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

Perceptual Learning of Natural and Sinewave Voices¹

**Sonya M. Sheffert, David B. Pisoni, Jennifer M. Fellowes²
and Robert E. Remez²**

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

¹ This research was supported by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University-Bloomington, and NIH-NIDCD Research Grant DC00308 to Barnard College, Columbia University.

² Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027.

Perceptual Learning of Natural and Sinewave Voices

Abstract. This report describes the results of a perceptual training study that was designed to explore how listeners learn to categorize novel voices and how knowledge of a familiar voice generalizes to novel utterances. The speech samples from which the listeners learned to identify individuals were of two kinds: Naturally produced English sentences and sinewave replicas of these sentences. The sinewave items were nonspeech tonal patterns that preserved coarse-grained properties of the talker's vocal tract transfer function while eliminating traditional cues to voice quality. Listeners were trained over several days to identify by name ten talkers from sentence length sinewave or natural speech utterances. Knowledge about the talker's voice was then assessed using two generalization tests in which listeners heard a novel set of sentences and were required to identify the speaker. In one generalization test, the sentences were sinewave replicas whereas in the other generalization test, the sentences were naturally produced. The results showed that perceptual learning of a talker's voice can occur even when specific acoustic products of vocal articulation are eliminated from the signal. The data also showed that speaker-specific knowledge acquired during this perceptual training task generalized to novel natural and novel sinewave sentences. Variability in the degree of perceptual learning affected generalization of speaker knowledge. The results of this study show that listeners can learn about a talker's voice from highly impoverished acoustic signals when the products of vocal articulation are eliminated, and that this knowledge generalizes to novel utterances produced by these same talkers.

Introduction

When a speaker produces an utterance, the listener recovers from the acoustic signal not only information about the consonants and vowels that compose the message, but information about the speaker's vocal tract morphology, affect and pronunciation habits. Historically, the perception and representation of the linguistic content of an utterance has been thought to be separate and independent from the processes needed for the encoding of speaker information (Halle, 1985; Laver & Trudgill, 1979). Recent findings, however, suggest a different perspective on the relation between linguistic and "indexical" information in speech, namely, one in which these two sets of attributes interact. The line of evidence most relevant to the current investigation comes from a recent series of perceptual learning studies conducted by Nygaard and colleagues (Nygaard & Pisoni, 1995; see also Nygaard, Sommers & Pisoni, 1994). In their procedure, listeners are trained over several days to identify a set of talkers from sentence length utterances. Subjects then are given a speech intelligibility test in which they are asked to transcribe new sentences presented in white noise. Nygaard and Pisoni (1995) found that subjects who had become familiar with the talkers were able to transcribe sentences more accurately than subjects who were unfamiliar with the speakers. This finding demonstrates that familiarity with the speaker who produced a sentence facilitates perceptual analysis of novel sentences, and supports the inference that linguistic and nonlinguistic information are not processed independently (see also Mullennix & Pisoni, 1990). This interaction between speaker and phonetic information also suggests similarities in the nature of the perceptual operations and neural representations that underlie voice recognition and speech perception.

Although the perceptual dimensions in which a voice is represented in memory are largely unspecified, several acoustic-phonetic cues have been traditionally assumed to underlie speaker recognition

(Bricker & Pruzansky, 1976; Laver & Trudgill, 1979). These include fundamental frequency, vocal tract resonances, glottal source, harmonic structure and the fine-grained power spectra of nasals and vowels. Recent findings reported by Remez, Fellowes and Rubin (in press) call into question the necessity of these traditional cues for speaker recognition. They demonstrated that listeners can identify familiar talkers in the absence of such variables using the technique of sinusoidal speech synthesis, which generates a nonspeech pattern that specifies the dynamics of a talker's vocal tract transfer function. Sinewave utterances are time-varying sinusoidal patterns that track the changing center frequencies of the naturally produced utterance from which they are modeled. These nonspeech tonal stimuli can be thought of as a highly simplified representation of the frequency and amplitude changes present in speech, an "acoustic caricature" of the original utterance. Although the signal is highly impoverished, most listeners are nevertheless able to perceive the linguistic content of the utterance (Remez, Rubin, Pisoni & Carroll, 1981), indicating that the dynamic properties of the sinewaves are sufficient to support phonetic perception.

The recent findings of Remez et al. (in press) further show that the global time-varying properties of the sinewaves also preserve speaker-dependent aspects of speech. In Remez et al. (in press), sinewave utterances modeled from the natural productions of ten speakers were presented to listeners in a voice recognition task. The listeners were members of the staff at Haskins Laboratories who had become highly familiar with the speakers over many years of social contact. In this task, they were required to identify the speaker from which the replica originated. To perform this voice recognition task, listeners had to draw on their long-term knowledge of a talker's voice and speaking style and compare this to the information preserved in the sinewave signals. The results showed that listeners were generally successful at identifying the familiar voices of their colleagues from the sinewave utterances. Identification accuracy exceeded chance for 6 of the 10 speakers. This finding indicates that information in the sinewaves specifying changes in the talker's vocal tract is sufficient to support talker identification among listeners who were highly familiar with the speakers, and that speaker identification can take place even when traditional voice recognition cues are eliminated from the signal.

The results also showed considerable variability in the recognizability of different speakers, as evidenced by the fact that recognition accuracy was below chance for some speakers in the set. Because Remez et al. (in press) did not experimentally manipulate or control familiarity, it is impossible to know whether the observed variability was due to differences in the degree of familiarity, or to other factors present in the speaker ensemble, such as perceptual distinctiveness or discriminability of the stimuli. The present study uses the same test items as Remez et al. (in press), but controls for the amount of speaker familiarity by using a laboratory-based training procedure (cf. Nygaard, et al., 1994) in order to examine in detail how variability in the rate and degree of perceptual learning affects the identifiability of different speakers.

The present investigation has several objectives. First, in Experiment 1, we wanted to establish whether perceptual learning of voices can take place in the absence of the acoustic cues typically assumed to underlie talker identification. Second, we wanted to assess the effects of perceptual learning of sinewave speakers on the generalization to novel utterances. In particular, will perceptual learning of speaker information be context-specific, generalizing only to other sinewave replicas, or will training also generalize to novel natural speech utterances? A final objective was to assess the effect of the degree of perceptual learning on the generalization of speaker-specific knowledge.

Experiment 1

Familiarity of the sinewave speakers was experimentally manipulated by training listeners to identify by name the ten talkers producing the original sentences. In Experiment 1, the sentences were sinewave replicas of the natural utterances. Subjects were trained until they were able to identify the ten talkers with at least 70% accuracy. Speaker knowledge was then assessed using two generalization tasks in which listeners heard a novel set of sentences and were required to identify the speaker. In one generalization test, the sentences were sinewave replicas whereas in the other generalization test, the sentences were naturally produced utterances. In both cases, the generalization tests used utterances that the subjects had not heard before during training.

Based on the findings of Remez et al. (in press), which demonstrated talker identification from sinewave utterances, we predicted that with appropriate training and feedback, the higher-order relational information preserved in the sinewave replicas would support the perceptual learning of speakers. We also expected that the speaker-specific knowledge acquired during sinewave training would not be dependent on the specific training items but would instead generalize to novel natural and novel sinewave sentences. We did, however, expect that generalization would be based on the similarity to known examples, and consequently, would be greatest in the condition in which the perceptual form of the sentences was the same across training and test. Specifically, talker identification would be higher on the sinewave generalization test as compared to the natural speech generalization test.

Method

Subjects

Nineteen adult subjects were recruited from the Bloomington, Indiana, community. Of these, five subjects failed to complete the study due to work or school commitments, and six were excused because of extremely slow progress during the initial training sessions. The remaining eight subjects completed the sinewave training phase and the two generalization tests. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. Subjects were paid for their participation.

Test Materials

The natural and sinewave sentences used in the present experiments were the same materials developed by Remez et al. (in press). The stimulus materials consisted of two sets of sentences. The first set contained nine natural utterances produced by five male and five female talkers. Each talker produced all nine sentences, for a total of 90 items. Audio recordings were obtained by asking speakers to read the sentences aloud in their natural speaking style. The sentences were then recorded on audiotape in a sound-proof booth and were low-pass filtered at 4.5 kHz, digitally sampled at 10 kHz, equated for root mean squared (RMS) amplitude and stored as sampled data with 12-bit resolution.

The second set of sentences were sinewave replicas of the original natural speech tokens. To create these items, the frequencies and amplitudes of the first three formants were derived at 5 msec intervals interactively relying on two representations of the spectrum: 1) linear predictive coding (LPC), and 2) discrete fourier transforms (DFT). Three time-varying sinusoids were then synthesized based on the center frequencies and amplitudes of the formants (Rubin, 1980). The synthesis algorithm preserved higher-order patterns of spectro-temporal change of the vocal tract transfer function, while eliminating the fundamental

frequency, harmonic relations and fine-grained spectral information. Subjectively, the sentences were difficult to understand and sounded very unnatural.

Three sentences were randomly selected (without replacement) for each of the three phases of the experiment (training, natural speech generalization and sinewave speech generalization). All sentences were rotated through all conditions for each listener to ensure that the observed effects were not due to any specific subset of the sentences or any order effects.

Procedure

Training Phase

Listeners were trained over several days to learn to identify the names of the 10 speakers using the sinewave utterances. Subjects were tested in groups of three or fewer in a quiet listening room. During each training session, subjects heard a random ordering of five repetitions of three sentences from each talker (150 items total). The same three sentences were used for each talker in each training session, and subjects were told before hand which three sentences they would be hearing. The sinewave training sentences were presented binaurally to subjects at 75 dB SPL over matched and calibrated stereophonic headphones (Beyerdynamic DT100). Subjects were asked to listen carefully to each sentence and to pay close attention to the talkers' voices. Each time a sentence was presented, the subject was required to press one of ten keyboard buttons labeled with each speaker's name. Keys 1-5 were labeled with female names and keys 6-10 with male names. Each time a subject made a response, the accuracy of that response and the name of the correct talker was displayed on the computer screen in front of the subject and recorded in the computer. Each training session lasted approximately 30 minutes. Training was continued until subjects achieved an average of 70% correct speaker recognition performance.

Familiarization Phase

Before beginning each of the generalization tests, subjects completed a brief familiarization task to remind them of the correspondence between the sinewave tokens and the speakers. The familiarization task was simply an abbreviated version of a training session in which subjects listened and responded to one instance of each sentence in each talker (30 items total). The items were presented in a random order and subjects received feedback after each response. The familiarization task took approximately 8 minutes.

Generalization Tests

After reaching a 70% correct criterion in the sinewave training phase, subjects completed two generalization tests. One generalization test presented three novel sinewave sentences, whereas the second test presented three novel naturally produced sentences. All of the sentences presented during the generalization tests were new to the subject. Half the subjects received the natural generalization test before the sinewave generalization test, while the other half received the tests in the opposite order. Each generalization test presented five repetitions of each of the three sentences in a random order (150 items total). Subjects were informed of the sentences they would be hearing before the start of each test. Subjects were asked to attend specifically to the talker's voice and to identify the speaker by pressing one of the ten buttons on the keyboard as they had done in the previous training phase. Subjects did not receive feedback during either of the two generalization tests.

Results and Discussion

Training Performance

Analysis of the training data revealed that listeners were, in fact, able to learn to identify the ten speakers from these highly impoverished sinewave signals. Their performance showed continuous improvement during the training phase. After the first training session, talker identification performance was above chance and steadily increased by an average of 5% each day. By the last day of training, listeners were able to identify the speakers with a mean accuracy of 76%. Figure 1 displays the mean identification performance as a function of training days and speaker sex. Figures 2a and 2b display each subject's talker identification performance as a function of training days and speaker sex. The latter figures illustrate that learning progressed at different rates for different subjects. The number of days needed to reach the 70% criterion varied among the subjects from 9 days to 16 days. Because the number of training days differed across subjects, speaker recognition improvement was assessed using a sample of four training days. Specifically, we used the first day of training, two evenly separated days in the middle, and the last day of training for the statistical analysis. For example, 16 days of training were reduced to Days 1, 6, 11 and 16.

Insert Figure 1 about here

Insert Figures 2a and 2b about here

A repeated measures analysis of variance with the factors of days of training and speaker sex was conducted on the accuracy data (shown in Figure 1). The analysis revealed significant effects for days of training, $F(3, 42) = 221.49$, $p < .0001$, and speaker sex, $F(1, 14) = 34.02$, $p < .0001$. The latter reflected the advantage in the identifiability of the female talkers during training.

The data from the last day of training also showed variability in the identifiability of different speakers within the training set. Figure 3 shows speaker recognition performance on the last day of training as a function of speaker. Examination of the graph shows the two most accurately identified talkers were female ("F1" and "F2") whereas the two least accurately identified talkers were male ("M2" and "M5"). An ANOVA comparing talker recognition performance on the last day of training revealed a significant effect of speaker sex, $F(1, 78) = 18.69$, $p < .0001$, confirming that the female speakers were more accurately identified than the male speakers.

Insert Figure 3 about here

We also observed considerable variability among the speakers within each sex. An ANOVA with speaker as a factor was conducted on the training scores for each sex. Differences were found among the female speakers, $F(4, 28) = 6.01$, $p < .001$, and the male speakers, $F(4, 28) = 8.35$, $p < .0001$. Taken together, the training data demonstrate that listeners can learn to identify individuals from sinewave replicas

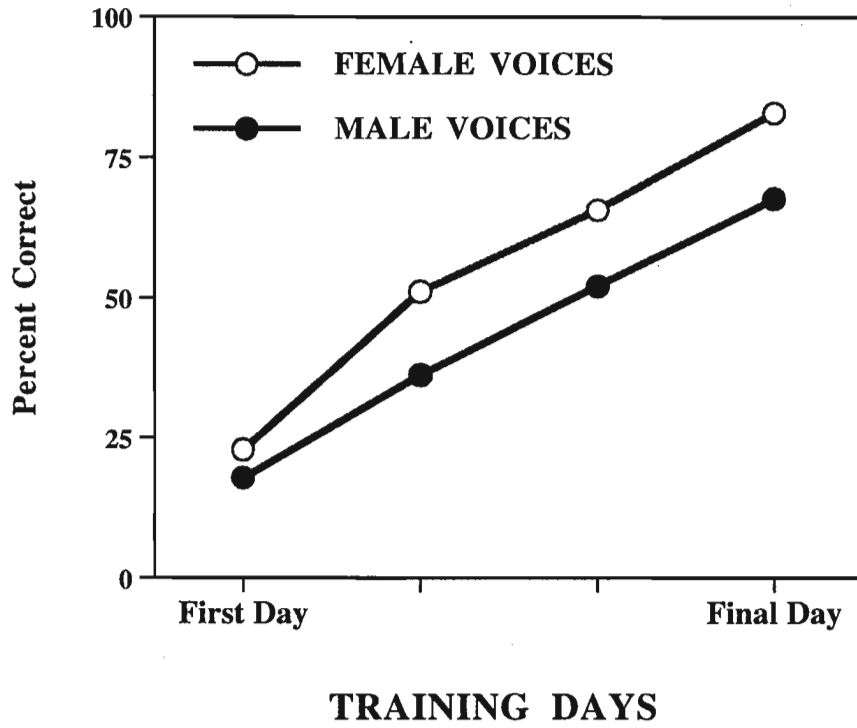
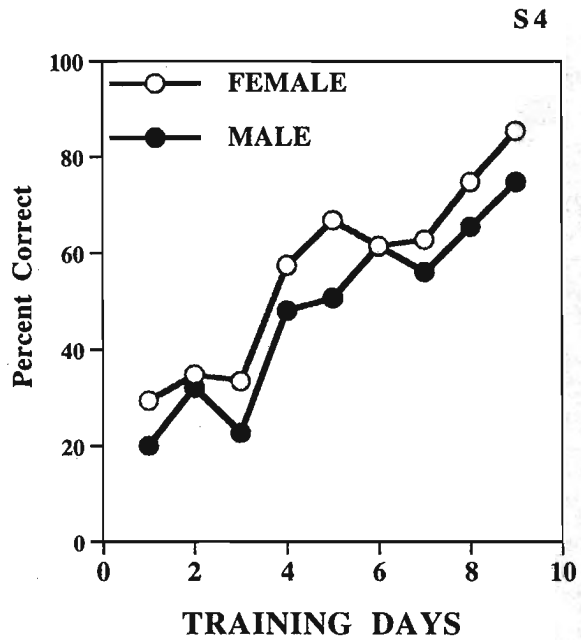
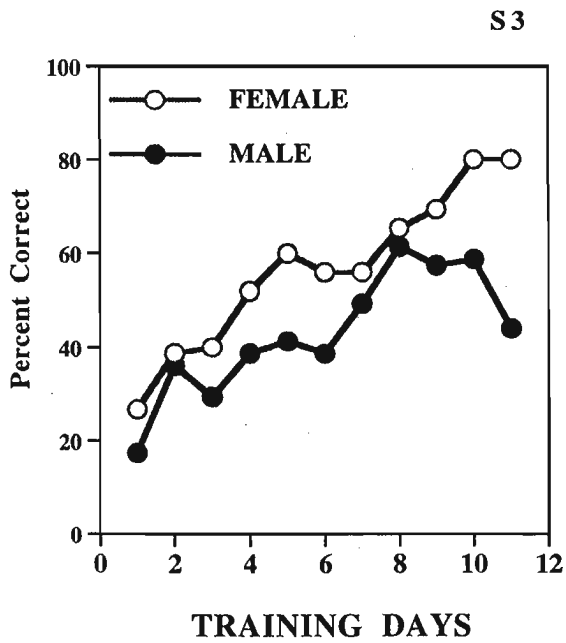
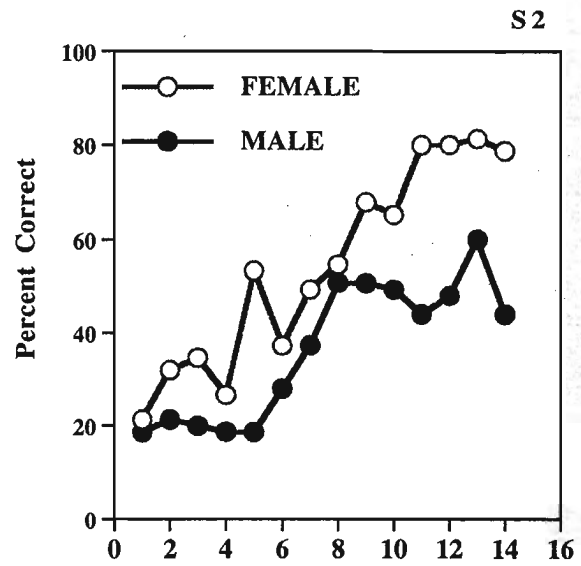
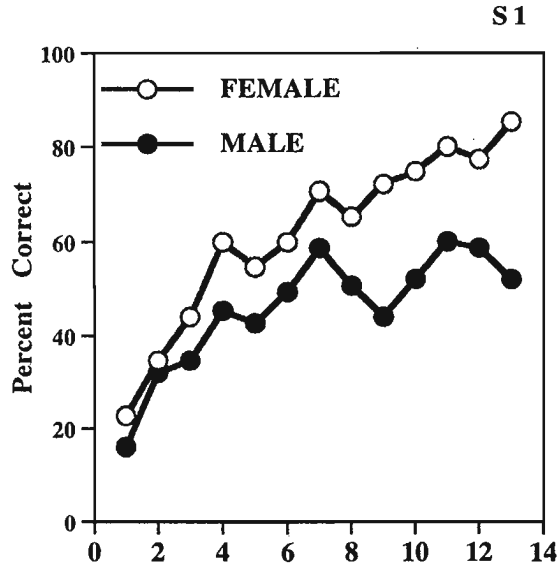
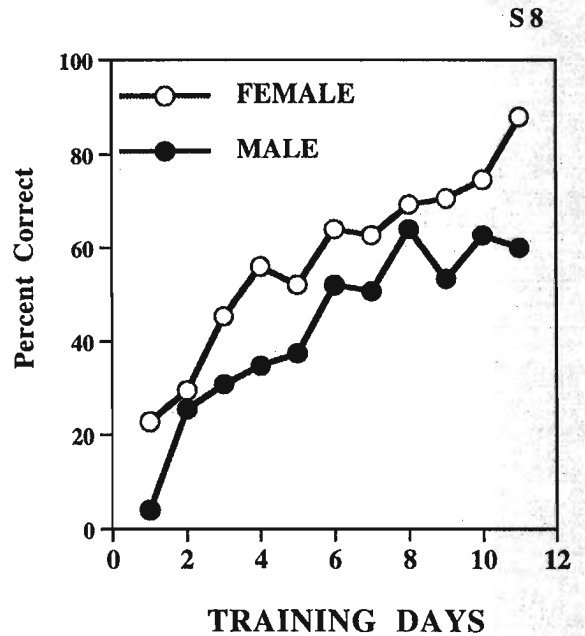
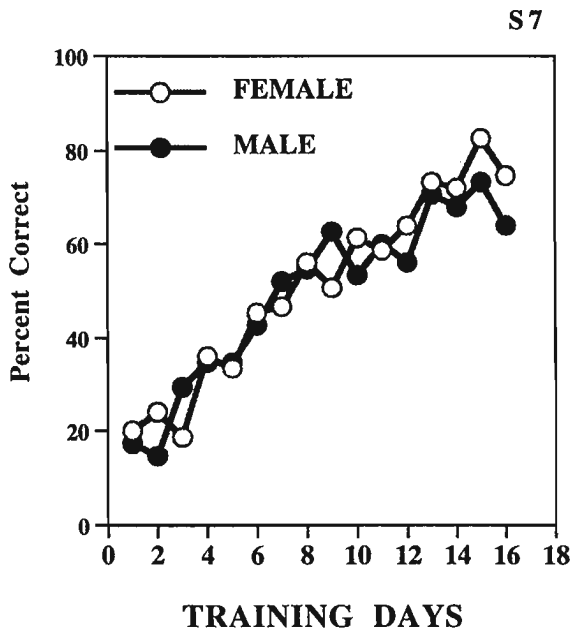
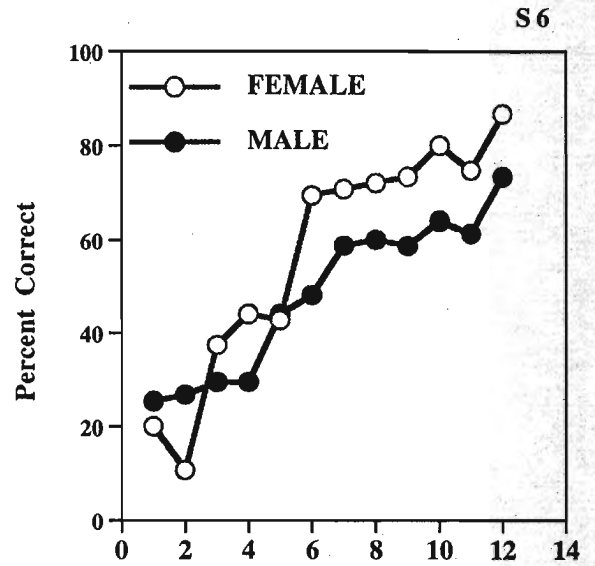
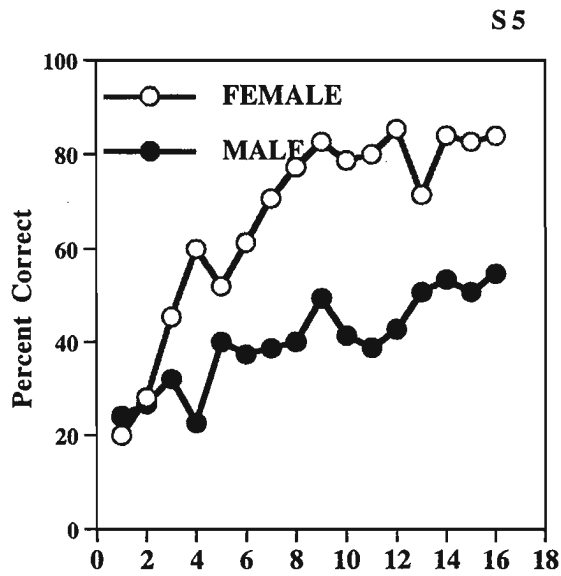


Figure 1. Mean speaker identification performance on sinewave replicas as a function of training days and speaker sex.



Figures 2a. Mean speaker identification performance on the sinewave training for subjects 1-4 as a function of training days and speaker sex.



Figures 2b. Mean speaker identification performance on the sinewave speech training for subjects 5-8 as a function of training days and speaker sex.

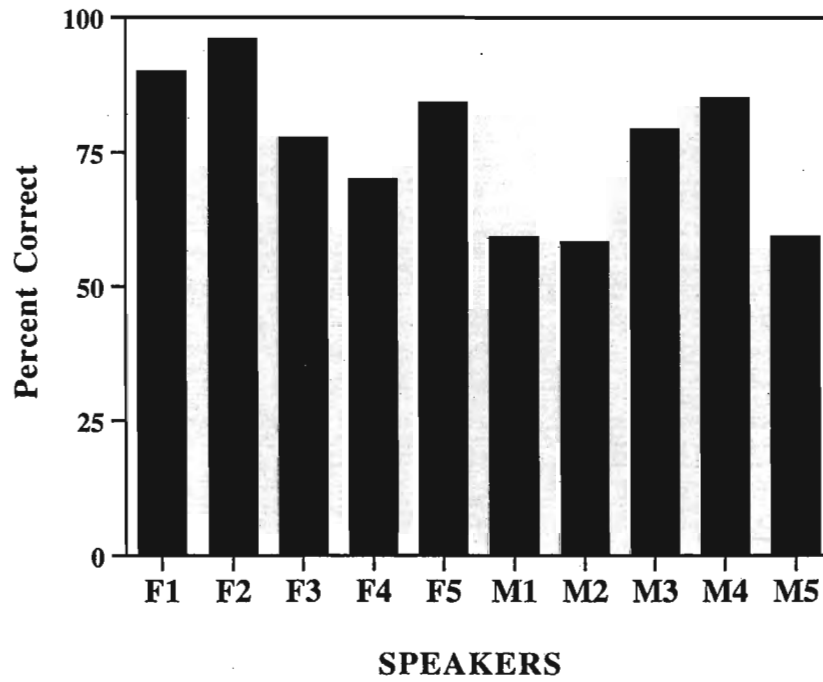


Figure 3. Speaker identification performance on sinewave replicas for the last day of training as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

of natural speech, although the rate and degree of learning varies as a function of the identifiability of different speakers within the training set.

Generalization Performance

Because there were large differences in the identifiability of different speakers within the training set, the statistical analysis of the generalization tests was conducted using generalization scores to normalize for different levels of performance. These scores were obtained by dividing the talker identification accuracy on the generalization test performance by talker identification accuracy on the training task.

At the time of the generalization testing, half of the subjects received the natural speech generalization test before the sinewave generalization test, whereas the other half completed the tests in the opposite order. To assess whether the order of the tests affected speaker recognition, an ANOVA on test order was conducted on the generalization scores from each generalization test. In both cases, the effect of test order was not significant. Consequently, the data from the two groups were pooled and all subsequent analyses ignore this factor.

Figure 4 displays the generalization scores for the natural speech and sinewave generalization tests for each speaker. The generalization data from both tests show that speaker-specific knowledge acquired during perceptual learning of sinewaves was not dependent on the specific sample used during training, but generalized to novel sentences including both natural and sinewave materials. Moreover, the same level of generalization occurred in both tests, despite the fact that in the natural speech condition, both the content and the acoustic form of the sentences differed from the items used during training. Specifically, the data showed that listeners' ability to recognize speakers decreased from 76% correct at the end of training to 46% correct for the natural test and 44% correct for the sinewave test. An ANOVA comparing the overall means from each of the three conditions revealed a significant effect, $F(7, 2) = 5.23$, $p < .0001$. Recognition was significantly different from training for the natural speech trials [$t(7) = 7.34$, $p < .001$] and for the sinewave replica trials [$t(7) = 11.18$, $p < .0001$]. However, the difference in generalization between the two generalization tests was not significant.

 Insert Figure 4 about here

One question we were interested in concerns whether the differences listeners exhibit in their ability to identify speakers during training generalizes to novel utterances. During training, female speakers were recognized more accurately than the male speakers. In contrast, although there was a numerical trend for the female speakers to be recognized more accurately than the male speakers in both generalization tests, these differences were not reliable in either condition, as shown in Figure 5.

 Insert Figure 5 about here

The training data also showed within-sex variability in speaker identification. However, smaller within-sex differences were found for the generalization test data. An ANOVA with the factor of speaker

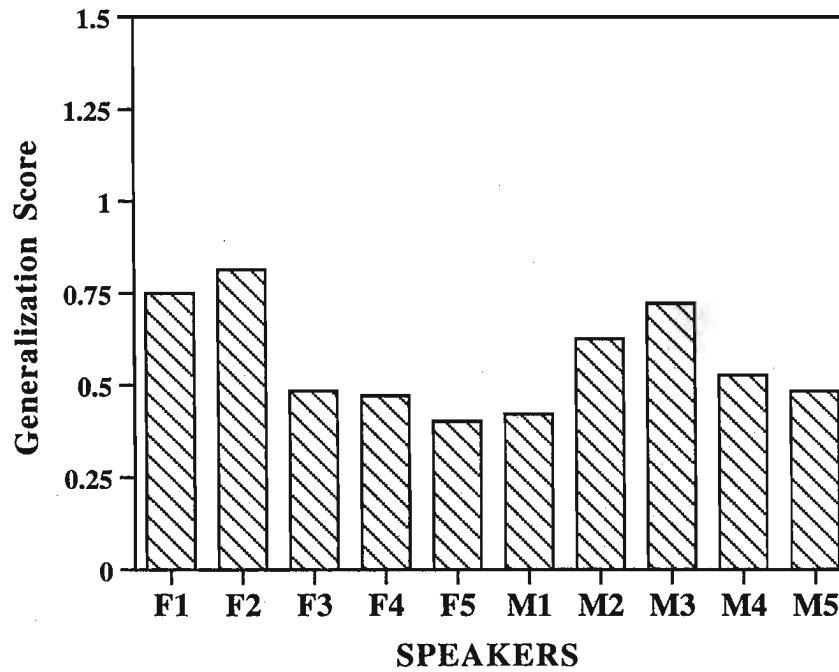
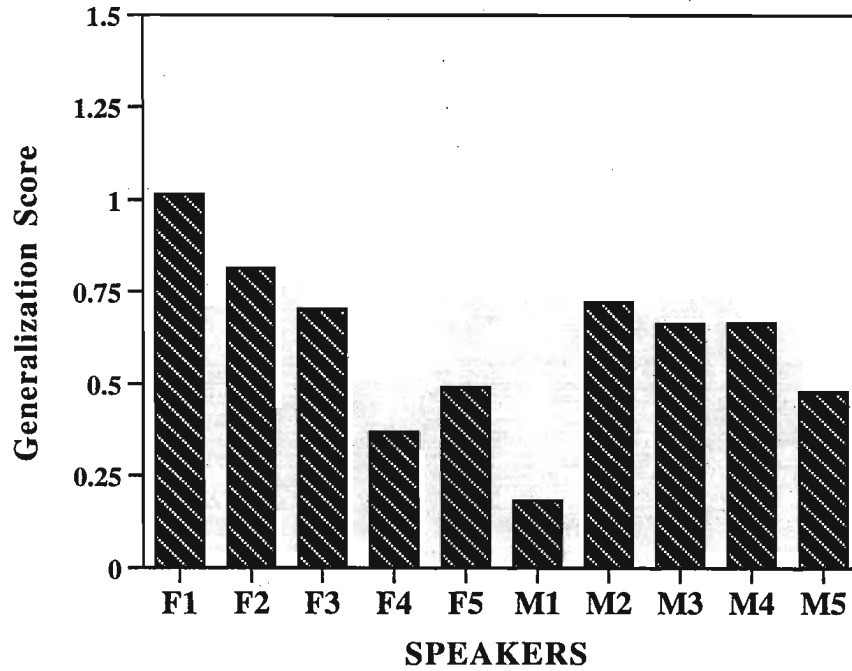


Figure 4. Mean speaker identification performance on the natural speech generalization (top panel) and sinewave replica generalization (bottom panel) as a function of training days and speaker sex. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

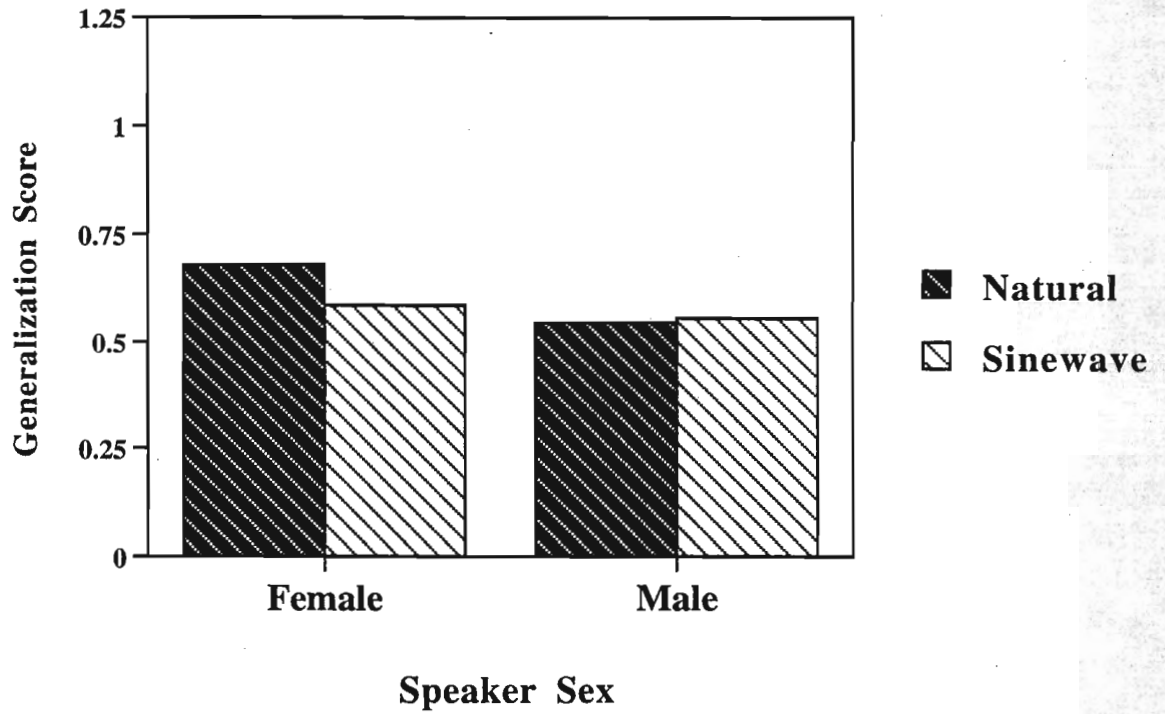


Figure 5. Mean generalization scores on the natural speech and sinewave replica generalization tests (following sinewave training) as a function of speaker sex.

was conducted on the natural speech generalization scores for each sex. The effect of speaker was marginally significant for the female speakers, $F(4, 28) = 2.63$, $p < .06$, but statistically significant for the male speakers, $F(4, 28) = 2.38$, $p < .05$. For the sinewave generalization task, reliable differences were found only among the female speakers, $F(4, 28) = 7.68$, $p < .001$. The effect of speaker did not approach significance for the male speakers.

Another way to examine the relationship between perceptual learning and speaker generalization is to examine the relative ranking of the speakers across each condition. The data show a modest overall rank order correlation for the ten speakers' identifiability at training and at test (Spearman's $\rho = .542$ for natural speech; $.492$ for sinewave stimuli), suggesting that the speakers that were most easily identified at training were also likely to be well identified at test. Curiously, the correlation between each of the generalization tests was quite high (Spearman's $\rho = .830$), indicating that the identifiability of a particular talker was more similar across the two generalization tests than across the sinewave training and generalization test.

In summary, we controlled speaker familiarity in this experiment by training all our listeners to a specific talker identification accuracy level. The motivation for this was to determine if the variability in speaker identification found by Remez et al. (in press) arose from differences in the discriminability of speakers within the speaker ensemble, or, instead, from differences in a priori speaker familiarity. In the present study, as in Remez et al., we observed striking differences in the identifiability of speakers within our training set. This suggests that perceptual distinctiveness or discriminability of the speakers in the set is the source of the discrepancies in speaker identification performance. In addition, we found that listeners can be trained to identify different speakers from nonspeech tonal analogs of speakers' utterances. This shows that talker identification can be accomplished solely from phonetic attributes present in these sinewave replicas, in the absence of voice quality. We also found evidence that perceptual learning from the sinewave training task generalized to novel natural and novel sinewave sentences. The fact that generalization was equivalent across the two generalization tests shows that the perceptual learning was not context-specific, and suggests the possibility that the same acoustic correlates of voice identity are being utilized in both the sinewave and natural speech generalization tests. It would be useful to determine whether the symmetry in generalization performance that occurred following training on the sinewave utterances is particular to that perceptual learning task, or if the pattern of generalization can be found following perceptual learning of natural voices. Experiment 2 provides this comparison.

Experiment 2

The design and method of Experiment 2 was identical to Experiment 1, except that the training sentences were natural speech utterances, rather than sinewave replicas. The natural sentences were used to train listeners until they were able to identify the 10 speakers with at least 70% accuracy. Generalization of speaker knowledge was then assessed using natural speech and sinewave generalization tasks in which listeners heard new sentences and identified the speaker. Based on previous research (Nygaard & Pisoni, 1995), we predicted that the perceptual learning of talkers from natural speech sentences would proceed rapidly and would readily generalize to novel natural speech sentences. Furthermore, the findings from Remez et al. (in press) and from Experiment 1 lead us to expect that training on the natural speech samples would facilitate speaker identification from sinewave replicas.

Method

Subjects

Eight new subjects were recruited from the Bloomington, Indiana, community. All subjects were native speakers of American English and reported no history of a speech or hearing disorder. Subjects were paid for their participation.

Test Materials

The 90 natural and 90 sinewave sentences used in Experiment 2 were identical to the sentences used in Experiment 1. Three sentences were randomly selected (without replacement) for the training, natural speech generalization and sinewave speech generalization tasks. All sentences were rotated through all conditions for each subject to eliminate any effects due to specific stimulus items or order effects.

Procedure

Training Phase

Listeners were trained to learn to explicitly identify the names of 10 speakers from the natural speech sentences uttered under the same training conditions used in Experiment 1. Again, training was continued until subjects achieved an average of 70% correct talker identification performance.

Familiarization Phase

The familiarization task preceded the natural speech and sinewave replica generalization tests. A random ordering of 30 natural speech training items was presented for speaker identification. Subjects received feedback after each response using the same methods as in Experiment 1.

Generalization Tests

The two generalization tests (natural speech and sinewave replica) were identical to those used in Experiment 1. Each generalization test presented five repetitions of three novel sentences in a random order (150 items total) for speaker identification. Subjects were informed about the sentences they would be hearing before the start of each test. They were also instructed to attend to the talker's voice and to identify the speaker by pressing the appropriate button on the keyboard. There was no feedback. The test order was counterbalanced across the subjects.

Results and Discussion

Training Performance

As expected, listeners learned to identify the ten different speakers from the naturally produced sentences very rapidly. Identification performance reached criterion for five of the eight subjects after only one training session. The remaining three listeners reached criteria by the end of the second session. Speaker identification performance averaged 78%.

The training data also revealed differences in the identifiability of female and male speakers. Figure 6 displays identification performance on the last day of training as a function of speaker. The graph shows that overall, female talkers were identified better than male talkers (87% vs. 71% correct for female and male speakers, respectively). In fact, only one male speaker ("M4") exceeded the identification accuracy of the female speakers. An ANOVA comparing speaker recognition performance on the last day of training revealed a significant effect of speaker sex, $F(1, 78) = 17.5, p < .0001$.

Insert Figure 6 about here

Variability in talkers' identifiability was also observed within each sex. An ANOVA with the factor speaker was conducted on the training scores for each sex. Reliable differences were found among the female speakers, $F(4, 28) = 2.90, p < .05$, and the male speakers, $F(4, 28) = 3.14, p < .05$. These findings with natural speech are similar to the observed in Experiment 1 using sinewave replicas. In sum, the natural speech training data reveal rapid, although somewhat variable, perceptual learning of the individual talkers within our training ensemble.

Generalization Performance

As in Experiment 1, because we found no differences in the mean test performance as a function of test order, the two test order groups were pooled to form a single composite test group. The statistical analysis of the generalization tests following training on natural speech was conducted on generalization scores.

Figure 7 displays the generalization scores for the natural speech and sinewave replica generalization tests for each speaker. The data for the two generalization tests differed markedly. Listeners' ability to recognize individuals was 88% for the natural speech generalization test but only 27% for the sinewave generalization test. An ANOVA comparing the overall means from each of the three conditions (training, natural and sinewave) revealed a significant effect, $F(7, 2) = 5.23, p < .0001$. Surprisingly, performance on the training task (78% correct) was reliably *lower* than performance on the natural speech test [$t(7) = 3.17, p < .05$]. This effect may be the result of the familiarization task preceding the generalization tests. The purpose of the familiarization task was to remind subjects of the correspondence between a particular name and talker, and is itself an abbreviated training task. Although the familiarization task only presents 30 items, it nevertheless increased listeners talker knowledge, allowing them to perform better on the subsequent generalization test than on training task. Training performance was, however, significantly higher than performance on the sinewave replica generalization test [$t(7) = 21.5, p < .0001$]. In addition, the two generalization tests differed reliably from each other [$t(7) = 22.89, p < .0001$].

Insert Figure 7 about here

Female and male voices were recognized with equal accuracy in both generalization tests, as shown in Figure 8. An ANOVA with the factor of speaker sex was conducted separately on the natural speech generalization scores and the sinewave replica generalization scores. The effect of speaker was not significant in either of the generalization tests.

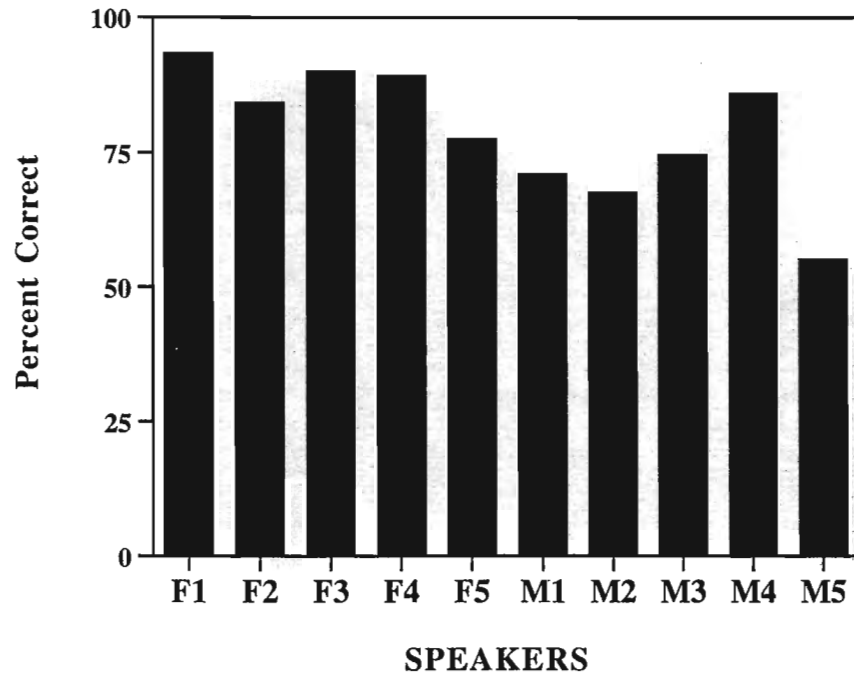


Figure 6. Mean speaker identification performance on natural speech for the last day of training as a function of speaker. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

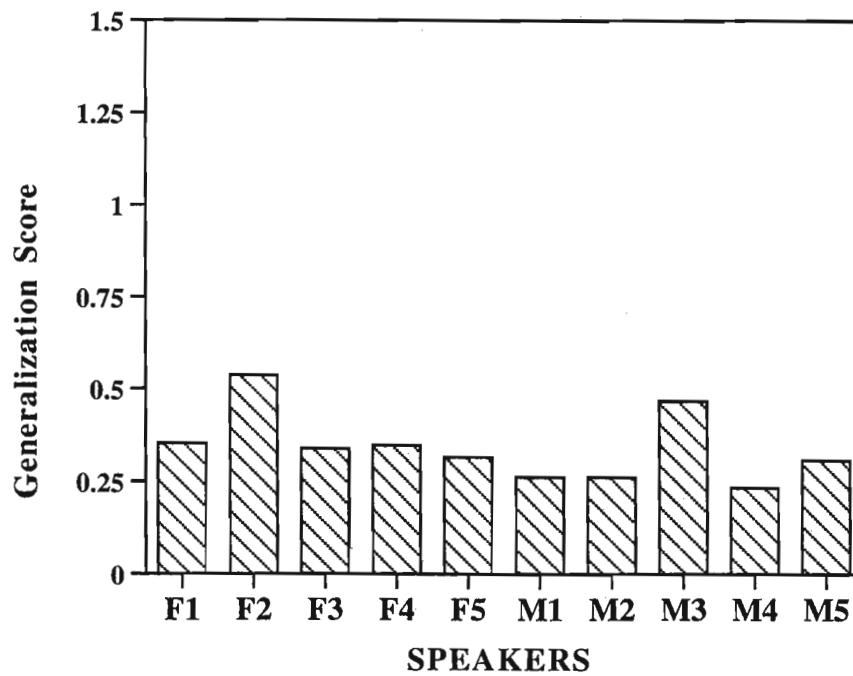
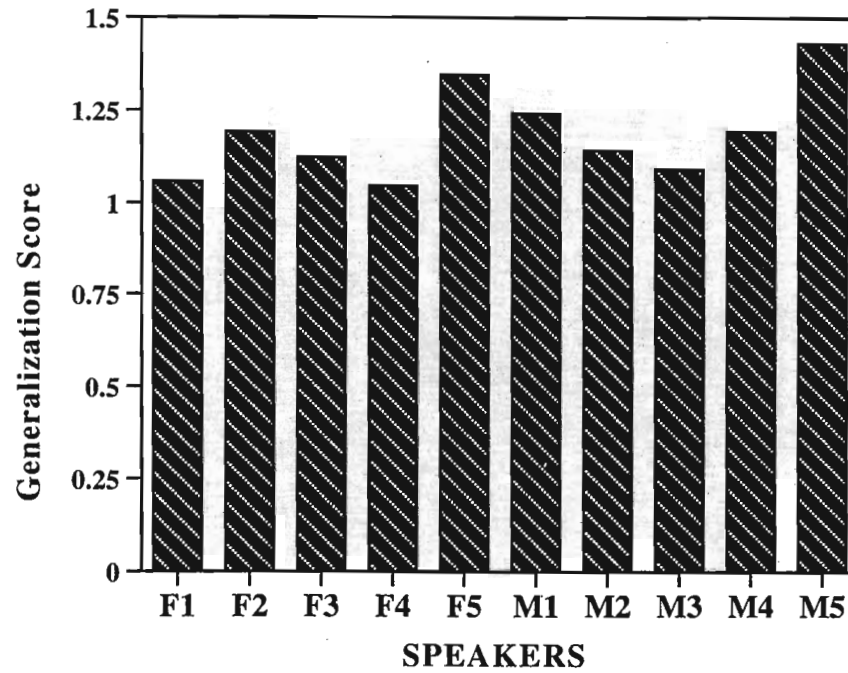


Figure 7. Mean speaker identification performance on the natural speech generalization (top panel) and sinewave replica generalization (bottom panel) as a function of training days and speaker sex. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

 Insert Figure 8 about here

There were also no differences in the identifiability of voices among the female and male speakers in either generalization test condition, as shown in Figure 8. An ANOVA with the factor speaker was conducted on the natural speech generalization scores for each sex. The effect of speaker was marginal for the female speakers $F(4, 28) = 2.58, p < .06$. No significant differences were found among the male speakers. For the sinewave replica generalization task, the effect of speaker did not approach significance for either the female or the male speakers.

A Spearman's rho correlation was conducted to assess the relationship between perceptual learning and generalization performance. The analysis indicated that the identifiability of a particular talker was most similar across the natural training and natural test conditions (Spearman's rho = .752), less similar between training and the sinewave test (Spearman's rho = .515), and least similar across the two generalization tests (Spearman's rho = .424).

In summary, Experiment 2 shows that subjects easily learned to identify a talker from naturally produced sentences, and that this knowledge generalized readily to novel natural speech utterances. In contrast, subjects had difficulty recognizing speakers from novel sinewave utterances, although overall performance was marginally above chance. The implications of these findings will be considered in detail in the general discussion section.

General Discussion

The findings from these two experiments demonstrate that perceptual learning of a talker's voice can take place in the absence of traditional acoustic correlates of voice quality. In Experiment 1, we found that listeners were gradually able to learn to identify and label speakers from sinewave replicas of naturally produced sentences. This result, in conjunction with the recent findings of Remez et al. (in press), demonstrate that information about the dynamics of a talker's vocal tract transfer function include talker-specific information, and that this information can be exploited during perceptual learning. The perceptual training data from both experiments also revealed considerable variability in the identifiability of different speakers. In particular, female speakers from this set of voices were identified more accurately than male speakers, and this advantage for female speakers was equal in both experiments (16%). However, variability in the identifiability of different speakers was less evident in the generalization tests. In both experiments, differences in the overall identifiability of female and male speakers in either test condition were either small or absent.

We also found evidence that the knowledge obtained from the perceptual learning task can be readily generalized to natural and sinewave speech. In particular, subjects who learned sinewave samples were able to identify a talker from novel sinewave replicas and from novel natural speech samples with equal accuracy. Moreover, variability in the learning of different speakers also generalized well to both tasks. In fact, the ranking of talkers based on identification accuracy was more similar between the sinewave training and natural speech test than between the sinewave training and the sinewave generalization test.

We found exactly the opposite pattern of results in Experiment 2. For subjects who were trained on the natural speech sentences, stimulus generalization was greatest in the condition in which the perceptual

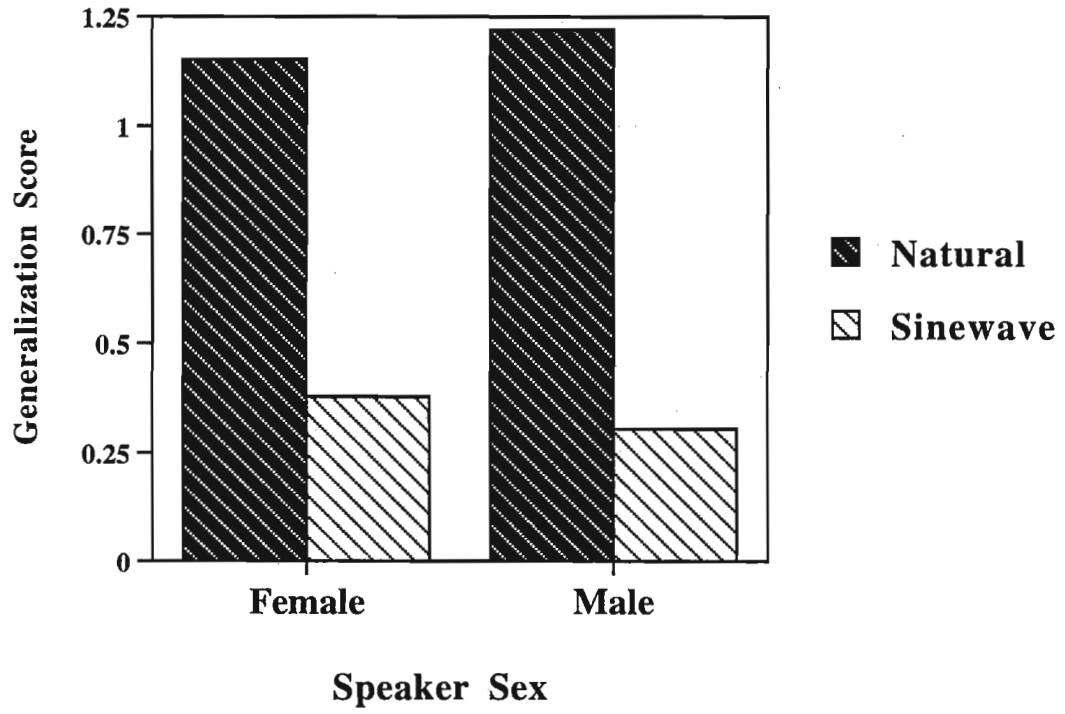


Figure 8. Mean generalization scores on the natural speech and sinewave replica generalization tests as a function of speaker sex.

form of the sentences was the same across training and test. The data show that talker identification at training was similar in magnitude and in patterning to the natural speech test, but quite different from the sinewave speech generalization test. Specifically, overall speaker identification performance and the identifiability of a particular talker was more similar across the natural training and natural test conditions than across the natural training and sinewave test conditions. However, accuracy identifying a speaker on the sinewave generalization test was very poor, although above chance.

A question to ask, then, is why the performance of our subjects was considerably worse than the performance of subjects in Remez et al. (in press). In the present study, listeners' ability to recognize individuals was only 27% for the sinewave generalization test, but approximately 55% in the talker identification test used by Remez et al (see their Figure 4). Four methodological differences in our experiment and theirs may account for the difference in sinewave speaker identification performance across these experiments. First, in our task, listeners heard three novel sentences three times during the generalization test, whereas in Remez et al. (in press) the same sinewave sentence from each of the ten talkers was presented six times. In this case, the fact that the linguistic content of the utterance was the same from trial to trial may have improved listeners ability to discriminate among the ten individuals. Second, our listeners did not have any knowledge or familiarity with sinewave speech before the generalization test task, whereas several of the listeners used by Remez et al. (in press) were familiar with what sinewave speech was and how it sounded. This may have increased subjects' ability to quickly focus their attention on the talker-specific phonetic information present in the sinewave replicas. Finally, in our experiment, familiarity was acquired through perceptual training with a small number of sentences over a few days. In contrast, the listeners in Remez et al. (in press) had acquired their speaker knowledge naturally from many hundreds or even thousands of utterances over the course of many years. We would expect, therefore, that they would have qualitatively richer speaker representations than the listeners in the present study.

One implication of these observations is that talker-specific generalization depends not only on familiarization with the specific acoustic attributes of a talker's voice, but also familiarization with the particular acoustic format in which a speaker is presented. For example, in Experiment 1, subjects became familiar with the sinewave materials during the course of training and were already highly familiar with natural speech. Consequently, they may have been better able to focus attention on the acoustic-phonetic information specific to each speaker during the generalization test phase. However, subjects in Experiment 2 had no prior experience with sinewave utterances. The novelty of the sinewave test items may have diverted their attention away from the voice-specific phonetic information in these highly unnatural sound patterns. In addition, because sinewaves are only an abstract representation of familiar natural speech properties, subjects may have had difficulty perceiving commonalities across the two different perceptual formats. A particular sinewave speaker may have failed to remind listeners of a familiar natural voice simply because the former was not perceived to be perceptually similar to a known speaker category. Experiments designed to determine the role of attention on the acquisition and generalization of speaker knowledge are currently underway in our laboratory.

In summary, the present experiments reveal an asymmetry in the generalization of speaker knowledge from natural and sinewave utterances. Speaker-specific knowledge acquired during training on sinewaves shows generalization to novel sinewave sentences as well as naturally produced utterances, whereas speaker-specific knowledge acquired during training on natural speech does not show generalization to sinewave utterances. The results also showed that variability in the degree of perceptual learning affected generalization of speaker knowledge to novel natural and sinewave sentences. Finally, the experiments showed that perceptual learning of a talker's voice can occur even when specific acoustic

products of vocal articulation are eliminated from the signal, and suggest that attention plays an important role in learning and generalization of speaker-specific knowledge.

References

- Bricker, P.D., & Pruzansky, S. (1976). Speaker recognition. In N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 295-326). New York: Academic Press.
- Halle, M. (1985). Speculations about the representation of words in memory. In V.A. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 101-114). New York: Academic Press.
- Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K.R. Scherer & H. Giles (Eds.), *Social Markers in Speech* (pp. 1-32). Cambridge, UK: Cambridge University Press.
- Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, *47*, 379-390.
- Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42-46.
- Nygaard, L.C. & Pisoni, D.B. (1995). Talker and task-specific perceptual learning in speech perception. *Proceedings of the XIIIth International Congress of Phonetic Sciences*. Stockholm: Stockholm University, *1*, 194-197.
- Remez, R.E. Rubin, P.E., Pisoni, D.B., & Carroll, T.D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947-950.
- Remez, R.E., Fellowes, J.M., & Rubin, P.E. (In Press). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*.
- Rubin, P.E. (1980). *Sinewave Synthesis*. Internal Memorandum, Haskins Laboratories, New Haven CT.