

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 21 (1996-1997)
Indiana University

**Effects of Talker, Rate and Amplitude Variation
on Recognition Memory for Spoken Words¹**

Ann R. Bradlow², Lynne C. Nygaard³ and David B. Pisoni

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

¹ This research was supported by NIH-NIDCD Training Grant DC-00012 and NIH-NIDCD Research Grant DC-00111 to Indiana University. We are grateful to Luis Hernandez for technical support and to Thomas Palmeri for programming assistance. An earlier version of this study was presented at the 131st meeting of the Acoustical Society of America in Indianapolis, IN, May, 1996.

² Now at Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL.

³ Now at Department of Psychology, Emory University, Atlanta, GA.

Effects of Talker, Rate and Amplitude Variation on Recognition Memory for Spoken Words

Abstract. This study investigated the encoding of spoken words using a continuous recognition memory task. In Experiment 1, subjects judged whether each word in a list of spoken words was "old" (had occurred previously in the list) or "new." Subjects were more accurate at recognizing a word as "old" if it was repeated in the same voice, and at the same speaking rate; however, there was no recognition advantage for words repeated at the same overall amplitude. In Experiment 2, if subjects judged a word as "old" they were then required to provide an additional explicit judgment as to whether it was repeated in the same voice, rate, or amplitude. Subjects again showed an advantage in recognition memory for words repeated in the same voice and same speaking rate, but no advantage occurred for the amplitude condition. However, in all three conditions, subjects were able to detect whether an "old" word was repeated in the same voice, rate or amplitude. These data suggest that information about all three properties of spoken words is encoded and retained in memory and can be used in recognition tasks requiring explicit judgments.

Introduction

A growing body of research has begun to identify the effects of stimulus variability on a variety of speech perception and spoken word recognition tasks (e.g., Mullennix et al., 1989; Sommers et al., 1994). Other studies have also shown effects of stimulus variability on memory for spoken words (e.g., Martin et al., 1989; Goldinger et al., 1991; Palmeri et al., 1993; Nygaard et al., 1995; for reviews see Pisoni, 1993; 1997). These findings represent a novel approach to the long-standing issue of "perceptual constancy" in the face of a highly variable speech signal. Rather than actively seeking acoustic, articulatory or relational invariants that are supposed to guide the listener in accessing phoneme- and ultimately word-sized units (e.g., Joos, 1948; Ladefoged and Broadbent, 1957; Stevens and Blumstein, 1978; Kewley-Port, 1983; Halle, 1985; Nearey 1989; Johnson, 1990 and many others), this research directly investigates the effects of various sources of stimulus variability in the test materials. The general orientation of this research regards the inherent variability in the speech signal due to different talker- and other instance-specific characteristics as a useful source of information to the listener about the communicative situation (Laver, 1989; Laver and Trudgill, 1979), rather than a source of "noise" in the signal that is "stripped away" by the processes of speech perception and spoken word recognition (see Pisoni, 1997).

With respect to spoken word recognition, Mullennix et al. (1989) showed that word recognition accuracy decreased and response times increased when subjects were presented with lists of words produced by multiple talkers relative to a condition where subjects were presented with the identical words produced by only a single talker. Sommers et al. (1994) replicated this result with a different set of words and talkers. Additionally, in an attempt to understand the nature of the talker-variability effect found by Mullennix et al. (1989), Sommers et al. (1994) also investigated the effects of speaking-rate and overall amplitude variability on word recognition. The results of Sommers et al. (1994) replicated the findings reported by Mullennix et al. (1989). They showed a decrease in word identification scores for mixed-talker lists relative to single-talker lists. Furthermore, Sommers et al. (1994) showed a comparable decrease in word identification scores for mixed-rate lists relative to single-rate lists, but no decrease in word identification scores for mixed-amplitude lists relative to single-amplitude lists. These findings indicated that all sources of variability in the test materials do not produce similar effects on word recognition scores.

Sommers et al. (1994) suggested that the effects of talker and rate variability on word recognition may be due to the relevance of these dimensions for the perception of phonetic contrasts (Ladefoged and Broadbent, 1957; Miller, 1987). In contrast, variability in overall amplitude does not signal a phonetic contrast, and therefore variability along this dimension does not exert costly processing demands for word recognition.

With respect to memory for spoken words, Martin et al. (1989) found that subjects performed better in a serial recall task when the words within lists were produced by a single talker than when the words within each list were produced by multiple talkers. This difference in serial recall of spoken words was located in the primacy portion of the serial recall curve, that is, for the first three words in ten-word lists. Martin et al. (1989) proposed that this finding arose from the increased processing demands incurred by increased stimulus variability, and that these additional processing requirements interfered with subjects' abilities to maintain and rehearse information in working memory and to transfer this information to long-term memory.

Goldinger et al. (1991) investigated further the nature of talker variability effects on recall of spoken word lists by varying the rate of presentation of the items in the list to be recalled. Goldinger et al. (1991) hypothesized that rate of presentation would affect the subject's ability to encode the distinctive voice information for multiple-talker lists. If given enough rehearsal time, it was thought that subjects might be able to use the distinctive talker information as a retrieval cue, and thus the multiple-talker lists would be more accurately recalled than the single-talker lists. Indeed, Goldinger et al. (1991) found that at fast presentation rates (one word every 250 ms), words in the primacy portion of the single-talker lists were more accurately recalled than those from multiple-talker lists; whereas at slow presentation rates (one word every 4000 ms), this difference in recall accuracy was reversed. These results showed that information about a talker's voice is encoded and can be used as an effective retrieval cue under optimal conditions.

In a subsequent study, Nygaard et al. (1995) found that at fast presentation rates, items presented early in lists spoken either by a single talker or at a single speaking rate were better recalled than the same items spoken by multiple talkers or at multiple speaking rates, respectively. At a slow presentation rate, early items in the multiple-talker lists were better recalled than those in the single-talker lists; however, this reversal of recall accuracy was not obtained for the items in the multiple-rate lists relative to those in the single-rate lists. Rather, at the slow presentation rate, there was no difference between recall of items in the multiple- and single-rate lists. Furthermore, Nygaard et al. (1995) found no differences between serial recall of single- and multiple-amplitude lists at fast, as well as at slow presentation rates. Taken together, these results suggest that distinctive talker information is encoded in the long-term memory representation of spoken words, and if given sufficient rehearsal time, this additional distinctive information can be used as a retrieval cue by the listener. In contrast, the data from these serial recall experiments did not provide any evidence that either speaking rate or overall amplitude are encoded in long-term memory along with the linguistic content of a spoken word.

In a study of recognition memory for spoken words, Palmeri et al. (1993) found that detailed information about a talker's voice is retained in memory and facilitates recognition of a previously encountered word. Specifically, Palmeri et al. (1993) found that listeners were better at recognizing a word as a repeated item in a continuous list of spoken words when the word was repeated in the same voice that it was originally spoken in than when the voice differed from first to second repetition. Furthermore, Palmeri et al. (1993) showed that, when listeners recognized that the word was a repeated word in the list, they were also able to explicitly recognize whether the voice was the same or different as the first occurrence of the word.

Taken together, the findings of Goldinger et al. (1991), Nygaard et al. (1995) and Palmeri et al. (1993) have shown that specific talker characteristics can affect recall and recognition of spoken words. (See also Craik and Kirsner, 1974; Schacter and Church, 1992; Church and Schacter, 1994; Sheffert and Fowler (1995)). Furthermore, Nygaard et al. (1995) showed that variability in speaking rate can produce effects on the recall of spoken words, but that variability in overall amplitude does not. The purpose of the present study was to further investigate the role of different sources of variability in the encoding of spoken words in memory by comparing the effects of talker, rate and amplitude variability using a continuous recognition memory task. We hypothesized that a recognition memory task might be more sensitive in revealing the retention in long-term memory of stimulus dimensions such as speaking rate and overall amplitude than the serial recall task used by Nygaard et al. (1995) because a recognition task was thought to be less resource demanding than a recall task. Another goal of the present study was to provide additional data regarding the effects of different sources of variability on a variety of speech perception and word recognition tasks. Specifically, we wanted to know whether the distinct effects of talker, rate and amplitude variability on word identification found by Sommers et al. (1994) and by Nygaard et al. (1995) using serial recall tasks, would also be obtained in recognition memory. Thus, we hoped to be able to develop a more comprehensive understanding of the effects of different item-specific features on speech perception and spoken word recognition.

EXPERIMENT 1

Experiment 1 investigated whether subjects were more accurate at recognizing a word as "old" (i.e., had occurred previously in a list of spoken words) if it was repeated in the same voice (Condition 1), at the same speaking rate (Condition 2), and at the same amplitude (Condition 3). The voice condition was a replication of Palmeri et al. (1993); the rate and amplitude conditions were designed to extend the findings on voice to conditions where the stimuli incorporated other sources of variability.

Method

Subjects

One hundred and twenty students enrolled in undergraduate introductory psychology courses at Indiana University served as subjects. All subjects received partial course credit for their participation. All were native speakers of American English with no history of speech or hearing disorder at the time of testing.

Stimuli

The stimuli used in Experiment 1 came from a database of 200 words spoken by two talkers (one male and one female) at three different rates of speech (fast, medium, and slow). The words were selected from four 50-item phonetically balanced (PB) word lists (ANSI, 1971), and were originally recorded embedded in the carrier sentence, "Please say the word _____." For each rate of speech, the full set of 200 sentences was presented to the talkers in random order on a CRT screen located in a sound-attenuated booth (IAC 401A). Productions were monitored via a loudspeaker located outside the recording booth so that the mispronounced sentences could be noted and re-recorded. The stimuli were transduced with a Shure (SM98) microphone, and digitized on line in real-time via a 12-bit analog-to-digital converter (DT2801) at a sampling rate of 10 kHz. The stimuli were then low-pass filtered at 4.8kHz and the target words were digitally edited from the carrier sentences. The average root mean square amplitude of each of

the stimuli was equated using a signal processing software package (Luce and Carrell, 1981). In order to create different presentation levels for the amplitude condition (Condition 3), high and low amplitude versions of the medium rate tokens from each of the two talkers were created. These tokens were generated by setting the maximum waveform amplitude level to a specified value. The remaining amplitude values in the digital files were then rescaled relative to this specified maximum. For the high and low amplitude sets, the maximum amplitude values were set at 60 dB SPL and 35 dB SPL, respectively. All other stimuli were leveled at 50 dB SPL.

For each of the three conditions (talker, rate, and amplitude) eight separate word lists were constructed in which each test word was presented and then repeated once after a lag of 2, 8, 16 or 32 intervening words. Each list began with 15 practice trials, which were used to familiarize the subjects with the test procedure. None of these 15 words was repeated in the experiment. The next 30 trials were used to establish a memory load and were not used in the final data analyses. The rest of the list consisted of 144 test word pairs, and 21 filler items which were not included in the analysis. The test pairs were distributed evenly across the four lags, with half of the repetitions at each lag in the same voice, rate or amplitude and half in a different voice, rate or amplitude as the original presentation of the test word. The total number of words in each list was 354.

For all three conditions, the lag between the first and second repetition of a word was manipulated as a within-subject variable (2, 8, 16 or 32 words). For the talker condition (Condition 1), only the medium rate tokens were used, and the voice of the talker for the second repetition of the target words was a within-subject variable (same vs. different voice). Forty-two subjects participated in Condition 1. For the rate condition (Condition 2), only the fast and slow rate tokens from both talkers were used. For this condition, Talker was a between-subjects variable, with half the subjects responding to tokens produced by the male talker ($n=20$) and half responding to tokens produced by the female talker ($n=20$). The speaking rate of the second repetition of the target words was a within-subject variable (same vs. different rate). Finally, for the amplitude condition (Condition 3), only the medium rate tokens from both talkers were used, and Talker was a between-subjects variable, with half the subjects responding to tokens produced by the male talker ($n=19$) and half responding to tokens produced by the female talker ($n=19$). The overall amplitude of the second repetition of the target words was a within-subject variable (same vs. different amplitude).

Procedure

Subjects were tested in groups of five or fewer in a quiet room used for speech perception experiments. The presentation of stimuli and collection of responses was controlled by a PDP-11/34 computer. Each digital stimulus was output using a 12-bit digital-to-analog converter and was low-pass filtered at 4.8 kHz. The stimuli were presented binaurally over matched and calibrated headphones (TDH-39) at a comfortable listening level. On each trial, subjects heard a spoken word and had up to five seconds to enter a response of "old" (i.e., the word had appeared previously in the list of spoken words) or "new" (i.e., the word was new to the list). Subjects entered their responses on appropriately labeled two-button response boxes. If no response was entered after five seconds, that trial was not recorded and the program proceeded to the next trial. No feedback was provided. The entire session of 354 trials lasted approximated 25-35 minutes.

Results and Discussion

Figure 1 shows the item recognition accuracies for the same-talker and different-talker repetitions (Figure 1a), same-rate and different-rate repetitions (Figure 1b), and same-amplitude and different-

amplitude repetitions (Figure 1c) as a function of lag. For the Talker condition, a 2-factor ANOVA with Lag (2, 8, 16, 32) and Repetition (same-talker, different-talker) as factors showed significant main effects for both factors. Accuracy decreased with increasing lag ($F(3,328)=24.518$, $p<.0001$), and same-talker repetitions were recognized better overall than different-talker repetitions ($F(1,328)=5.516$, $p<.0194$). The two-way interaction was not significant. This result replicates the previous findings of Palmeri et al. (1993) that there is a same-voice advantage for recognizing a word as a repeated item without any explicit instructions to the subjects to attend to the talker's voice.

Insert Figure 1 about here

For the Rate condition, a 3-factor repeated measures ANOVA with Lag (2, 8, 16, 32), Repetition (same-rate, different-rate), and Talker (male, female) as factors showed significant main effects for Lag and Repetition but not for Talker (indicating no difference in recognition memory for words spoken by a male or a female talker). Accuracy decreased with increasing lag ($F(3,152)=17.057$, $p<.0001$) and same-rate repetitions were better recognized than different-rate repetitions ($F(1,152)=39.895$, $p<.0001$). There was no main effect of Talker ($F(1,152)=.323$, $p=.5708$) and none of the interactions were significant indicating that regardless of the talker, there were consistent and reliable effects of Lag and Repetition. This finding extends the same-voice advantage found by Palmeri et al. (1993) to a different item-specific characteristic of speech, and thus demonstrates that both talker and rate information are encoded in memory along with the symbolic/linguistic information about a spoken word.

For the Amplitude condition, a 3-factor repeated measures ANOVA with Lag (2, 8, 16, 32), Repetition (same-amplitude, different-amplitude), and Talker (male, female) as factors showed significant main effects for Lag and Talker. Accuracy decreased with increasing lag ($F(3,144)=38.474$, $p<.0001$), and accuracy was generally higher for the male talker than for the female talker ($F(1,144)=4.319$, $p<.0395$). However, there was no main effect of Repetition, and none of the interactions were significant. Thus, while recognition accuracy decreased with increasing lag, there was no difference in recognition accuracy between the same-amplitude and different-amplitude trials. Furthermore, this pattern of results was obtained for both talkers even though the overall accuracy scores for the male talker were slightly higher than for the female talker (91.2% and 88.9% correct item recognition, respectively). The fact that there was no same-amplitude advantage for both talkers suggests that overall amplitude information may not be a property of speech that is encoded into long-term memory in the same way as talker and rate information, and that different item-specific stimulus characteristics can have distinct effects on speech perception and spoken word recognition.

In order to compare the overall level of discrimination between "old" and "new" items across the three conditions, we computed d' scores for each subject in each condition. The mean d' score in all three conditions was significantly greater than zero ($p<.0001$ in all three conditions by a one-sample t-test), indicating good discrimination in all conditions (see Table I). Furthermore, a one-factor ANOVA with Condition as the factor, showed a significant main effect of Condition ($F(2,117)=5.198$, $p<.007$). Post-hoc comparisons (Fisher's PLSD) showed a significant difference in d' for the Talker and Rate conditions ($p<.002$), and for the Rate and Amplitude conditions ($p<.039$). However, there was no difference in d' for the Talker and Amplitude conditions.

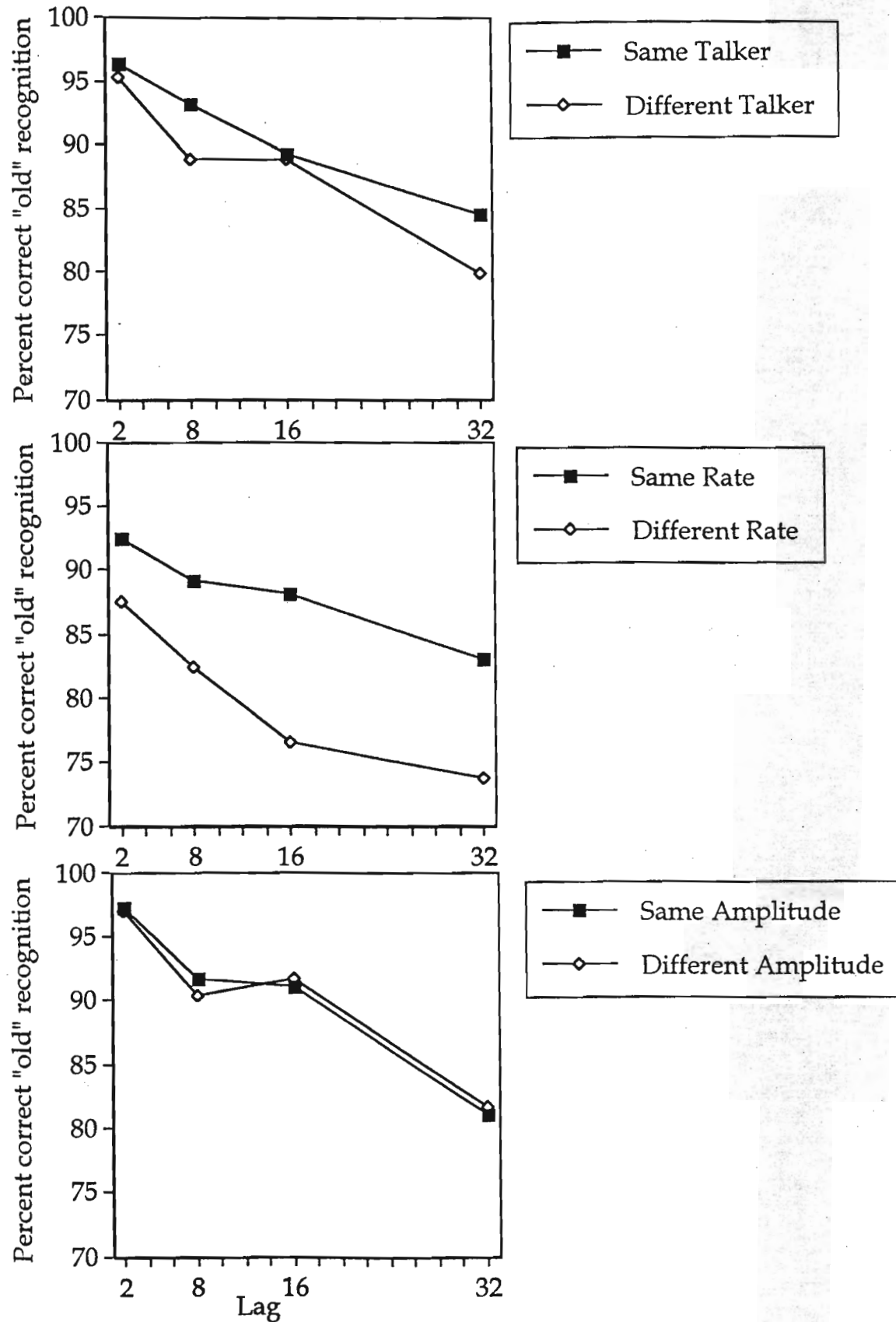


Figure 1. Item recognition accuracy scores as a function of lag from Experiment 1 for (a) the Talker condition, (b) the Rate condition, and (c) the Amplitude condition.

Table I
False Alarm Rates for Experiment 1.

Condition	Hit Rate	False Alarm Rate	d prime
Talker	81.0%	11.6%	2.17
Rate	80.3%	16.4%	1.91
Amplitude	81.5	13.5%	2.08

In summary, the results of Experiment 1 demonstrate that same talker and same rate trials were recognized better than different talker and different rate trials, respectively. In contrast, there was no difference in recognition memory for same and different amplitude trials. Thus, information about the talker's voice and speaking rate are encoded in the long-term memory representation of spoken words. However, there was no evidence that information about the overall amplitude of a spoken word is encoded in memory. The possibility remains, of course, that overall amplitude information may be retained in memory, but that when subjects are instructed to recognize the item as "new" or "old" they are unable to use this information as an implicit retrieval cue in this task. In order to evaluate this alternative another experiment was carried out.

EXPERIMENT 2

Experiment 2 was designed to investigate whether listeners can explicitly recognize changes in talker, rate and amplitude for a repeated word. Whereas in Experiment 1, subjects were not required to pay explicit attention to the voice, rate or amplitude of the test item, in Experiment 2, subjects were required to make an explicit judgment regarding a change in voice, rate or amplitude. We hypothesized that this would be a more sensitive test of the extent to which detailed information about the instance-specific characteristics of a spoken word are encoded in long-term memory. Specifically, we were interested in investigating the possibility that subjects are able to detect changes in overall amplitude even though a change in overall amplitude did not produce changes in recognition accuracy scores for words in the amplitude condition of Experiment 1.

Method

Subjects

One hundred and nineteen students enrolled in undergraduate introductory psychology courses at Indiana University served as subjects. All subjects received partial course credit for their participation. All were native speakers of American English and reported no history of speech or hearing disorder at the time of testing.

Stimuli and Procedure

The stimulus materials for Experiment 2 were identical to those used in Experiment 1. All aspects of the stimulus presentation and test conditions were identical to Experiment 1 except that in this experiment subjects were given three response categories rather than two. In Experiment 2, after hearing the spoken word, subjects had 5 seconds to identify the word as "new" if it had not occurred in the list

before, as “old-same” if it had occurred before and was repeated with the same voice (Condition 1), rate (Condition 2) or amplitude (Condition 3), or as “old-different” if it was repeated with a different voice (Condition 1), rate (Condition 2) or amplitude (Condition 3). Thus, in Experiment 2, in addition to recognizing a word as “old” or “new,” subjects were also required to make an explicit judgment for the items recognized as “old” regarding voice, rate or amplitude variation from the first to second repetition of the word. A group of 33 subjects participated in the talker condition. For the rate condition, a group of 21 subjects was tested on stimuli spoken by the male talker, and a separate group of 21 subjects was tested on stimuli spoken by the female talker. For the amplitude condition, a separate group of 22 subjects was tested on each of the two stimulus sets (one from the male talker, one from the female talker).

Results and Discussion

Figure 2 shows the overall percentage of correct “old” item recognition for the talker condition (Figure 2a), the rate condition (Figure 2b) and the amplitude condition (Figure 2c). The accuracy scores shown in this figure represent all cases of correct “old” item recognition regardless of accuracy on the “same-different” judgment. This analysis allowed us to compare the pattern of results on the item recognition task across Experiments 1 and 2.

 Insert Figure 2 about here

As shown in Figure 2, same talker trials were recognized better than different talker trials, same rate trials were recognized better than different rate trials, but there was no difference in recognition accuracy for same and different amplitude trials. This pattern of results is consistent with the results of Experiment 1. For the talker condition, a 2-factor ANOVA with Repetition (same talker or different talker) and Lag (2, 8, 16, 32) as factors showed main effects of both factors. Same-talker trials were better recognized than different-talker trials ($F(1,256)=4.541, p=.0340$), and recognition accuracy decreased with increasing lags ($F(3,256)=13.258, p<.0001$). The two-way interaction was not significant.

For the rate condition, a 3-factor repeated measures ANOVA with Repetition (same rate or different rate), Lag (2, 8, 16, 32), and Talker (male, female) as factors showed main effects of all three factors. Same-rate trials were better recognized than different-rate trials ($F(1,160)=26.973, p<.0001$), recognition accuracy decreased with increasing lags ($F(3,160)=20.906, p<.0001$), and recognition accuracy was slightly better for tokens produced by the male talker than for those produced by the female talker (mean difference = 2.94%, $F(1,160)=4.815, p=.0297$). None of the interactions involving Talker as a factor was significant indicating that the pattern of decreasing recognition accuracy with increasing lags, and across same-rate and different-rate trials, was consistent across both talkers. Similarly, the two-way interaction between Repetition and Lag was not significant.

For the amplitude condition, a 3-factor repeated measures ANOVA with Repetition (same amplitude or different amplitude), Lag (2, 8, 16, 32), and Talker (male, female) as factors showed a main effect of Lag, but no main effects of Repetition or Talker. None of the interactions was significant. As expected from the results of Experiment 1, recognition accuracy decreased with increasing lags ($F(3,168)=48.870, p<.0001$) but there was no same amplitude

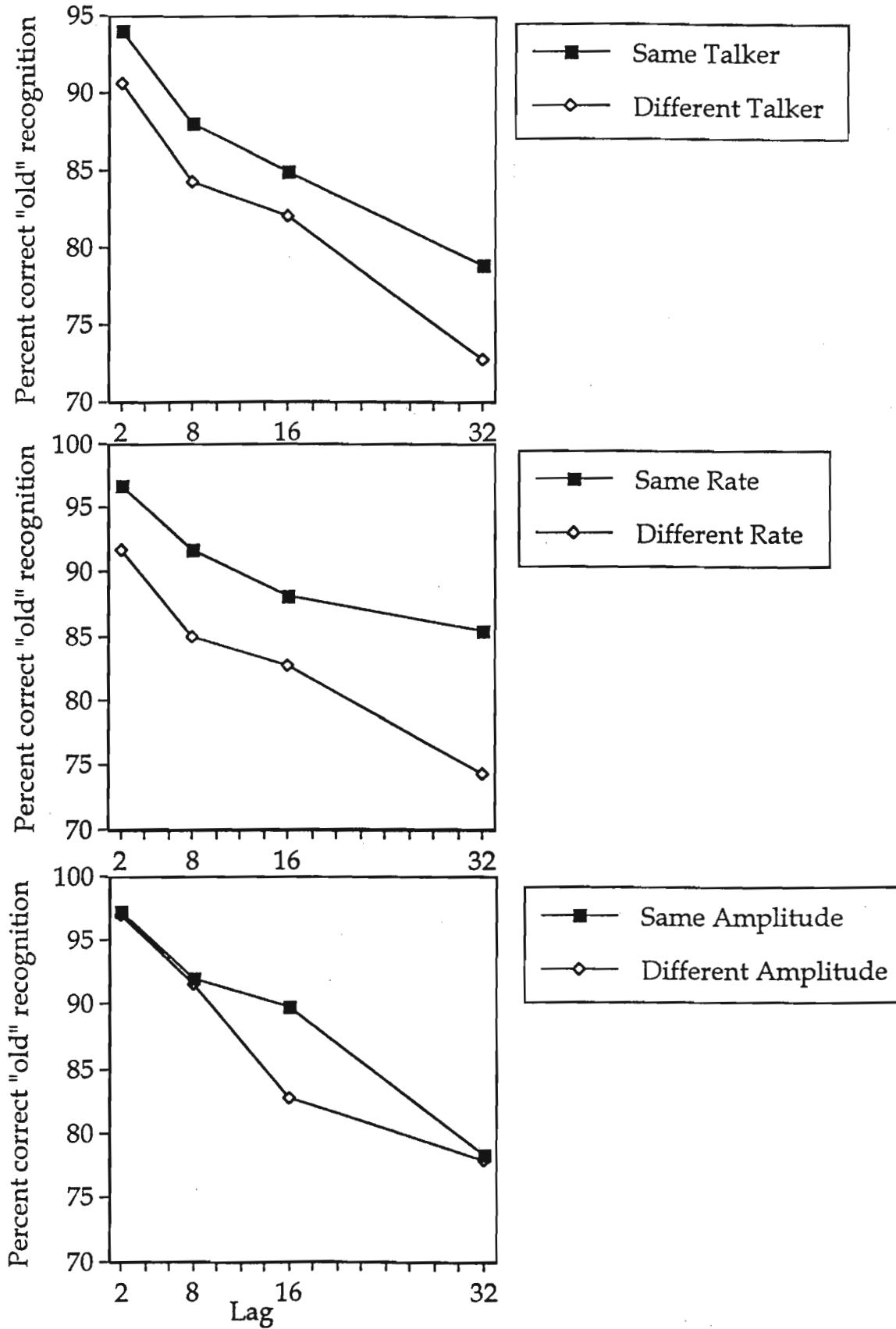


Figure 2. Item recognition accuracy scores as a function of lag from Experiment 2 for (a) the Talker condition, (b) the Rate condition, and (c) the Amplitude condition.

linguistic information about a spoken word. In contrast, once again, there was no evidence that information about overall amplitude is retained in long-term memory.

The similarity between the patterns of item recognition accuracy scores for the two experiments indicates that the additional response category for Experiment 2 did not alter the main effects of Lag and Repetition on item recognition accuracy. In order to assess directly the effect of the additional response category, separate repeated measures ANOVAs for each of the three conditions with Experiment (1 or 2) as the repeated measure were performed. For the Talker condition, the analysis showed the expected main effects of Lag ($F(3,256)=33.364$, $p<.0001$) and Repetition ($F(1,256)=8.552$, $p=.0038$). The two-way interaction between Lag and Repetition was not significant. There was also a significant main effect of Experiment ($F(1,256)=12.059$, $p=.0006$) due to generally higher accuracies for Experiment 1 than for Experiment 2 (means = 88.55% and 84.36%, respectively). None of the interactions involving the Experiment factor were significant indicating that the patterns of decreasing accuracy with increasing lag, and of higher accuracy for same-voice repetitions, were consistent across both experiments. For the Rate condition, there were main effects of Lag ($F(3,344)=40.025$, $p<.0001$) and Repetition ($F(1,344)=73.220$, $p<.0001$), but there was no effect of Experiment and none of the interactions were significant. Finally, for the amplitude condition, the main effect of Lag was significant ($F(3,304)=76.150$, $p<.0001$), and the main effect of Experiment was significant ($F(1,304)=7.398$, $p=.0069$) but there was no main effect of Repetition. As for the Talker condition, the effect of Experiment for the Amplitude condition was due to generally higher accuracies for Experiment 1 than for Experiment 2 (means = 90.21% and 87.82%, respectively). Thus, the additional response category in Experiment 2 resulted in slightly lower overall recognition accuracy scores for the Talker and Amplitude conditions. However, across all three conditions, the general pattern of results for the two experiments was consistent in showing a same-voice and same-rate advantage relative to different-voice and different-rate trials, respectively. Similarly, both experiments showed no same-amplitude advantage relative to different-amplitude trials.

In order to determine whether subjects can explicitly recognize variation in talker, rate and amplitude for items that were correctly identified as "old," d' scores were calculated for each condition at each lag. In this analysis, a "Hit" was defined as a response of "old/same" to a stimulus that was repeated with the same voice, rate or amplitude. A "False Alarm" was defined as a response of "old/same" to a stimulus that was repeated with different voice, rate or amplitude. Using this measure, we were able to determine if listeners can discriminate changes in talker, rate and amplitude, and thus establish whether detailed information along each of these dimensions was retained in memory.

 Insert Figure 3 about here

Figure 3 shows the d' scores for all three conditions as a function of lag. Two main findings are shown in this figure. First, for all three conditions at all lags, the d' scores differed significantly from zero indicating that subjects were able to discriminate "old/same" from "old/different" trials in all cases. One sample t-tests for each condition at each lag confirmed that these d' scores were all significantly different from zero at the $p<.0001$ level. This finding suggests that, regardless of whether the instance-specific information affected recognition memory accuracy in the "old-new" task, listeners do retain highly detailed information in memory to the extent that variability along each of the three dimensions was explicitly detected. Second, variability along each of the three dimensions was detected with a different degree of accuracy: talker variability was detected better than rate variability which was detected better than amplitude variability. A two factor ANOVA with Condition (talker, rate, amplitude) and Lag (2, 8, 16, 32)

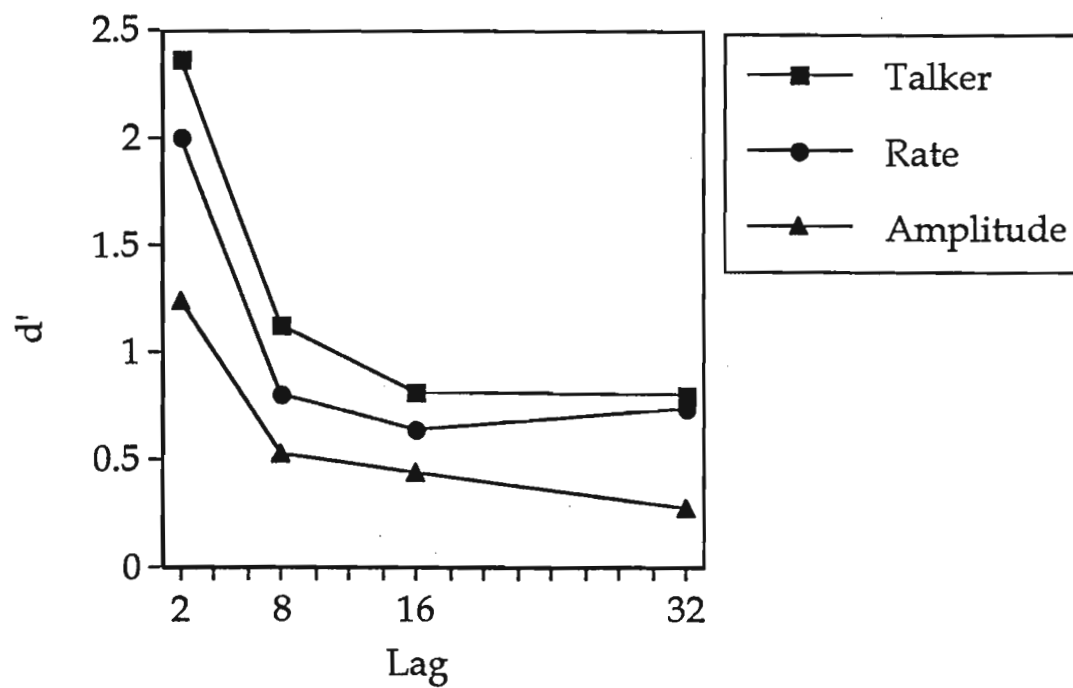


Figure 3. d' scores for all three conditions of Experiment 2 as a function of lag.

as factors showed main effect of both factors (Condition: $F(2,476)=45.459$, $p<.0001$; Lag: $F(3,476)=110.988$, $p<.0001$). The two-way interaction was also significant ($F(6,476)=3.264$, $p=.0037$). This finding suggests that, although fine details of the stimulus dimensions are retained in memory, certain dimensions represent more perceptually salient characteristics than others, and thus may produce more substantial effects on speech perception and spoken word recognition performance in different tasks.

The main goal of Experiment 2 was to investigate whether subjects were able to explicitly discriminate changes in talker, rate or amplitude for items that they recognized as repeated items (i.e., "old" items). In particular, we were interested in the results of this task for the amplitude condition where inconsistent amplitude information did not affect recognition memory performance. The results showed that subjects were indeed able to explicitly detect changes in talker, rate and amplitude. Thus, this task provided evidence that, even though all sources of variability do not function identically with respect to spoken word recognition, detailed information about the instance-specific characteristics of a spoken word is retained in memory along with the more abstract linguistic content of the word. These highly detailed memory representations even include information along an apparently linguistically irrelevant dimension such as overall amplitude.

GENERAL DISCUSSION

The overall goal of this study was to investigate the extent to which the neural representation of spoken words encodes detailed, instance-specific information. The results that emerged from this study complement the findings of earlier studies that have investigated the effects of talker, rate and amplitude variability on speech perception and memory for spoken words. The general pattern of results that has emerged from this set of experiments (summarized in Table II) suggests that information about all sources of variability is retained in long-term memory. However, the processing costs incurred by trial-to-trial variability along different stimulus dimensions varies for different properties of the speech signal.

A comparison of the effects of talker, rate, and amplitude variability on the tasks listed in Table II reveals a hierarchy in which amplitude, rate and talker variability have increasingly profound effects on speech perception and memory for spoken words. The relatively weak effect of amplitude variability is seen by the fact that experiments using all three tasks (word identification, serial recall, and continuous recognition) failed to find an effect of trial-to-trial changes in overall amplitude. In fact, the only evidence that overall amplitude information is retained in long-term memory comes from the task in which subjects were asked to *explicitly* identify variability along this dimension (present study, Experiment 2). In contrast, the stronger effect of rate variability was evident in all three tasks, where trial-to-trial changes in speaking rate resulted in decreased performance relative to trials with no change in speaking rate. For instance, word lists in which each word was spoken at a constant speaking rate were better identified when embedded in noise than identical lists spoken with multiple speaking rates (Sommers et al., 1994). Similarly, single-rate word lists were more accurately recalled than multiple-rate lists (Nygaard et al., 1995); and, consistent rate trials were better recognized in a continuous recognition memory task than trials in which the rate changed (present study, Experiment 2). The effects of talker variability are comparable to the effects of rate variability, however, a difference between the effects of talker and rate variability emerged in the serial recall task with long ISI's (Nygaard et al., 1995). When given enough time, the talker's voice was apparently encoded by the listeners in the long-term memory representation of the spoken words, and thus served as an identifying feature of the words. In this manner, the talker's voice functioned as a retrieval cue and aided the listener in the serial recall task to the extent that multiple talker lists were *better* recalled than single talker lists. In contrast, at long ISI's, the detrimental effect of multiple speaking rates was diminished only to the extent that multiple rate lists were recalled as well as single rate lists. Thus, the results of these

studies lead us to postulate a hierarchy of effects of stimulus variability on speech perception and memory for spoken words with talker variability having the most pervasive effects, rate variability having intermediate effects, and amplitude variability having the weakest effects.

Table II.

Summary of findings regarding the impact of talker, rate and amplitude variability on speech perception and memory for spoken words.

Source of Variability	Word Identification	Serial Recall		Recognition Memory	
		Short ISI's	Long ISI's	Item Recognition	Attribute Recognition
Talker	Single>Multiple ^{1,2}	Single>Multiple ^{3,4,5}	Multiple>Single ^{4,5}	Same>Different ^{6,7}	Yes ^{6,7}
Rate	Single>Multiple ²	Single>Multiple ⁵	Multiple=Single ⁵	Same>Different ⁷	Yes ⁷
Amplitude	Single=Multiple ²	Single=Multiple ⁵	Multiple=Single ⁵	Same=Different ⁷	Yes ⁷

¹ Mullennix et al. 1988

² Sommers et al. 1994

³ Martin et al. 1989

⁴ Goldinger et al. 1991

⁵ Nygaard et al. 1995

⁶ Palmeri et al. 1993

⁷ Present study

At this point we can speculate as to the mechanism that underlies these different effects for different sources of variability. It is possible that these differences in the effects of talker, rate, and amplitude variability reflect differences in the complexity of the acoustic correlates of changes along these dimensions. In all of the experiments listed in Table II that investigated the effects of amplitude variability, a change in amplitude was achieved by simply setting the maximum level for each waveform to a specified value and then rescaling the remaining amplitude levels relative to that maximum. Thus, amplitude variability was a constant, uni-dimensional adjustment. In contrast, rate variability was more naturally achieved, and was thus variable and multi-dimensional in its acoustic correlates. Rate variability within a given speaker is not achieved by a constant "stretching" or "shrinking" of the acoustic waveform. Rather, certain acoustic segments are more dramatically reduced in duration than others when overall speaking rate is increased, and various other acoustic-phonetic changes (e.g., vowel reduction) occur in response to changes in speaking rate (e.g., Lehiste, 1972; Klatt, 1973, 1976; Port, 1981; Picheny et al., 1986, 1989; Uchanski et al., 1996). Thus, an increase or decrease in speaking rate is clearly a dynamic, multi-dimensional transformation of the speech signal. Similarly, a change in talker leads to a wide variety of

acoustic-phonetic changes. Not only do different talkers differ in vocal tract shape and size, which leads to different spectro-temporal characteristics, but different talkers also differ in articulatory "style" (including speaking rate, dialect, and other idiosyncratic differences) which can lead to large differences in the acoustic waveform of a given word across various talkers (e.g., Fant, 1973; Joos, 1948; Peterson and Barney, 1952).

Thus, the extent of the effects of variability in talker, rate, and amplitude investigated by the experiments listed in Table II appear to be directly related to the complexity of the acoustic correlates that result from these sources of variability. From the listener's point of view then, it is possible that the simpler the acoustic transformation related to a given source of variability, the fewer the processing resources required to compensate for that variability, and consequently the less the impact of this variability on speech perception and memory for spoken words.

Another explanation for the differential effects of the different sources of variability on speech perception and memory for spoken words takes into account the relevance of each source of variability for the perception of phonetic contrasts. Variability in talker characteristics has been shown to have a significant impact on phonetic contrast perception. For example, Ladefoged and Broadbent (1957) found that vowel identification could be altered depending on the perceived talker characteristics of a precursor phrase, and Johnson (1990) showed that perceived speaker identity plays an important role in the F0 normalization of vowels. Similarly, several studies have demonstrated the rate dependency of phonetic processing for both vowels and consonants (e.g., Port, 1981; Summerfield, 1981; Miller, 1987; Miller and Volaitis, 1989). In contrast, overall amplitude variability does not, by itself, signal phonetic contrasts, and there does not appear to be an amplitude-dependency in speech perception that is comparable to talker- and rate-dependent phonetic processing. Thus, it is possible that the observed differences between the effects on speech perception and memory for spoken words of talker and rate variability on the one hand, and amplitude variability on the other, is due to differences in their phonetic relevance to the listener.

A wider range of sources of variability needs to be investigated in order to provide conclusive evidence for one of these two alternative explanations for the different effects of different sources of variability. For example, it may be enlightening to investigate the effects of variations in dialect, vocal effort, emotional state and other such para-linguistic characteristics, as well as the effects of non-linguistic factors such as filtering characteristics due to different microphones or recording conditions. Nevertheless, so far as we can tell from the available data, all instance-specific stimulus attributes appear to be retained in memory to the extent that listeners are able to detect such changes. There is now a growing body of converging evidence demonstrating that the processes of speech perception and spoken word recognition operate in the context of highly detailed representations of the acoustic speech signal, rather than on idealized abstract symbolic representations of abstract linguistic information. We believe these are important new observations about speech and spoken language processing that have broad implications for future research and theory about speech perception.

References

- American National Standards Institute (1971). *Method for measurement of monosyllabic word intelligibility*. (American National Standard S3.2-1960 [R1971]). New York: Author.

- Church, B. A. & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 521-533.
- Craik, F. I. M. & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274-284.
- Fant, G. (1973). *Speech sounds and features*. Cambridge, MA: MIT Press.
- 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, NY: Academic Press.
- Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, **88**, 642-654.
- Joos, M. A. (1948). Acoustic phonetics. *Language*, **24**, 1-136.
- 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. (1973). Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, **54**, 1102-1104.
- 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. (1989). Cognitive science and speech: A framework for research. In H. Schnelle and N. O. Bensen (Eds.), *Logic and linguistics: Research directions for cognitive science. European Perspectives*, (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, UK: Cambridge University Press.
- Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, **51**, 2018-2024.
- Luce, P. A. & Carrell, T. D. (1981). *Creating and editing waveforms using WAVES* (Research in Speech Perception, Progress Report No. 7). Bloomington, IN: Indiana University Speech Research Laboratory.

- 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.
- 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., & Volaitis, L. E. (1989). Effects of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, *46*, 505-512.
- 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. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088-2113.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception and Psychophysics*, *57*, 989-1001.
- Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309-328.
- Peterson, G. E. & Barney, H. L. (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, *24*, 175-184.
- Picheny, M. A., Durlach, N. I. & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, *29*, 434-446.
- Picheny, M. A., Durlach, N. I. & Braida, L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *Journal of Speech and Hearing Research*, *32*, 600-603.
- Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*, 109-125.
- Pisoni, D. B. (1997). Some thoughts on "Normalization" in speech perception. In J. Mullennix and K. A. Johnson (Eds.), *Talker variability in speech processing*, (pp. 9-32). Academic Press.
- Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, *69*, 262-274.
- Schacter, D. L. & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 915-930.
- Sheffert, S. M. & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Learning and Memory*, *34*, 665-685.

- Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-1324.
- Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **64**, 1358-1368.
- Summerfield, Q. (1981). On articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 1074-1095.
- Uchanski, R. M., Choi, S., Braida, L. M., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research*, **39**, 494-509.