speech Research Laboratory, Indiana University, Bloomington.

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

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

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1993). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In J. Goodman and H. Nusbaum (Eds.) *Development of speech perception: The transition from recognizing speech sounds to spoken words*. Cambridge, MA: MIT Press.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). *Journal of Experimental Psychology: Human Perception and Performance,* **8**, 297-314.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic and, acoustic contributions. *Journal of the Acoustical Society of America*, **89**, 2961- 2977.

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics*, **52**, 37-52.

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

Posner, M. & Keele, S. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, **83**, 304-308.

Price, J. P. (1981). A cross-linguistic study of flaps in Japanese and in American English. Unpublished doctoral dissertation, University of Pennsylvania.

Pruitt, J. S. (in press). Comments on "Training Japanese listeners to identify /r/ and /l/: A first report" [J. S. Logan, S. E. Lively, and D. B. Pisoni, J. Acoust. Soc. Am. 89, 874-886 (1991)] [71]. *Journal of the Acoustical Society of America.*

Sawusch, J. R. (1986). Auditory and phonetic coding of speech. In E. C. Schwab & H. C. Nusbaum (Eds.) *Pattern recognition by humans and machines: Vol. 1, Speech perception* (pp. 51-88). New York: Academic Press.

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

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

Strange, W. (1972). *The effects of training on the perception of synthetic speech sounds: Voice onset time*. Unpublished 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 & Psychophysics*, **36**, 131-145.

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

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

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

Werker, J. (1989). On becoming a native listener. *American Scientist*, **77**, 54-59.

Werker, J. & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, **37**, 35-44.

Werker, J. & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, **75**, 1866-1878.

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Speech Perception as a Talker-Contingent Process[1]

Lynne C. Nygaard, Mitchell S. Sommers, and David B. Pisoni

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

[1]This research was supported by NIH Training Grant #DC-00012-13 and NIH Research Grant #DC-00111-16 to Indiana University. We thank Carol Fowler, Robert Remez, Peter Eimas, Scott Lively and Steve Goldinger for their helpful comments and suggestions on an earlier version of this manuscript.

Abstract

To determine how familiarity with a talker's voice affects perception of spoken words, two groups of subjects learned to recognize a set of voices over a nine day period. One group then identified words at four signal-to-noise ratios that were produced by the same set of talkers. Control subjects identified words at the same signal-to-noise ratios that were produced by a different set of talkers. The results showed that the ability to identify a talker's voice improved intelligibility of novel words produced by the same talkers. The results suggest that speech perception may involve talker-contingent processes whereby familiarity with aspects of the vocal source facilitates the phonetic analysis of the acoustic signal.

Speech Perception as a Talker-Contingent Process

During the perception of speech, listeners must extract stable phonetic percepts from acoustic signals that are highly variable. Variations in talker characteristics, in particular, have been shown to produce profound effects on the acoustic realization of speech sounds (Nearey, 1978; Peterson & Barney, 1952). The consequences of this variability for the perception of speech is that phonetic segments, syllables, and words do not necessarily have any invariant acoustic form (but see, Stevens & Blumstein, 1978). Speech signals produced by different speakers may have quite different acoustic characteristics due to individual differences in size and shape of the vocal tract, dialect, and age. How then is this variability resolved by perceivers of speech? What role do talker-specific characteristics play in the perception of speech? Are representations of the linguistic content of speech separate from the talker-specific attributes inherent in any spoken utterance?

Traditionally, models of speech perception have characterized variation in the acoustic speech signal as a perceptual problem that perceivers must solve (Shankweiler, Strange, & Verbrugge, 1976). Listeners are thought to achieve stable phonetic percepts given variation in talker through a compensatory process in which specific voice characteristics are normalized to arrive at canonical linguistic units (Joos, 1957; Summerfield & Haggard, 1973). Unfortunately, this standard view of talker normalization begs the question of how processing of talker information is related to the processing of the phonetic content of speech. Although a talker's voice carries important indexical information into the communicative setting (Laver & Trudgill, 1979), the encoding of talker-specific information has generally been considered to be a problem that is quite separate from apprehending the linguistic content of an utterance (but see Garvin & Ladefoged, 1963; Johnson, 1990; Miller, 1989). On one hand, researchers have investigated the ability of listeners to explicitly recognize and discriminate familiar and unfamiliar voices (e.g., Legge, Grossmann, Pieper, 1984; Van Lancker, Kreiman, Emmorey, 1985). On the other hand, with a few exceptions, research in speech perception has been devoted to studying the linguistic content of speech independently of any variability in talker or source characteristics.

The theoretical and empirical dissociation of the encoding of talker characteristics and the processing of the phonetic content of an utterance assumes that these two kinds of information are independent (Laver & Trudgill, 1979). Only recently has this assumption been questioned based on a growing body of research demonstrating effects of talker variability on both perceptual and memory processes. For example, several studies (Mullennix, Pisoni, and Martin, 1988; Summerfield and Haggard, 1973) found that phoneme and word recognition performance was poorer for listeners presented with words produced by multiple talkers compared to words produced by a single talker. In addition, using a speeded classification task, Mullennix and Pisoni (1990) reported that subjects could not ignore irrelevant variation in a talker's voice when asked to categorize the initial phonemes of isolated words. Taken together, these findings suggest that variability in the speech signal due to voice is time and resource demanding and may not be entirely independent of processing of the phonetic content of the signal.

Talker variability has been found to affect memory processes as well. Martin, Mullennix, Pisoni, and Summers (1989) and Goldinger, Pisoni, and Logan (1991) both found that at relatively fast presentation rates the serial recall of words in initial list positions was superior for lists of words produced by a single talker as compared to lists of the same words produced by multiple talkers. At longer presentation rates, however, Goldinger et al. (1991) reported that words in initial list positions from multiple-talker lists were recalled better than words from single-talker lists. These results

suggest that at fast presentation rates, variation due to talker differences affects the initial encoding and subsequent rehearsal of the to-be-remembered lists. However, superiority in recall performance for multiple-talker lists at longer presentation rates suggests that given sufficient time listeners were able to fully process and encode each word and use the accompanying talker information for further elaboration. Consequently, as opposed to fast presentation rates where talker variability hinders recall performance, at longer presentation rates listeners were able to use the distinctive talker information to aid performance in the serial recall task.

Further evidence that talker-specific information is encoded and retained in long-term memory comes from recent experiments by Palmeri, Goldinger, and Pisoni (1993). Using a continuous recognition memory procedure, specific voice information was shown to be retained along with item information and these attributes were found to aid in later recognition. The finding that subjects were able to use talker-specific information suggests that this source of variation may not be discarded in the process of speech perception, but rather variation in a talker's voice may become part of a rich and highly detailed representation of the speaker's utterance. The decrements in performance due to talker variability would then be due to the additional attention and resources necessary to encode the indexical information conveyed by a talker's voice.

Although previous experiments have demonstrated a relationship between processing of talker information and processing of the phonetic content of a speaker's utterance, the question remains as to the nature of this relationship (Ladefoged & Broadbent, 1955). The purpose of the present experiment was to address this question more directly by investigating the effects of long-term memory representations of voice on phonetic encoding. To accomplish this, two groups of listeners were asked to learn to recognize the names of ten talkers over a nine day period. At the end of the training period, we evaluated the role of talker recognition on the perception of spoken words to determine if the ability to identify a talker's voice was independent of phonetic analyses, as assumed by current theoretical accounts of speech perception (Liberman & Mattingly, 1985; McClelland & Elman, 1986; Stevens & Blumstein, 1978). One group of listeners identified a set of novel words at four signal-to-noise ratios produced by the same ten talkers that they had been exposed to during the nine days of training. The other group of listeners also received a set of novel words at the same signal-to-noise ratios produced by ten unfamiliar talkers that had not been presented in training. If familiarity with a talker's voice is found to affect subsequent word recognition performance, the results would argue for a link in processing between encoding of talker information and phonetic perception and would suggest that speech perception may be a talker-contingent process.

METHOD

Subjects

Subjects were thirty-eight undergraduate and graduate students at Indiana University. Nineteen subjects served in each condition--experimental and control. Each subject participated in ten one-hour sessions. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. The subjects were paid for their services.

Stimulus Materials

Three sets of stimuli were used in this experiment. All were selected from a database of 360 monosyllabic words produced by ten male and ten female talkers. Word identification tests in quiet showed greater than 90% intelligibility for all words. In addition, all words were rated to be highly familiar (Nusbaum, Pisoni, & Davis, 1984). The stimuli were originally recorded on audiotape and

digitized at a sampling rate of 10 kHz on a PDP 11/34 computer using a 12-bit analog-to-digital converter. The root mean squared (RMS) amplitude levels for all words were digitally equated.

The training stimuli consisted of a subset of 160 words produced by five male and five females. The words and talkers used in the training phase were selected randomly from the larger database. Generalization stimuli consisted of a subset of 100 words produced by the same ten talkers used in training. These items were also chosen randomly from the original database. Stimuli used in the word recognition test consisted of the remaining 100 words from the original stimulus set. For the experimental condition, the 100 words used in the word recognition test were produced by the same ten talkers used in the training set. For the control condition, the same 100 words used in word recognition were produced by ten talkers that were different from those used in the training phase.

Procedure

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

On each of the nine training days, both groups of listeners completed three different phases. The first phase consisted of a familiarization task. Five words from each of the ten talkers were presented in succession to the listeners. Following this, subjects heard a ten-word list composed of one word from each talker in succession. Each time a token was presented to the listeners, the name of the appropriate talker was displayed on a CRT screen. Listeners were asked to listen carefully to the words presented and to attend specifically to the talker's voice so they could learn the name.

The second phase of training consisted of a recognition task in which subjects were asked to identify the talkers who had produced each token. Ten words from each of the ten talkers were presented in random order to listeners who were asked to recognize the voice by pressing the appropriate button on a keyboard. The keys were labeled with ten names. Keys 1-5 were labeled with male names; keys 6-10 were labeled with female names. On each trial, after all subjects had entered their responses, the correct name appeared on the CRT screen.

After subjects completed two repetitions of the first two phases of training, we administered a test phase on each day. As in the training phase, ten words from each of the ten talkers were presented in random order. However, feedback was not given for a correct response.

Although the words used in the test phase were drawn from the same one hundred words used in the training phase, on each day of training there was no overlap of individual tokens between the test phase and the training phase. That is, subjects never heard the same item produced by the same talker in both the test and training phase. In addition, training stimuli were re-selected from the database on each day so that subjects never heard the same word produced by the same talker in training. This training procedure was designed to expose listeners to a diverse set of tokens from each of the talkers.

Generalization. On the tenth day of the experiment, both groups of subjects completed a generalization test. One hundred new words produced by each of the ten familiar talkers were used. As in the test phase used during training, ten words from each of the ten talkers were presented in

random order. No feedback was given. Thus, the generalization test was identical to the training test phase, except that listeners had never heard any of the words before.

Word Intelligibility. In addition to the generalization test, we also administered a speech intelligibility test in which subjects were asked to identify words presented in noise. In this transfer task, each trial began with a warning prompt displayed in the middle of the CRT screen. Five hundred ms after offset of the prompt, continuous, white noise low-pass filtered at 4.8 kHz was presented at 70 dB (SPL). Fifty ms after the onset of the noise, a stimulus word was presented at either 80, 75, 70, or 65 dB (SPL) yielding four signal-to-noise ratios; +10, +5, 0, and -5. The noise was terminated 50 ms after the offset of the stimulus word. Equal numbers of words were presented at each of four signal-to-noise ratios. Also, the number of words produced by a talker at each signal-to-noise ratio was counterbalanced. In this test, subjects were simply asked to identify the *word* itself (rather than explicitly recognize the talker's voice) by typing in their response on a keyboard. Only exact phonetic matches to the presented stimuli were counted as correct.

RESULTS AND DISCUSSION

Training. Overall, the training procedure was successful. Subjects showed continuous improvement across the nine days in their ability to recognize talkers from isolated words. However, substantial individual differences were found in performance. Consequently, we selected a criterion of 70 percent correct for talker recognition on the last day of training for inclusion in the experiment. Our rationale for choosing this criterion was simply that to determine whether learning a talker's voice affects perceptual processing, we needed to ensure we had identified a group of subjects that did, in fact, learn to successfully recognize the talkers' voices from isolated words. Using the 70 percent correct talker identification criterion, nine subjects from each training group were included in the experiment.[2] Figure 1 shows the results of the training procedure for both groups of listeners. Percent correct recognition of talkers is plotted as a function of day of training. Chance performance on this task is defined as 20 percent assuming that subjects are able to reliably discriminate gender and thus limit their choice on a given trial to the five male or five female talkers. Both groups of listeners identified talkers consistently above chance even on the first day of training and performance rose to nearly 80 percent correct by the last day of training. A repeated measures ANOVA with learning and days of training as factors showed a significant main effect of day of training [$F(9,144) = 73.55$, $p < .0001$] but no difference between the two groups over days of training [$F(1,16) = 0.14$, $p > .7$].

Insert Figure 1 about here

Generalization. The results of the generalization test are also presented in Figure 1. Percent correct name recognition is shown for both groups by the filled symbols. Recognition of voices from novel words was almost identical to that found on the final day of training. The implication of this result is that listeners acquired detailed knowledge about the talkers' voices that was not necessarily dependent on the specific words or specific tokens that carried that information. That is, the perceptual learning

[2]It should be noted that the task of learning to identify voices from isolated words is extremely difficult (see Williams, 1964). Therefore, it was necessary to set a somewhat arbitrary training period and then select subjects who had learned to our criterion by the end of that period. Given additional training with isolated words or the use of sentences or larger passages of speech, a greater percentage of our subjects would have reached a criterion level of performance.

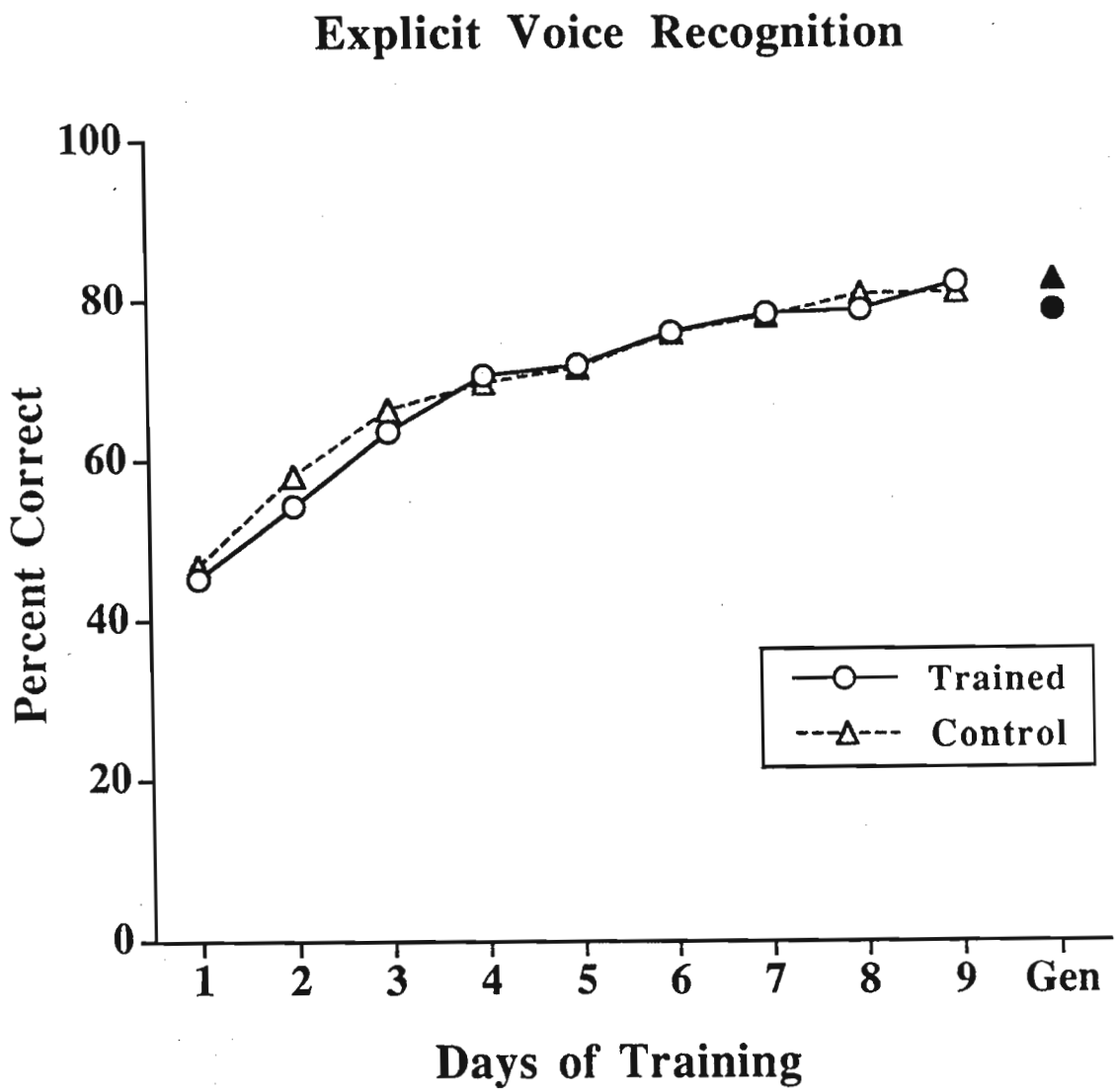


Figure 1. Explicit voice recognition performance for both groups of listeners. Percent correct talker recognition is plotted as a function of days of training (open circles and open triangles). The results of a generalization test administered on the tenth day are also included. Percent correct talker recognition on the generalization test is plotted for both groups of learners (closed circles and triangles).

that takes place in the course of the nine training sessions is not dependent on the training stimuli but rather readily generalizes to novel utterances produced by the same set of talkers.

Word Intelligibility. Figure 2 shows percent correct word identification as a function of signal-to-noise ratio for both groups of trained subjects. Perceptual identification performance of subjects that were tested with familiar talkers was significantly better than control subjects who were tested with unfamiliar voices. As expected, identification performance decreased from the +10 to the -5 signal-to-noise ratio for both groups of subjects. However, subjects who were tested with words produced by familiar voices were significantly better in recognizing novel words at each signal-to-noise ratio than controls. A repeated measures ANOVA with training and signal-to-noise ratio as factors revealed highly significant main effects of both signal-to-noise ratio [$F(3,48) = 173.27$, $p < .0001$] and training condition [$F(1,16) = 13.62$, $p < .002$].[3]

To ensure that the overall intelligibility of the two sets of voices did not differ, two additional groups of eighteen untrained subjects who were not familiar with either set of talkers were also given the same word intelligibility test. One untrained group received the stimulus tokens produced by the talkers who were used in the training phase; the other untrained group received the stimulus tokens from the talkers that were presented to the control group. Identification performance for the trained and untrained control groups did not differ. A separate repeated measures ANOVA including only the control conditions revealed only a significant main effect of signal-to-noise ratio [$F(2,38) = 0.57$, $p > .5$]. This finding confirms that the difference in performance between the experimental group and the trained controls was not due to inherent differences in the intelligibility of the voices or words used.

Insert Figure 2 about here

CONCLUSIONS

The present study found that voice recognition and the processing of the phonetic content of a linguistic utterance were not independent. Rather, listeners who learned to recognize a set of talkers apparently retained talker-specific information in long-term memory that facilitated the subsequent perceptual processing and identification of novel words produced by the same talkers. These findings provide the first demonstration that exposure to a talker's voice facilitates perceptual processing of the phonetic content of that speaker's novel utterances. Not only does the perceptual learning that results from the talker recognition task generalize to novel words produced by familiar talkers, but it also affects the perceptual processing of the phonetic content of novel words produced by the same talkers in a speech intelligibility test. Listeners that were exposed to one set of voices but were tested with another set of voices failed to show any benefit of the experience gained by learning to explicitly recognize those voices. Only experience with the specific voices used in the intelligibility test facilitated the phonetic processing of novel words. The implication of this result is that phonetic perception and spoken word recognition appear to be affected by knowledge of specific information about a talker's voice. Exposure to specific acoustic attributes of a talker's voice appears to contribute to the processes used to recognize spoken words.

---

[3]Four items from both the trained and untrained control conditions were eliminated from the overall analyses. After the experiments had been run, these items were found to be mispronounced.

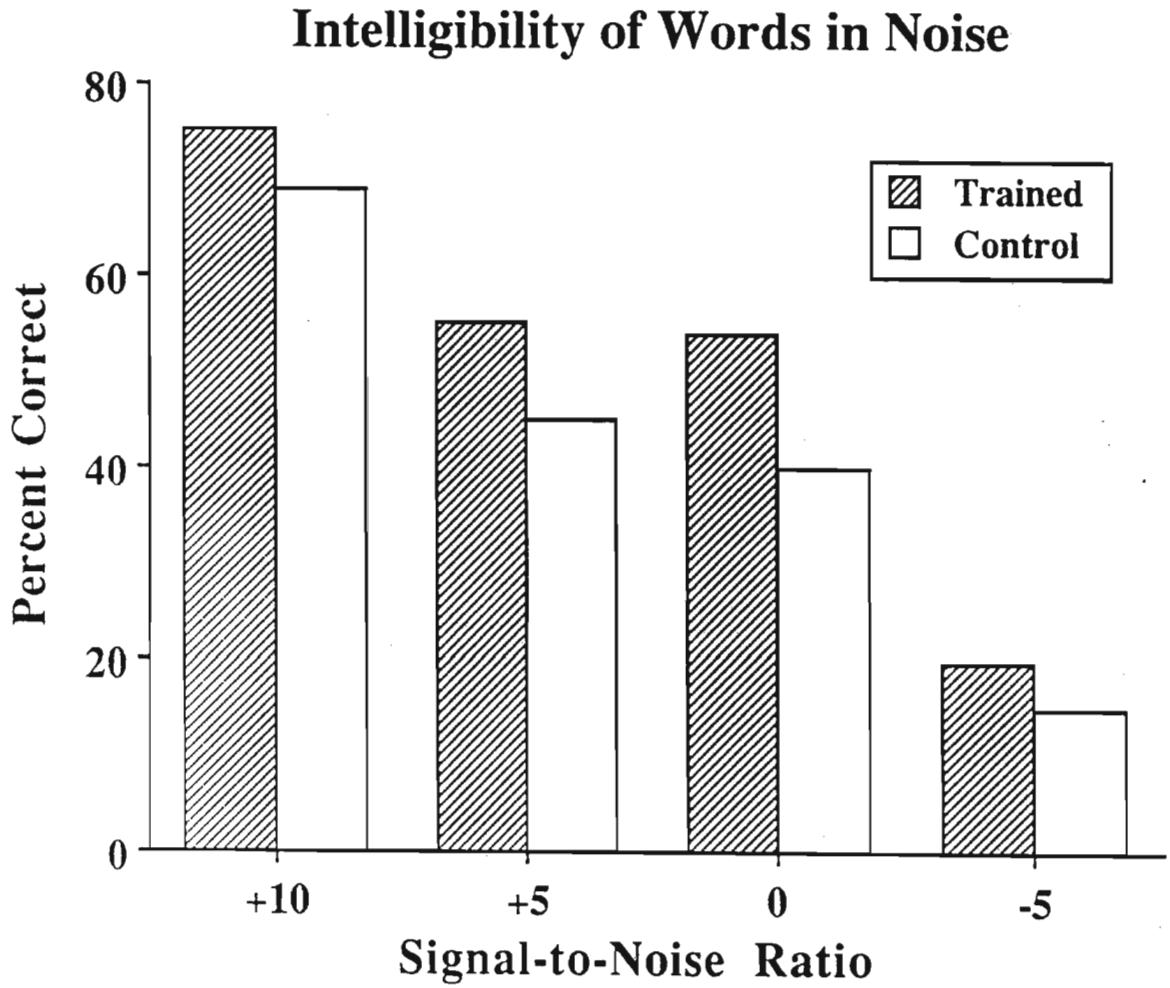


Figure 2. Mean intelligibility of words presented in noise for trained and control subjects. Percent correct word recognition is plotted at each signal-to-noise ratio.

What kind of knowledge is acquired when listeners learn to recognize voices? One possibility is that the procedures or perceptual operations necessary for the extraction of voice information from the speech signal are encoded in a type of procedural memory (Kolers, 1976; Kolers & Roediger, 1984). That is, the specific perceptual operations used to compensate or normalize for each talker's voice may be retained in memory. This procedural knowledge would then allow listeners more efficient processing of novel words produced by familiar talkers. We believe this situation may be very similar to the case of reading and remembering inverted text. Kolers and Ostry (19974) found that the operations necessary to read inverted text were retained in long-term memory and facilitated subsequent tasks involving reading inverted text. The type of detailed procedural knowledge that they describe may be responsible for subjects' superior performance in identifying words spoken by familiar talkers in the present experiment. Alternatively, specific sets of features or attributes of a talker's voice may be merely listed in memory (Johnson, 1990; Miller, 1989). Attributes such as average fundamental frequency, relative formant spacing, and glottal characteristics, for example, may all be stored in a memory representation for a talker's voice and used as a reference or template for subsequent phonetic processing. Another possibility is that the characteristic scale or time-varying properties of a talker's vocal tract may be directly perceived and remembered (Fowler, 1986). In learning about a talkers' voice, listeners may become sensitive to information in the acoustic signal about specific dynamic properties of the talker's vocal tract as an acoustic source. Whatever the exact nature of the talker-specific information, it is important to emphasize here that exposure to these voices during the training phase allowed listeners to use this knowledge to help them perceive and recognize novel spoken words.

The present findings have several implications for current theories of speech perception and spoken word recognition. First, the ability of subjects to acquire information about a talker's voice that they can subsequently use to facilitate phonetic processing suggests that the representation of spoken words in memory may be much more detailed than previously thought (Palmeri, Goldinger, & Pisoni, 1993). Listeners appear to be sensitive to the unique aspects and distinctive characteristics of a talker's voice and this sensitivity to form may not be independent of the processing of the phonetic content of the talker's utterance. The interaction observed between knowledge of a talker's voice and speech perception suggests that the mental representations of spoken words may incorporate both phonetic content and talker-specific source characteristics. At the very least, both types of information may be retained (Mullennix & Pisoni, 1990). This proposal differs from traditional accounts of speech perception which assume that talker information as well as other sources of variability in speech are discarded by the processes used to arrive at the canonical, abstract linguistic units that are presumably encoded into long-term memory.

Second, whatever the exact nature of the process that compensates for variation in voice characteristics, the present findings demonstrate that this process can be modified by experience and training with a specific talker's voice. The interaction of learning to identify a talker's voice and the processing of the phonetic content of a talker's utterance suggests that any proposed mechanism of perceptual compensation is susceptible to general processes of perceptual learning and attention. Thus, the processing of a talker's voice may demand time and resources if the voice is unfamiliar to a listener (Martin, Mullennix, Pisoni, & Summers, 1989; Mullennix, Pisoni, & Martin, 1989; Summerfield & Haggard, 1973), but may become much more efficient if the voice can be identified as a familiar talker (Lightfoot, 1989).

Finally, this study provides the first direct demonstration of the role of long-term memory for source characteristics in speech perception and spoken word recognition. Experience with a talker's voice facilitates perceptual processing of the acoustic-phonetic attributes of an utterance. Indeed, the

present findings suggest that the phonetic coding of speech is carried out in a talker-contingent manner. Knowledge of the characteristics of a talker's vocal tract and consequently, attributes of a talker's voice, appear to be used by listeners to facilitate phonetic perception and spoken word recognition.

References

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

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

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

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

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

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

Kolers, P. A., & Roediger, H. L., III. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, **23**, 425-449.

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

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

Legge, G. E., Grossmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 98-303.

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

Lightfoot, N. (1989). Effects of talker familiarity on serial recall of spoken word lists. *Research on Speech Perception Progress Report No. 15.* Bloomington, IN: 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**, 676-681.

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

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

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

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

Nearey, T. (1978). *Phonetic features for vowels.* Bloomington, IN: Indiana University Linguistics Club.

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

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

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.

Shankweiler, D. P., Strange, W., & Verbrugge, R. R. (1976). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, knowing: Toward an ecological psychology.* Hillsdale, NJ: Erlbaum.

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.

Summerfield, Q., & Haggard, M. P. (1973). Vocal tract normalisation as demonstrated by reaction times. *Report on Research in Progress in Speech Perception, 2,* Belfast, Northern Ireland: The Queen's University of Belfast.

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

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

Training Listeners to Perceive Novel Phonetic Categories: How Do We Know What is Learned?[1]

John S. Logan[2], Scott E. Lively and David B. Pisoni

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

[1]We wish to thank two anonymous reviewers for their insightful comments on a previous draft of this manuscript. This research was supported, in part, by NIH Research Grant DC00111-16 and, in part by NIDCD Research Training Grant DC00012-14 to Indiana University in Bloomington, IN.
[2]Now at the Department of Psychology, Carleton University, Ottawa, Ontario, Canada.

Abstract

Logan, Lively, and Pisoni (1991) carried out a perceptual learning experiment in which Japanese listeners were trained to identify English words containing /r/ and /l/. Pruitt (1992) has criticized several aspects of our procedures and conclusions. First, he argues that the lack of appropriate control groups make interpretation of our results problematic. Second, he asserts that our generalization test was methodologically flawed. Although Pruitt raises some important issues that are worth pursuing in future research, we argue below that the methodology that we employed and the conclusions that we drew were valid.

Training Listeners to Perceive Novel Phonetic Categories: How Do We Know What is Learned?

The Effect of Training on Pretest/Posttest Performance

Pruitt contends that the design of our experiment did not allow us to disentangle a number of factors that may have contributed to the success of our training regimen. We trained native Japanese listeners in an identification task and evaluated their performance using a pretest/posttest design. During training, subjects were presented words which contained /r/ and /l/ in five different phonetic contexts. Each of the words was produced by five different talkers. In addition, extensive feedback was provided. Feedback not only included information about whether subjects' responses were correct or incorrect, but subjects were also re-presented any stimulus on which they made an error. Logan et al. demonstrated that such a training procedure produced a statistically reliable improvement in identification performance for naturally produced English words containing /r/ and /l/. This finding contrasts sharply with the result obtained by Strange and Dittmann (1984), who found very little improvement on the natural speech tokens that were used during the pretest-posttest phase of their experiment. Moreover, we also found that performance varied systematically as a function of phonetic environment and talker.

Pruitt takes issue with the fact that we did not systematically explore the relative contribution of each of these factors to the improved identification performance we observed. His complaint, however, neglects several important issues. First, it was never our intention in this initial study to examine every possible permutation of stimulus and procedural variables. Instead, our goal was to see if we could develop a laboratory-based training procedure that would facilitate the identification performance of Japanese listeners presented with naturally produced English words containing /r/ and /l/ (Logan et al., p. 874). Based on an analysis of previous efforts (Jamieson & Morosan, 1986, 1989; Strange & Dittmann, 1984), we determined that sufficiently variable stimuli combined with an appropriate training task were critical for the formation of robust new phonetic categories. Our results convincingly demonstrated both of these points.

Second, we ourselves acknowledged that substantial work remained to determine the extent to which factors such as talker variability and phonetic variability contributed to the improved performance we obtained (see Logan et al., p. 883). To this end, we recently have completed several experiments that examined the separate contribution of phonetic variability and talker variability (see Lively, Logan, & Pisoni, submitted). In one experiment, we trained Japanese listeners with tokens from the three most difficult phonetic environments for Japanese listeners (initial singleton, initial consonant clusters and intervocalic positions). Tokens containing /r/ and /l/ from these environments were produced by five different talkers. We replicated our results. Improvements in identification performance were obtained from the pretest to the posttest and during training. In addition, generalization performance was not reliably different when listeners responded to tokens produced by an old talker compared to tokens produced by a new talker; performance with both generalization talkers was equivalent to performance at the conclusion of training. In a second experiment, we trained another group of listeners with a single talker who produced the /r/-/l/ contrast in all five phonetic environments. Improved performance in some phonetic environments was observed both in the posttest and during training. However, the limitation of this training strategy was revealed by the results of the tests of generalization. Subjects were significantly more accurate in identifying tokens

from a familiar talker. However, their mean level of performance with the old talker was only equivalent to performance during Weeks 1 and 2 of training. Overall, the results of these new training experiments suggest that talker variability plays a larger role than phonetic variability, thus addressing one of the major concerns raised by Pruitt in his criticism of our original study.

Pruitt suggests that the improvements in identification performance obtained by Logan et al. in the posttest could be due to "... subjects' aptitudes, the testing conditions, or the quality of the stimuli...". First, our subjects were not so homogeneous in terms of 'aptitude' that this characteristic alone caused them to perform in an identical manner during the experiment. As shown in Table A2 of Logan et al., average pretest performance ranged from 58.3% to 95.8% correct, not at all what one would expect if initial performance was equivalent across subjects. Moreover, the fact that all of our subjects showed a consistent improvement from pretest to posttest can be taken as evidence that the effects of training were real and not due to characteristics specific to the group of subjects we tested.

We are in complete agreement with Pruitt when he suggests that 'testing conditions' and 'the quality of the stimuli' were responsible for the improvements we observed. After all, providing the appropriate 'testing conditions' and ensuring 'quality of the stimuli' was one of the major goals of our experiment to begin with! It is important to note here that our pretest-posttest materials were the same words that were used by Strange and Dittmann, who failed to obtain any reliable changes from the pretest to the posttest. The implication of our results is that we used an effective training program whereas previous efforts did not.

Finally, Pruitt notes that for words containing /r/ and /l/ in initial position, the pretest performance for the subjects tested in our experiment was "...quite good when compared to other studies with native Japanese speakers and naturally produced stimuli (e.g., Strange and Dittmann, 1984, showed only 64% correct identification on word-initial singletons at pretest)." The point of this statement appears to be that our subjects' initial level of performance was solely responsible for the improvements we observed. We take issue with this on the grounds that the level of word-initial singleton performance obtained in Strange and Dittmann's pretest is not representative of all Japanese subjects. For example, Mochizuki (1981) reported a figure of approximately 88% correct for seven Japanese subjects presented words containing /r/ and /l/ in word-initial singleton positions, a value higher than the 80% we observed for word-initial singletons in the pretest phase of our experiment. Furthermore, we have recently replicated our earlier findings with a group of monolingual Japanese speakers (N = 19) from Kyoto, Japan (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1992). Mean performance on the pretest across all phonetic environments was only 63% for these subjects and was as low as 52% for contrasts in initial consonant clusters. In short, there is no systematic evidence to back Pruitt's claims that initial levels of performance were responsible for the results obtained in Logan et al. We conclude that our training procedure was effective, even for subjects whose initial performance may have been very poor.

Generalization Performance

Pruitt's second major point concerns the tests of generalization that we used to assess the listeners' ability to transfer what they had learned during training to novel words and novel talkers. The tests of generalization were administered to subjects after they had completed training and the posttest. It consisted of two parts, novel words produced by a novel talker and novel words produced by a familiar talker whom subjects had heard during training. Pruitt argues that generalization cannot be assessed using our procedure for two reasons. First, he states that there was no control for the

intelligibility of the two talkers. Second, he states that there was no mention of controlling the phonetic contexts across the two parts of the test. Finally, he argues that we "...overstated the results of the test of generalization.". Designing a test of generalization to assess what listeners have learned from a training task poses several difficulties. Pruitt correctly points out that the inherent intelligibility of talkers is one factor that could play a significant role in determining listeners' performance. However, the complexity of this issue is illustrated by the following example. In our experiment we pretested all of the stimuli used in the generalization tests with native speakers of English to insure that all of the stimuli were equated for intelligibility. Yet, when the Japanese listeners were presented with the same stimuli, their performance varied as a function of talker and phonetic context. The obvious point of this example is that intelligibility depends upon whether the listener is a native speaker or nonnative speaker. The less obvious point relates to how stimuli can be equated for intelligibility when they are presented to nonnative listeners. If one were to attempt to equate the intelligibility of talkers for the Japanese listeners, what criteria would be used to demonstrate equivalent intelligibility? In the case of English listeners, all of the tokens selected for use in our experiment produced no /r/ or /l/ errors. Unfortunately, given the variability in performance due to phonetic context for the Japanese listeners, it is unclear how talkers could be best matched for intelligibility.

Instead, it seemed to us to be more productive to deal directly with variability in intelligibility and its effect on generalization. Thus, we compared the most intelligible talker in training to a talker that our subjects had never heard before. We tested three Japanese listeners and found that their performance was marginally better for the familiar talker than for the unfamiliar talker. Although it is possible that this effect was due to differences in the inherent intelligibility of the talkers, it is also consistent with a growing body of evidence suggesting that listeners retain quite detailed information about unique perceptual episodes, such as the specific talker producing a specific word (e.g., Craik & Kirsner, 1974; Goldinger, 1992; Palmeri, Goldinger, & Pisoni, in press; Pisoni, 1992; Schacter & Church, in press). In addition, Lively et al.'s (submitted) recent training results demonstrate that generalization performance appears to be governed more by the composition of the training items, rather than inherent differences in intelligibility. These findings indicate that we should consider the contribution of different sources of variability to phonetic category acquisition before trying to identify an intelligibility metric for nonnative listeners.

Pruitt's second criticism of the test of generalization used in Logan et al. was that it was not clear whether the words produced by the two talkers were comparable in terms of phonetic context. The stimuli in our tests of generalization were minimal pairs of English words that contrasted /r/ and /l/ in the same five phonetic environments used during training. The majority of these words contained /r/ or /l/ in initial singleton position or initial consonant clusters. Thus, the distribution of items used during training and generalization tests were similar across phonetic environments. Unfortunately, syllabic and phonetic contexts could not be controlled precisely across stimulus sets due to the limited number of English words that contrast /r/ and /l/ in each potential context. The fact that none of the words used in the pretest and posttest or in training were repeated in either of the tests of generalization imposed further constraints on the distribution of words according to context. In short, although the tokens used in these three sets of words virtually exhaust the set of /r/-/l/ minimal pairs available in English, the composition of words in each set was comparable.

The final criticism Pruitt makes regarding our test of generalization is that we overstated our results. We disagree. Pruitt's argument is based, in part, on the marginal difference that we observed when we compared generalization performance for words produced by an old talker and words produced by a new talker. The most convincing evidence against Pruitt's position is that Lively et al.

(1992) have replicated the generalization results obtained by Logan et al. In their replication, Lively et al. used the same stimuli and procedures used in our original experiment to test Japanese listeners living in Japan. In addition to obtaining a reliable improvement between pretest and posttest as a function of training, the listeners also reliably identified novel words produced by a familiar talker more accurately than novel words produced by a unfamiliar talker. Given this replication of our earlier findings, we do not think that we overstated our original claims.

As a final comment on the issue of generalization, Pruitt's statements suggest that any difference in performance between familiar and unfamiliar talkers is at odds with our claims regarding the robust nature of our training procedure -- "...it is paradoxical that Logan et al. insist that this nonsignificant trend indicates a lack of generalization considering their claims regarding stimulus variability and robust learning.". We do not think that any of our statements were paradoxical. Instead, Pruitt's statements suggest that he considers generalization to be an all or none phenomenon. While such a position would immeasurably simplify the evaluation of training methods, it is unlikely that this is the case. As our data demonstrate, generalization depends on the degree of similarity between the training stimuli and the test stimuli. As the test stimuli become more similar to the training stimuli, they will be identified more accurately. Similarly, as test stimuli diverge from training stimuli, they will be identified less accurately (cf. Strange and Dittmann, 1984). Thus, discrepancies between pretest/posttest and generalization performance are not paradoxical unless one relies on a narrow definition of what generalization means.

Conclusion

It is rare in science for one experiment to address each and every variable that may affect the outcome of an experiment. If scientists delayed publication until they had meticulously examined every possible combination of experimental variables that might affect the outcome of an experiment, the progress of science would be ill-served. Logan et al. described an initial effort to examine the role of stimulus variability in training listeners to perceive nonnative phonetic categories. We have taken this initial effort and used it as a starting point for systematically exploring some of the variables that may have contributed to the effects we obtained (e.g., Lively et al., submitted; Lively et al., 1992). Contrary to Pruitt, we believe that this study has advanced our understanding of some very basic processes involved in learning novel phonetic categories. Our results address several important issues in speech perception and, more generally, in the field of perceptual categorization. First, we have demonstrated an effective means for modifying phonetic perception in nonnative speakers of English. Second, our findings suggest that listeners develop context-sensitive representations for new phonetic categories, rather than abstract, idealized canonical representations. Talker variability and variability due to phonetic environment appear to be important factors in acquiring new phonetic categories. Finally, our results demonstrate that selective attention to the contrastive cues of novel phonetic categories can be quickly and robustly modified in the laboratory with relatively simple training procedures. The important role of stimulus variability in acquiring new phonetic categories contrasts sharply with the traditional assumption that listeners develop abstract context-invariant prototypes for nonnative speech sounds and that these units are acquired by focusing listeners' attention only on the contrastive or criterial cues used by native speakers of a language.

References

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

Goldinger, S.D. (1992). Words and voices: Implicit and explicit memory for spoken words. *Research on speech perception: Technical report no. 7.* Bloomington, IN: Indiana University Press.

Jamieson, D. & Morosan, D. (1986). Training non-native speech contrasts in adults: Acquisition of the English / /-/ / contrast by francophones. Perception & 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.

Lively, S.E., Logan, J.S., & Pisoni, D.B. (submitted) Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning perceptual categories.

Lively, S.E., Pisoni, D.B., Yamada, T, Tohkura, Y., & Yamada, R. (1992). Training Japanese listeners to identify English /r/ and /l/: III. Long-term retention of new phonetic categories. *Research on Spoken Language Processing Progress Report No. 18.* Bloomington, IN: Speech Research Laboratory

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

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

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

Pisoni, D.B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. *ICSLP92.*

Pruitt, J.S. (1992). Comments on "Training Japanese listeners to identify /r/ and /l/: A first report" [J. S. Logan, S. E. Lively, and D. B. Pisoni, *J. Acoust. Soc. Am.* **89**, 874-886 (1991)] [71].

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

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

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 18 (1992)
Indiana University

A New PC-based Real-time Experiment Control System[1]

Luis R. Hernandez, Thomas D. Carrell, James G. Reutter, Robert H. Bernacki

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

[1]The development of this system was supported by NIH Research Grant DC00111-16 to Indiana University in Bloomington, IN.

Abstract

This report describes the expansion and development of the computer hardware and software facilities of the Personal Computer (PC) based real-time experiment control system. We have outlined requirements and described an implementation plan of a PC configuration that will eventually replace our current PDP-11/34-based experimental control system.

A New PC-based Real-time Experiment Control System

Overview

This project has several basic goals. The first is to design a replacement for the aging PDP-11/34-based real-time experiment control systems used in our laboratory (see Forshee & Nusbaum, 1984). This is necessary because support for these systems has become virtually nonexistent both in terms of hardware and software. Furthermore, the architecture is 25 years old and, by modern standards, it is an extremely difficult environment in which to develop robust applications. The second goal of the project is to develop an environment in which a variety of experimental tasks are possible that are difficult to perform with existing equipment. The third goal of this project is to develop a system which is useful enough and inexpensive enough for researchers in speech perception and other related behavioral testing fields to implement in their own laboratories. Individual scientists' productivity may be greatly enhanced by the sharing of common tools. This will benefit both the developers of this project and the recipients of our development efforts.

Design Criteria

With these goals in mind, we have decided on a number of design criteria. The system should rely, as much as possible, on standard, off-the-shelf components so that all users may take advantage of competitive pricing as well as simplified maintenance. The software we create should be made available to researchers at little or no cost. We will probably make it available via the Internet to reduce our effort and expenses. The software should be made available in source code as well as executable format because the end users will be scientists who need to know the exact details of the experiments they are running. Scientists may also need to extend and modify the system for their own experimental paradigms or hardware configurations. The system should be designed in such a way so that it can be implemented inexpensively for simple frequently used experimental paradigms. However, as the sophistication of the experiments increases, the cost of additional equipment will increase as well. For example, a researcher will be able to use this system to inexpensively setup identification or discrimination experiments in speech perception. Whereas the software and hardware additions necessary to run complex cross-modal auditory/visual paradigms with precise timing will be substantially more expensive.

General Guidelines

Software

Modular software design is essential to insure well written code, ease of support, ease of maintenance, portability and reusability of our libraries. The software library will have at a minimum capabilities analogous to those of the existing RBC calls (see Forshee, 1975). However, modified or new routines might have to be added to accommodate new capabilities.

The experimental paradigms could be written in any Microsoft link-standard language (e.g., QuickC or QuickBasic). As a long term goal, software will be developed under a graphical user interface (GUI).

Hardware

We will attempt to keep abreast of the rapidly expanding availability of commercial technology. However, due to funding constraints, this may not always be possible. It may be necessary to have separate hardware configurations for routine experiments and advanced experiments. One reason for this is that it might not be necessary to duplicate configurations of advanced experiments that might require expensive devices or custom built attachments.

The base-level system will be a single PC. Multiple subject stations will be based on a PC-based client-server configuration, with Ethernet inter-connection. The goal will be to permit the individual client PC's to run completely independently for single-subject experiments and groups of PCs to be synchronized for multiple-subject experiments.

Live-video program sources will be provided by laser disc and permit either disc audio source, or computer D/A synchronized audio output.

System Requirements

System Configuration

The ideal configuration will consist of a Master PC (server) with read/write CD-ROM and a DAT backup device. Storage capacity will be large (on the order of Gigabytes), with Ethernet connections to the subject stations. It will also have a single PC (client) at each subject station. The subject stations will have a hard disk with a capacity of at least 100 Meg. The Master PC will have the ability to switch from wide to local TCP/IP connectivity. See Figure 1.

Audio I/O

The audio system will be compatible with CD standards which use a sampling rate of 44.1 kHz and an amplitude resolution of 16 bits for each of two channels. The system will be capable of continuous output from disk, but with at least 10 seconds of single of onboard buffering. Simultaneous A/D during D/A will also be supported.

Response Latencies and Feedback

We will take advantage of the personal computer to utilize diverse forms of input and feedback; such as monitors, keyboards, mice and touch-screens. Additional devices such as response boxes and voice activation boxes (VOX) will also be needed for flexible data collection.

Our current testing stations employ response boxes with up to 7 buttons. Generally, fewer buttons (e.g., 2) have been used in identification and discrimination experiments. More buttons (typically all seven) have been used in rating experiments. The system currently under development will employ responses boxes with a maximum of eight buttons. This turns out to be a convenient power of two, and more buttons do not increase the precision of rating experiments. Button response will be signaled on the down press only, with a short travel distance. Button size is not critical, but will be consistent. All hardware must support measurement of response times with an accuracy of one millisecond. For multiple responses during trial, an interrupt driven "time-stamp" technique will be employed. In experiments where latency is measured, keyboard responses will not be acceptable due to small,

unpredictable variations in response time. Moreover, mouse, voice operated switches, and touch-screen inputs will never be used to measure latency.

The VOX will not be acceptable for latency measures. It will be acceptable for control only, since accuracy is well outside an order of 1 ms. We also plan to have touch screens available in all the subject booths under the same criteria.

We will use a standard "super VGA" video monitor with refresh rates on the order of 10 to 20 ms for visual stimulus presentation and response feedback. We will attempt to synchronize the display with the response timing routines in order to improve reaction time accuracy. However, we currently do not know whether this will turn out to be possible.

Video disc

Multimedia capabilities of personal computers are changing daily. We will monitor these developments and implement the most reasonable system as late in the design cycle as possible. It seems that some packages may offer the type of flexibility and provisions requested. A 10 to 20 msec accuracy would be sufficient for latency control and D/A Audio should be synchronized to within 15 ms of video.

Communications/Network

Ethernet boards will be used to connect the master PC to workstations (see "System Configuration" section). A bridge and/or router should be added to alleviate or stop traffic from the network.

The master server would feed all of the appropriate data files and programs to its clients. Each single-subject workstation should be able to run an experiment without the server. Although, in some case there might be a need to synchronize blocks among all single-subject workstations. Bloc-by-block synchronization accuracy should be preserved across the network.

Control Software

Software representing the experimental paradigm will be developed under DOS and Microsoft Windows 3.1 and NT (or "Chicago") as they become available. Subject's instructions, training, and feedback might be developed under Windows 3.1.

DOS and Windows will work together so that the controlling program may "switch off" the GUI and gain access to the low-level control of DOS if necessary (e.g., certain software directly addresses IO and is incompatible with Windows).

We will establish a support library of high-level-language (HLL) callable routines, using the MicroSoft link standard. Applications written in Assembly, BASIC, C, FORTRAN or Pascal could all run under the same environment. Experiment control programs will be written in the MicroSoft "QuickX" family (i.e. QuickC and QuickBasic).

A management utility will be installed to set up subject stations, verify accuracy, and download data from Master PC, and monitor the progress of the experiment with statistics. Since the Master PC is only for management and control, it would be possible to run single subject experiments on individual subject stations.

See Appendix A for an outline of proposed hardware specifications.

Project Structure and Outline

Short-term Plans

In the short term, we will attempt to duplicate and slightly improve the existing experiment control facilities of the PDP-11/34 and LABLIB (Forshee, 1979) on a 386 PC. However, we will not duplicate the multiple subject capabilities of that system right away. We have divided the project into several phases.

Phase 1. The first phase of this project is straightforward. The goal will be to use a modern 386 personal computer to run simple identification and discrimination experiment control programs. The programs will run under DOS and the Data Translation analog-to-digital software. We will attempt to literally duplicate the necessary LABLIB routines in C from Forshee's RBC library routines. The final product at this phase will have the same timing and audio capabilities as the current PDP-11/34s. The hardware will consist of the PC itself, a Data Translation board for audio output, a response collection button box response system developed by the Psychology Department Electronic Shop, and a Genus video control software for text and graphic display. During this phase timing issues, hardware, and experiment control capabilities and deficiencies will be evaluated.

Phase 2. The goals for this phase will be the same as for Phase 1 with the exception that better and more accessible hardware will be used. The Gravis Ultrasound card will be used for audio output. Precisely timed display of text will be added at this point. These changes will improve the usability of the system and test the modularity of the code written in Phase 1.

Phase 3. At this point, the Phase 2 system will be used to test the accuracy of the system. Proof of performance guidelines will be written and carried out. Testing of timing, synchronization, acoustic stimulus quality, video timing and quality, and overall performance will be performed.

By the end of Phase 3, we will be able to make some concrete decisions about hardware and software. We will also be able to evaluate our assumptions and consider any unexpected outcomes.

Mid-term Plans

After we have duplicated our current capabilities in single-subject mode on a 386 platform we will modify the system to handle multiple-subject experiment sessions. The overall strategy will be to use one PC per testing station with each PC networked to a central "experiment server" PC. The central PC will coordinate the action of the testing station PCs on a block-by-block basis. It will also perform data and stimulus management and backup functions. The purpose of central control, as opposed to simply running a bunch of independent PC subject stations, is to make the experimenter's life easier. For example, facilities will be included to run multiple subjects in synchrony for simpler experiments.

Phase 1. Port experiment control system to a client/server model. Independent PCs should recognize that they are connected to a controller, and if they are, request services from that controller. If they are not, they should run the experiment locally in single-subject mode. In this phase, the subject station PCs should ask the server which experiment to run, which stimuli to use, where to put the data, and when to begin each block. Ideally communication will be via TCP/IP over 10-baseT Ethernet at this point, but we may decide to slip back to serial communication as a temporary measure.

Phase 2. Develop block-by-block data central data backup. Develop an experiment logbook database. Finalize network strategy.

Phase 3. Beta test of experiment control system. Debug and revise system based on conducting several real experiments.

Long-term Plans

The long-term plans consist of adding capabilities to the basic system. These include: (1) Synchronized audio and video (perhaps via Quicktime-like software if performance is acceptable), (2) Multiple reaction times within a trial, (3) Laser disk video control, and (4) Video mixing within Windows.

Current Status of System

At the present time we are in Phase 2 of the short term plan and our expectations have been met. We are nearly ready to start testing the accuracy of the system. After testing is completed we will be in a better position to adjust or replace software and hardware to match our system requirement. In an effort to rely more on off-the-shelf components, we are constantly investigating different hardware configurations, but at this stage we are concentrating our efforts on the accuracy of the system.

Our current PC configuration consists of a IBM-compatible 486 CompuAdd model 433e, 16MB RAM and 600MB hard disk. Non-interlace SVGA monitor. UltraSound Advance Gravis D/A board. A Metrabyte PIO-12 digital I/O interface board is also installed as part of the response box built at the Indiana University psychology electronics department.

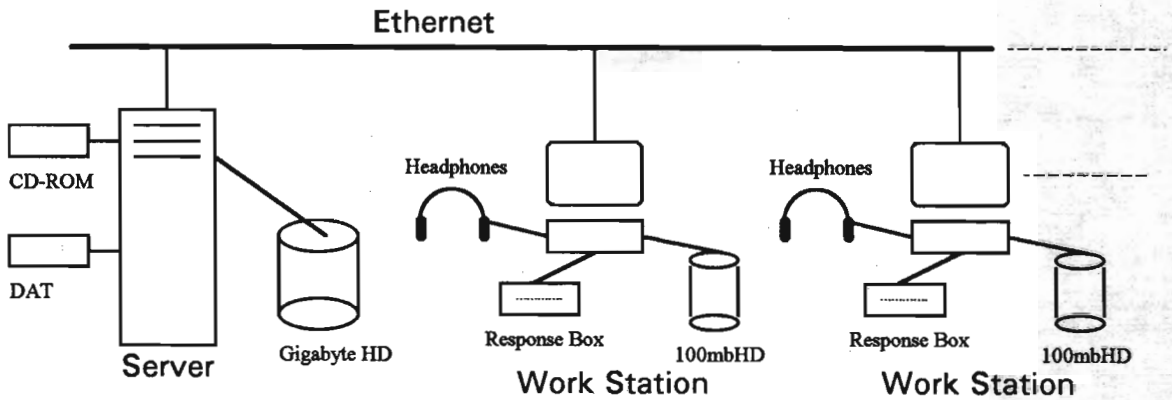


Figure 1. Local Area Network Configuration for PC based experimental system.

References

Forshee, J. C. (1975). Speech perception research laboratory: The state of the computer system. *Research on Speech Perception Progress Report No. 2*, Indiana University, Bloomington, IN.

Forshee, J. C. (1979). Speech perception research laboratory: The state of the computer system. *Research on Speech Perception Progress Report No. 5*, Indiana University, Bloomington, IN.

Forshee, J. C., & Nusbaum, H. C. (1984). An update on computer facilities in the Speech Research Laboratory. *Research on Speech Perception Progress Report No. 10*, Indiana University, Bloomington, IN.

Appendix A

Proposed Hardware Specifications

A/D/A
Under consideration: UltraSound Advanced Gravis, MediaVision, Microsoft Sound.

- Compatible with CD music standards
- Sample rate of 44.1 Khz, two channels with over sampling D/A
- On-board anti-alias filters
- Onboard buffer RAM of at least 256k bytes
- 16 bit resolution, > 80dB S/N
- Standard Audio "line-out" & "line-in"
- ISA bus compatible

Response Boxes
Under consideration: IU Psychology Dept. Electronic Shop response box

- Quiet
- Good tactile feel (ideal: HP-calculator-like buttons, not PC-keyboard-like)
- Register response on down only
- Travel distance on the order of a few mm, with non-linear pressure (press hard and fall through to end).
- Gold contacts, sealed, with debouncing.
- Size approximately 3/8" and consistent.
- Spacing should be approx. 1 inch.
- No "smarts" in button box. Use just switches and lights. No need for "fancier" feedback like LCDs.
- Touch screens on CRT display

Feedback Lamp

- Mechanically robust
- Low power, high efficiency
- at least 100 mcd
- simple and rapid replacement

Digital Input

- Latched level sensitive inputs
- Capable of maskable interrupt generation on bit change
- 8 bits or more
- Onboard timer chip (if available)

Timer

- resolution to 1 ms

- interrupt capability
- ease of programming
- Latched digital input lines (if available)

Video monitors
Under consideration: Sony Trinitron Models

- SVGA (1024x768)
- .28 dot pitch.
- Vertical refresh rate of 72 Hz or greater, non-interlaced.
- Horizontal, 55 Khz (18 microsecond refresh).
- trinitron to avoid convergence problem. (Tom Carrell)

Network
Under consideration: 3Com Ethernet boards
- thin wire or twisted pair Ethernet
- 10-base-T

IV. Publications

Papers Published:

Duffy, S.A., & Pisoni, D.B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, **35**, 351-389.

Goldinger, S.D., Luce, P.A., Pisoni, D.B., & Marcario, J.K. (1992). Form-based priming in spoken word recognition: Roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18** (6), 1210-1237.

Goldinger, S.D., Palmeri, T.J., & Pisoni, D.B. (1992). Words and voices: Perceptual details are preserved in lexical representations. In J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, & G.E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing*. Alberta, Canada: University of Alberta.

Jusczyk, P.W., Pisoni, D.B., & Mullennix, J.W. (1992). Effects of talker variability on speech perception by 2-month-old infants. *Cognition*, **43**, 253-291.

Lively, S.E., Pisoni, D.B., & Logan, J.S. (1992). Some effects of training Japanese listeners to identify English /r/ and /l/. In Y. Tohkura (Ed.), *Speech perception, production and linguistic structure* (pp. 175-196). Tokyo: Ohmsha Publishing Co. Ltd.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1992). Effects of speaking rate and talker variability on the representation of spoken words in memory. In J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, & G.E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing*. Alberta, Canada: University of Alberta.

Pisoni, D.B. (1992). Some comments on talker normalization in speech perception. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.) *Speech perception, production and linguistic structure* (pp. 143-151). Tokyo: Ohmsha Publishing Co. Ltd.

Pisoni, D.B. (1992). Some comments on invariance, variability and perceptual normalization in speech perception. In J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, & G.E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing*. Alberta, Canada: University of Alberta.

Sommers, M.S., Nygaard, L.C., & Pisoni, D.B. (1992). Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. In J.J. Ohala, T.M. Nearey, B.L. Derwing, M.M. Hodge, & G.E. Wiebe (Eds.), *Proceedings 1992 International Conference on Spoken Language Processing*. Alberta, Canada: University of Alberta.

Manuscripts Accepted for Publication (In Press):

Goldinger, S.D., Pisoni, D.B., & Luce, P.A. (In Press). Speech perception: Research and theory. In N.J. Lass (Ed.) *Principles of experimental phonetics*. Toronto, Canada: B.C. Decker.

Humes, L.E., Nelson, K.J., Pisoni, D.B., & Lively, S.E. (In Press). Effects of age and hearing loss on serial recall of natural and synthetic speech. *Journal of Gerontology*.

Lewellen, M.J., Goldinger, S.D., Pisoni, D.B., & Greene, B.G. (In Press). Word familiarity and lexical fluency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General*.

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

Lively, S.E., Pisoni, D.B., & Goldinger, S.D. (In Press). Spoken word recognition: Research and Theory. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics,* New York: Academic Press.

Lively, S.E., Pisoni, D.B., Summers, W.V., & Bernacki, R.H. (In Press). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *Journal of the Acoustical Society of America.*

Logan, J.S., Lively, S.E., & Pisoni, D.B. (In Press). Training listeners to perceive novel phonetic categories: How do we know what is learned? *Journal of the Acoustical Society of America.*

Mullennix, J.W., Goldinger, S.D., & Pisoni, D.B. (In Press). Some characteristics of talker normalization. In J. Charles-Luce, P.A. Luce, & J.R. Sawusch (Eds.) *Theories in spoken language: Perception, production and development*. Norwood, NJ: Ablex.

Nygaard, L.C. (In Press). Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. *Journal of Experimental Psychology: Human Perception and Performance.*

Nygaard, L.C., & Pisoni, D.B. (In Press). Speech perception: New directions in research and theory. In J.L. Miller & P.D. Eimas (Eds.), *Handbook of perception and cognition: Volume 11 Speech, language and communication.*

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

Pisoni, D.B. (In Press). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. In K. Hirose, S. Kiritani & G. Fant (Eds.) *Festschrift in honor of Hiroya Fujisaki*. Amsterdam: Elsevier North-Holland.

Pisoni, D.B., & Lively, S.E. (In Press). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research.* Timonium, MD: York Press.

Pisoni, D.B., Logan J.S., & Lively, S.E. (In Press). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In J. Goodman & H.C. Nusbaum (Eds.). *Development of speech perception: The transition from recognizing speech sounds to spoken words*. Cambridge: MIT Press.

Ralston, J.V., Pisoni, D.B., & Mullennix, J.W. (In Press). Comprehension of synthetic speech produced by rule. In R. Bennett, A. Syrdal, & S. Greenspan (Eds.) *Behavioral aspects of speech technology: Theory and applications*. New York: Elsevier.