Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning *

David B. Pisoni
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, USA

Abstract

This paper 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 strong evidence for 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.

Keywords

1. Introduction

My research in the early 1970's was directly motivated by Hiroya Fujisaki's proposal of the differential roles of auditory and phonetic memory codes in the perception of consonants and vowels (Fujisaki and Kawashima, 1969). The studies that I carried out at that time

---

*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 and 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. This paper is dedicated to Professor Fujisaki, one of the great pioneers in the field of speech research and spoken language processing.

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 a specialized perceptual mode for speech sounds that was distinct from other perceptual systems (Liberman et al., 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 (Pisoni and Luce, 1987). 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 et al., 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 et al., 1992a; Nygaard et al., 1992a). 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, 1992a).

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 as the "personal" characteristics of speech – to be independent of the linguistic content of the talker's message (Ladefoged, 1975; Laver and 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.

2. Experiments on talker variability in speech perception

A series of novel experiments has 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, 1992b). Our research on talker variability began with the observations of Mullennix et al. (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.

In a second experiment, we measured naming latencies to the same words presented in both test conditions (Mullennix et al., 1989). Table 1 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.

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 and 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.

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 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. 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 suggesting that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory.

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 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.

3. 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 et al. (1992b) 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.

Other experiments on serial recall have also been completed to examine the encoding and representation of speaking rate in memory. Nygaard et al. (1992b) 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., 1992b). 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.

4. Experiments on variability in perceptual learning

We have always maintained a strong interest in issues surrounding perceptual learning and development in speech perception (Aslin and Pisoni, 1980; Walley et al., 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 et al., 1991; Lively et al., 1992, 1993; Strange and 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 and 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 and 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 (Lively et al., 1992). 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 et al., in press). 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.

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.

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 and 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.

5. Abstractionist versus 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 and 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 and 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 and Brooks, 1984; Brooks, 1978; Tulving and Schacter, 1990; Schacter, 1990). According to this approach, stimulus variability is considered to be "lawful" and informative to perceptual analysis (Elman and McClellan, 1986). Memory involves encoding specific instances, as well as the processing operations used in recognition (Kolers, 1973, 1976). 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 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 and Brooks, 1984; Brooks, 1978). These criteria which are summarized below can be applied directly to speech perception and spoken language processing.

5.1. 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.

5.2. 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.

5.3. 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.

5.4. 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. Moreover, 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.

5.5. Three domains of speech

Among category systems, speech appears to be unusual 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.

6. 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 (Jacoby and Brooks, 1984; 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 or fragment completion, 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 suggests 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 and Trudgill, 1979). While there have been a number of recent studies on explicit voice recognition and identification by human listeners (Papcun et al., 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., in press). 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 damage can affect the perception of familiar and novel voices in very different ways (Kreiman and VanLancker, 1988; VanLancker et al., 1988, 1989).

Most researchers working in speech perception adopted a common set of assumptions about the units of linguistic analysis and the goals of speech perception. 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 suggests 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 and McClellan, 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 simply 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.

7. Conclusions

The results summarized in this paper 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 novel 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 is 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.

Acknowledgments

References

Aslin, RN.; Pisoni, DB. Some developmental processes in speech perception. In: Yeni-Komshian, G.; Kavanagh, JF.; Ferguson, CA., editors. Child Phonology: Perception and Production. New York: Academic Press; 1980. p. 67-96.

Brooks, LR. Nonanalytic concept formation and memory for instances. In: Rosch, E.; Lloyd, B., editors. Cognition and Categorization. Hillsdale, NJ: Erlbaum; 1978. p. 169-211.

Creelman CD. Case of the unknown talker. J. Acoust. Soc Amer. 1957; Vol. 29:655.

Eich JE. A composite holographic associative memory model. Psychol Rev. 1982; Vol. 89:627–661.

Elman, JL.; McClellan, JL. Exploiting lawful variability in the speech wave. In: Perkell, JS.; Klatt, DH., editors. Invariance and Variability in Speech Processes. Hillsdale, NJ: Erlbaum; 1986. p. 360-385.

Fowler, CA. Listener–talker attunements in speech. In: Tighe, T.; Moore, B.; Santroch, J., editors. Human Development and Communication Sciences. Hillsdale, NJ: Erlbaum; (in press)

Fujisaki, H.; Kawashima, T. Annual Report of the Engineering Research Institute. Vol. Vol. 28. Tokyo: Faculty of Engineering, University of Tokyo; 1969. On the modes and mechanisms of speech perception; p. 67-73.

Goldinger, SD. Research on Speech Perception Technical Report No. 7. Bloomington, IN: Indiana University; 1992. Words and voices: Implicit and explicit memory for spoken words.

Goldinger SD, Pisoni DB, Logan JS. On the locus of talker variability effects in recall of spoken word lists. J. Exp. Psychol: Learning, Memory, and Cognition. 1991; Vol. 17(No. 1):152–161.

Hintzman DL. Schema abstraction in a multiple-trace memory model. Psychol Rev. 1986; Vol. 93:411–423.

Jacoby, LL.; Brooks, LR. Nonanalytic cognition: Memory, perception, and concept learning. In: Bower, G., editor. The Psychology of Learning and Motivation. New York: Academic Press; 1984. p. 1-47.

Jusczyk, PW. Infant speech perception and the development of the mental lexicon. In: Nusbaum, HC.; Goodman, JC., editors. The Transition from Speech Sounds to Spoken Words: The Development of Speech Perception. Cambridge, MA: MIT Press; 1993.

Kakehi, K. Adaptability to differences between talkers in Japanese monosyllabic perception. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. Speech Perception, Production and Linguistic Structure. Tokyo: IOS Press; 1992. p. 135-142.

Klatt DH. Speech perception: A model of acoustic–phonetic analysis and lexical access. J. Phonet. 1979; Vol. 7:279–312.

Klatt, DH. The problem of variability in speech recognition and in models of speech perception. In: Perkell, JS.; Klatt, DH., editors. Invariance and Variability in Speech Processes. Hillsdale, NJ: Erlbaum; 1986. p. 300-324.

Kolers PA. Remembering operations. Memory & Cognition. 1973; Vol. 1:347–355.

Kolers PA. Pattern analyzing memory. Science. 1976; Vol. 191:1280–1281. [PubMed: 1257750]

Kreiman J, VanLancker D. Hemispheric specialization for voice recognition: Evidence from dichotic listening. Brain & Language. 1988; Vol. 34:246–252. [PubMed: 3401693]

Ladefoged, P. A Course in Phonetics. New York: Harcourt Brace Jovanovich; 1975.

Laver, J.; Trudgill, P. Phonetic and linguistic markers in speech. In: Scherer, KR.; Giles, H., editors. Social Markers in Speech. Cambridge: Cambridge University Press; 1979. p. 1-31.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychol Rev. 1967; Vol. 74:431–461. [PubMed: 4170865]

Lively, SE.; Pisoni, DB.; Logan, JS. Some effects of training Japanese listeners to identify English /r/ and /l/. In: Tohkura, Y., editor. Speech Perception, Production and Linguistic Structure. Ohmsha, Tokyo: 1992. p. 175-196.

Lively, SE.; Pisoni, DB.; Yamada, RA.; Tohkura, Y.; Yamada, T. Research on Speech Perception, Progess Report No. 18. Indiana University; 1992. Training Japanese listeners to identify English |r| and |l|: III. Long-term retention of new phonetic categories; p. 185-216.

Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. J. Acoust. Soc. Amer. 1993

Logan JS, Lively SE, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/: A first report. J. Acoust. Soc. Amer. 1991; Vol. 89(No. 2):874–886. [PubMed: 2016438]

Martin CS, Mullennix JW, Pisoni DB, Summers WV. Effects of talker variability on recall of spoken word lists. J. Exp. Psychol.: Learning, Memory and Cognition. 1989; Vol. 15:676–684.

Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics. 1990; Vol. 47(No. 4):379–390. [PubMed: 2345691]

Mullennix JW, Pisoni DB, Martin CS. Some effects Of talker variability on spoken word recognition. J. Acoust. Soc Amer. 1989; Vol. 85:365–378. [PubMed: 2921419]

Neisser, U. Cognitive Psychology. New York: Appleton–Century–Crofts; 1976.

Nygaard, LC.; Sommers, MS.; Pisoni, DB. Effects of speaking rate and talker variability on the representation of spoken words in memory. Proc. 1992 Internal. Conf. Spoken Language Processing; Banff, Canada. 1992a. p. 209-212.

Nygaard LC, Sommers MS, Pisoni DB. Effects of speaking rate and talker variability on the recall of Spoken words. J. Acoust. Soc Amer. 1992b; Vol. 91(No. 4):2340.

Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. Psychol. Sci. (in press).

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. J: Exp. Psychol: Learning, Memory and Cognition. 1993; Vol. 19(No. 2):1–20.

Papcun G, Kreiman J, Davis A. Long-term memory for unfamiliar voices. J. Acoust. Soc. Amer. 1989; Vol. 85:913–925. [PubMed: 2926007]

Peters, RW. Joint Project Repon No. 56. Pensacola, FL: U.S. Naval School of Aviation Medicine; 1955. The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise; p. 1-9.

Pisoni DB. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception & Psychophysics. 1973; Vol. 13:253–260.

Pisoni, DB. Speech perception. In: Estes, WK., editor. Handbook of Learning and Cognitive Processes. Vol. Vol. 6. Hillsdale, NJ: Erlbaum; 1978. p. 167-233.

Pisoni, DB. Effects of talker variability on speech perception: Implications for current research and theory. Proc. 1990 Internal. Conf. Spoken Language Processing; Kobe, Japan. 1990. p. 1399-1407.

Pisoni, DB. Some comments on invariance, variability and perceptual normalization in speech perception. Proc. 1992 Internal. Conf. Spoken Language Processing; Banff, Canada. 1992a. p. 587-590.

Pisoni, DB. Some comments on talker normalization in speech perception. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. Speech Perception, Production and Linguistic Structure. Tokyo: IOS Press; 1992b. p. 143-151.

Pisoni, DB.; Luce, PA. Speech perception: Research, theory, and the principal issues. In: Schwab, EC.; Nusbaum, HC., editors. Pattern Recognition by Humans and Machines. New York: Academic Press; 1986. p. 1-50.

Pisoni DB, Luce PA. Acoustic-phonetic representations in word recognition. Cognition. 1987; Vol. 25:21–52. [PubMed: 3581727]

Pisoni DB, Nusbaum HC, Luce PA, Slowiaczek LM. Speech perception, word recognition and the structure of the lexicon. Speech Communication. 1985; Vol. 4(Nos. 1–3):75–95.

Posner, MI. Abstraction and the process of recognition. In: Spence, JT.; Bower, GH., editors. The Psychology of Learning and Motivation: Advances in Learning and Motivation. New York: Academic Press; 1969. p. 43-100.

Posner M, Keele S. On the genesis of abstract. J. Exp. Psychol. 1986; Vol. 77:353–363.

Roediger HL. Implicit memory: Retention without remembering. Amer. Psychol. 1990; Vol. 45(No. 9):1043–1056. [PubMed: 2221571]

Schacter, DL. Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. In: Diamond, A., editor. Development and Neural Basis of Higher Cognitive Function. Vol. Vol. 608. Ann. New York Acad. Sci.; 1990. p. 543-571.

Schacter DL. Understanding implicit memory: A cognitive neuroscience approach. Amer. Psychol. 1992; Vol. 47(No. 4):559–569. [PubMed: 1595984]

Sommers, MS.; Nygaard, LC.; Pisoni, DB. Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude. Proc. 1992 Int. Conf. Spoken Language Processing; Banff, Canada. 1992a. p. 217-220.

Sommers MS, Nygaard LC, Pisoni DB. The effects of speaking rate and amplitude variability on perceptual identification. J. Acoust. Soc. Amer. 1992b; Vol. 91(No. 4):2340.

Stevens, KN. Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. Proc 7th Int. Cong. Phonetic Sciences; Mouton, The Hague. 1971. p. 206-232.

Stevens, KN. The quantal nature of speech: Evidence from articulatory acoustic data. In: David, EE., Jr; Denes, PB., editors. Human Communication A Unified View. New York: McGraw-Hill; 1972. p. 51-66.

Strange W, Dittmann S. The effects of discrimination training on the perception of /r–l/ by Japanese adults learning English. Perception & Psychophysics. 1984; Vol. 36(No. 2):131–145. [PubMed: 6514522]

Studdert-Kennedy, M. The perception of speech. In: Sebeok, TA., editor. Current Trends in Linguistics. Mouton, The Hague: 1974. p. 2349-2385.

Studdert-Kennedy, M. Speech perception. In: Lass, NJ., editor. Contemporary Issues in Experimental Phonetics. New York: Academic Press; 1976. p. 243-293.

Studdert-Kennedy M. Speech perception. Language and Speech. 1980; Vol. 23:45–66. [PubMed: 6999263]

Studdert-Kennedy M. On learning to speak. Hum. Neurobiol. 1983; Vol. 2:191–195. [PubMed: 6668235]

Tulving E, Schacter DL. Priming and human memory systems. Science. 1990; Vol. 247:301–306. [PubMed: 2296719]

VauLancker DR, Cummings JL, Kreiman J, Dobkin B. Phonagnosia: A dissociation between familiar and unfamiliar voices. Cortex. 1988; Vol. 24:195–209.

VanLancker DR, Kreiman J, Cummings J. Voice perception deficits: Neuroanatomical correlates of phonagnosia. J. Clin. Exp. Neuropsychol. 1989; Vol. 11(No. 5):665–674.

Walley, AC.; Pisoni, DB.; Aslin, RN. The role of early experience in the development of speech perception. In: Aslin, RN.; Alberts, J.; Petersen, MR., editors. The Development of Perception: Psychobiological Perspectives. New York: Academic Press; 1981. p. 2119-2255.
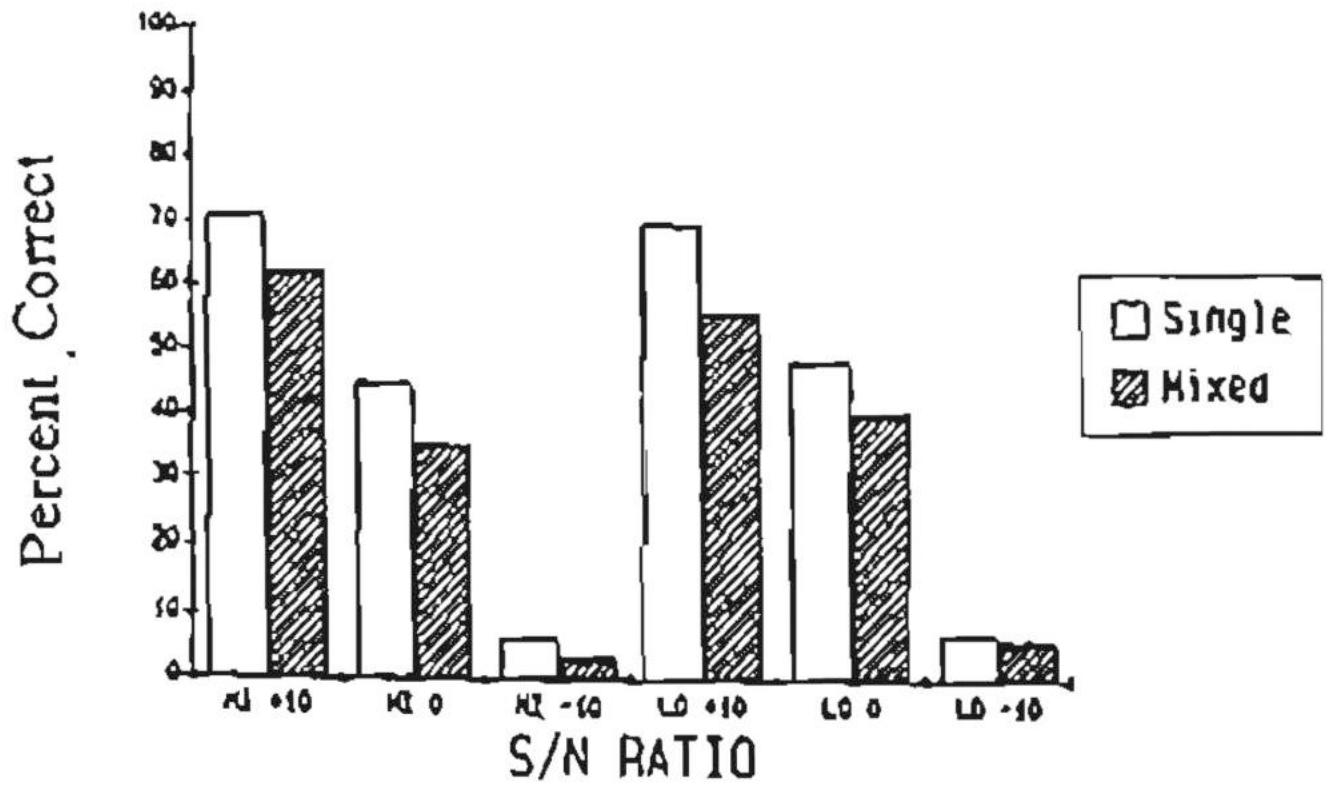
Fig. 1.
Overall mean percent correct performance collapsed over subjects for single- and mixed-talker conditions as a function of high- and low-density words and signal-to-noise ratio (from (Mullennix et al., 1989)).
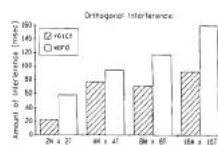
Fig. 2.
The amount of orthogonal interference (in milliseconds) across all stimulus variability
conditions as a function of word and voice dimensions (from (Mulleonix and Pisoni, 1990)).
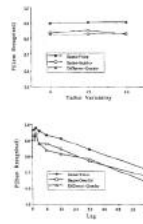
Fig. 3.
Probability of correctly recognizing old items in a continuous recognition memory experiment. In both panels, recognition for same-voice repetitions is compared to recognition for different-voice/same-gender and different-voice/different-gender repetitions. The upper panel displays item recognition 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)).
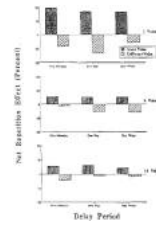
Fig. 4.
Net repetition effects observed in perceptual identification as a function of delay between sessions and repetition voice (from (Goldinger, 1992)).
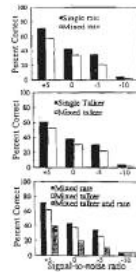
Fig. 5.
Effects of talker, rate, and combined talker and rate variability on perceptual identification (from (Sommers et al., 1992b)).
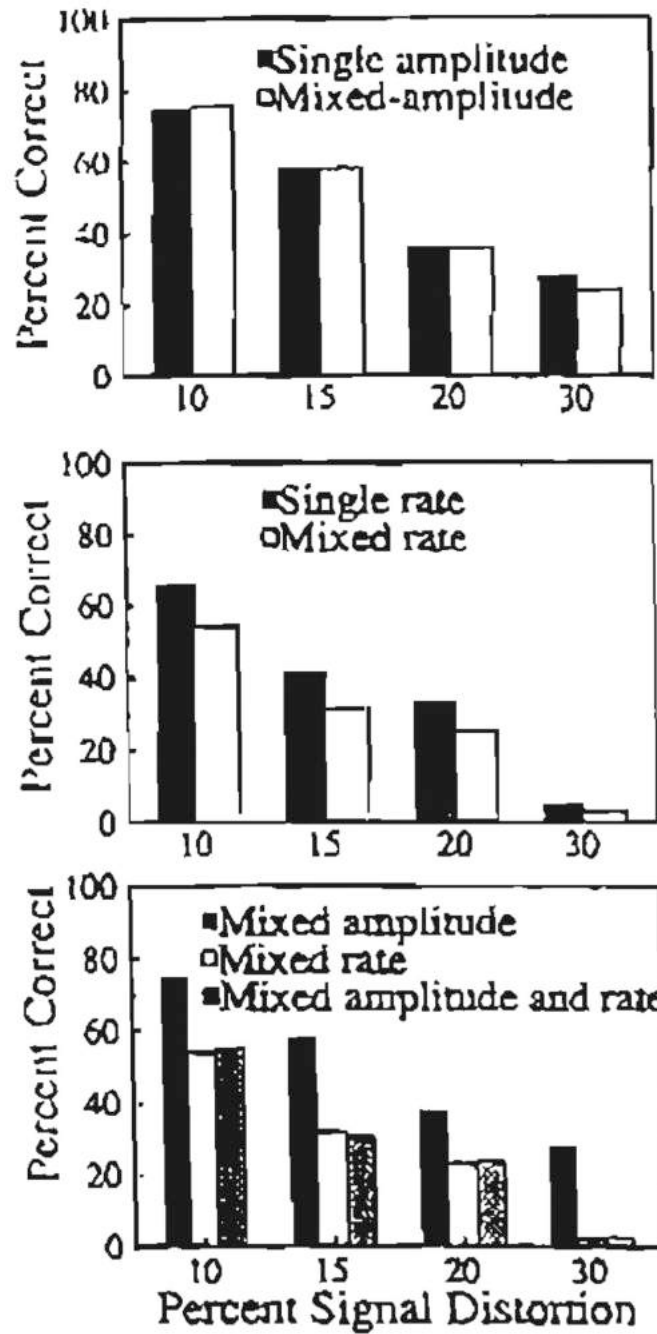
Fig. 6.
Effects of amplitude, rate, and combined amplitude and rate variability on perceptual identification (from (Sommers et al., 1992)).
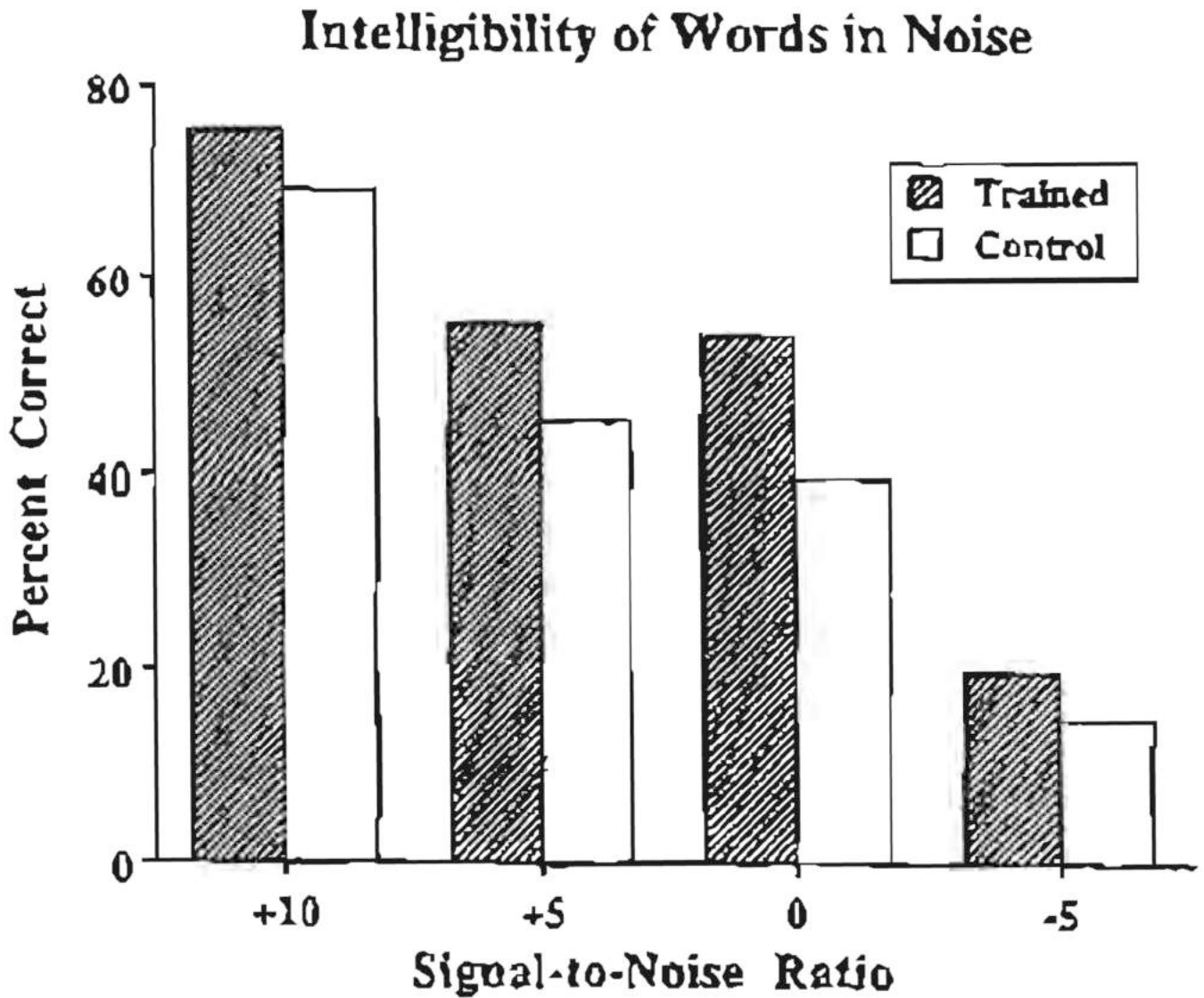
Intelligibility of Words in Noise

Fig. 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., 1992b)).

Table 1

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