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

Experience with Sinewave Speech and the Recognition of Sinewave Voices[1]

Sonya M. Sheffert, David B. Pisoni, Nathan R. Large, Jennifer M. Fellowes[2] and Robert E. Remez[2]

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, New York.
[2] Department of Psychology, Barnard College, New York, NY.

Experience with Sinewave Speech and the Recognition of Sinewave Voices

Abstract. This study explores the learning and generalization of voices using sinewave speech patterns. The present experiment was designed to address an asymmetry in generalization performance reported in a previous perceptual learning study (Sheffert, Pisoni, Fellowes & Remez, 1996). In that study, listeners who became familiar with a talker's voice through listening to sinewave replicas of naturally produced sentences showed excellent generalization to both novel sinewave and natural speech sentences, whereas speaker-specific knowledge acquired through natural speech only generalized to natural speech sentences. The primary difference between the two experiments was the amount of experience subjects had with sinewave materials before the generalization tests. The present experiment tested the hypothesis that the ability to recognize familiar voices from sinewaves depends on having prior phonetic experience with the unusual acoustic properties of sinewave signals. The experiment consisted of three phases: Training, transcription, and generalization. In the training phase, listeners learned to categorize ten talkers from naturally produced sentences. In the transcription phase, listeners were familiarized with sinewaves by transcribing sinewave sentences produced in an unfamiliar voice. In the generalization phase, listeners' ability to recognize the ten talkers from a novel naturally produced sentences and sinewave replicas was assessed. The results confirmed the earlier findings of Sheffert et al. (1996) by demonstrating that speaker-specific knowledge acquired during the perceptual training task generalized readily to novel natural utterance, but not to novel sinewave utterances. The data also show that prior exposure to the unusual nonspeech tonal patterns did not improve generalization performance. This pattern of results demonstrates that subjects' ability to exploit the talker-specific phonetic information present in the sinewave replicas does not depend on having phonetic experience with sinewave speech patterns.

Introduction

The present experiment is concerned with the perceptual learning and generalization of voice information. It is part of a series of experiments aimed at determining what properties of the acoustic signal support familiar voice recognition. This particular study is a control experiment designed to address an asymmetry in generalization performance found in two previous perceptual training experiments (Sheffert, Pisoni, Fellowes & Remez, 1996). In one study, sinewave replicas of naturally produced utterances were used to train listeners to identify 10 talkers. The sinewave signals were composed of three time-varying sinusoids that reproduced the center frequency and amplitude patterns of the oral, nasal and fricative resonances of the natural utterance. These nonspeech patterns preserved coarse-grained properties of the talker's vocal tract transfer function, including idiosyncratic phonetic variation, while eliminating cues to voice quality, such as fundamental frequency, harmonic structure and the fine-grained power spectra of nasals and vowels.

Surprisingly, subjects were able to learn to identify the ten talkers from these highly impoverished nonspeech sounds, although it took several training sessions. This result is important because it shows that perceptual learning of voices can be accomplished solely from the time-varying phonetic attributes encoded in the vocal tract transfer function. That is, the acoustic cues typically implicated in talker identification (Bricker, & Pruzansky, 1976; Laver, 1991) are not necessary for voice learning.

Our earlier experiment also showed that subjects who learned the voices from sinewave patterns were able to identify a talker from novel sinewaves and from novel natural speech samples with equal accuracy. This result indicates that the same acoustic correlates of voice identity were being utilized in both the sinewave and natural speech generalization tests, and supports the proposal that phonetic properties of an utterance jointly specify words and talkers (Remez, Fellowes & Rubin, 1997).

A very different pattern of generalization was found in a second experiment in which subjects learned to identify the talkers from natural utterances. The results showed that subjects quickly learned to identify a talker from naturally produced sentences, and that this knowledge generalized readily to novel natural speech utterances. However, speaker-specific knowledge did not generalize to novel sinewaves. In fact, the accuracy of identifying a speaker on the sinewave generalization test was only marginally above chance. Not only does this finding contrast with the generalization pattern in the first experiment, but it also contrasts with previous data from Remez et al. (1997).

A methodological difference that may account for the differences observed across the two training experiments is the amount of exposure subjects had to sinewave materials prior to the sinewave generalization test. In Experiment 1, subjects had several days of training with the sinewaves and were accustomed to the peculiar acoustic quality of the tonal analogs and were able to perceive the tonal analogs as speech-like. Similarly, the listeners used in Remez et al. (1997) were largely composed of speech scientists who had prior exposure to sinewave signals In contrast, subjects in Experiment 2 had no prior experience with sinewaves. It is possible that the novelty of the sinewaves led subjects in Experiment 2 to focus on the unusual auditory impressions the tonal signals evoke, rather than on the talker-specific phonetic information present in the signals. Failing to attend to the transfer-relevant phonetic information in the sinewaves may have prevented the listeners from showing learning on the sinewave voice recognition test.

The present experiment tested the hypothesis that the ability to recognize familiar voices from sinewaves depends on having prior experience with sinewave signals. We attempted to resolve this issue by providing experience to our listeners using a sinewave transcription task. This task follows the natural speech perceptual training task but precedes the generalization tests. The purpose of the sinewave transcription task was to accustom listeners to the unusual timbre of the signals and to facilitate the perception of the phonetic information in the sinewaves. Subjects were provided with several repetitions of 29 sinewave replicas. None of these items duplicated any of the items used in the training or transcription tasks. In addition, the sentences were produced by a male voice who was not one of the talkers in the training or generalization phases of the experiment (R.E.R). Subjects were informed that the sinewaves were analogs of the natural speech tokens and were encouraged to listen to them as they would speech. The participants simply listened to each sentence and transcribed what they heard.

The general design of the experiment was identical to Experiment 2 in Sheffert, et al. (1996), with the exception of the transcription task. Specifically, the familiarity of the ten speakers was experimentally manipulated by teaching listeners to identify by name the talkers producing the natural speech utterances using a feedback-driven supervised training procedure. The sentences were sinewave replicas of the natural utterances. Subjects were trained (with feedback) until they were able to identify the ten talkers from the natural speech samples with at least 70% accuracy. After the training phase, subjects completed a sinewave transcription task designed to accustom them to the unusual acoustic qualities of sinewave replicas.

We compared two transcription presentation methods that differed in the amount of trial-by-trial variation among the sentences. For half the subjects, three repetitions of each sentence were presented randomly in the list, and subjects transcribed the entire sentence after each trial ("random presentation" condition). For the other half of the subjects, five repetitions of each sentence was presented consecutively in a blocked fashion ("blocked presentation" condition). Subjects were required to transcribe the entire sentence after the last repetition, although they could note partial information during any repetition. We assumed the latter method would facilitate transcription performance because there would be fewer trial-to-trial changes in the items, and more opportunities to attend to the sentence.

Voice recognition was then assessed using two generalization tasks in which listeners heard a new 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. No feedback was provided. If prior exposure to sinewave speech is an important determinant of performance on the sinewave generalization test, there should be a gain in accuracy relative to the previous condition which did not include a transcription task (Experiment 2 from Sheffert et al., 1996).

Experiment 1

Method

Subjects

Twenty adult subjects were recruited from the Bloomington community. Of these, three subjects failed to complete the study due to work or school commitments, and one was excused because of a possible hearing impairment. The remaining sixteen subjects completed the natural speech training phase, the sinewave transcription task 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

Three types of sentences were used in the present experiment. The natural speech sinewave test sentences were the same items used by Sheffert at al. (1996), and developed by Remez et al. (1997). One set of sentences consisted of nine natural utterances produced by five male and five female talkers. Each talker produced all nine sentences, bringing the total number of sentences to 90. Audio recordings were obtained by asking speakers to read the sentences aloud in their natural speaking style. The sentences were 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. Formant values were obtained using two measures interactively: 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.

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.

The third set of sentences were sinewave tokens used for the transcription test. These items consisted of 29 tonal analogs derived from natural speech utterances produced by a male speaker who was not part of the training set. The sentences were recorded, digitized and stored in the same manner as the stimulus materials described above.

Procedure

Training Phase

Listeners were trained to identify the names of the 10 speakers using the natural speech utterances. Subject testing was conducted for groups of three or less in a quiet listening room. During each training session, subjects heard a random ordering of five repetitions of three sentences from each of the 10 talkers (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 in a given session. The natural speech 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' voice. Each time a sentence was presented, the subject was required to press one of ten keyboard buttons, each of which was labeled with a 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 by the computer. Each training session lasted approximately 30 minutes. Training was continued until subjects achieved an average of 70% correct speaker recognition performance at the end of a session.

Sinewave Transcription Task

In this phase of the experiment, subjects completed the sinewave transcription task under the same listening conditions as the training task. The transcription task presented 29 sinewave sentences based on the natural speech production of a talker (R.E.R.) who was not part of our training ensemble. Subjects were told that the purpose of the task was to measure their ability to identify words in synthesized sentences. We encouraged the participants to listen in their "speech mode" by telling them that although the sentences sound very unnatural, they were in fact real speech that had been processed by a computer. Subjects were told that there was only one "voice" speaking the sentences. The listeners were instructed to transcribe the sentences as accurately as possible on a response form. They were not given feedback on their performance in this phase of the experiment.

Listeners in the "random presentation" group were told that each sentence would be repeated three times throughout the list. They were instructed to write down as many words or partial words as they could after each sentence, and guess if needed. The task was self-paced in order to give the listeners plenty of time to process and transcribe the sentences.

Subjects in the "blocked presentation" group were given the same instructions, with the exception that they were told that each sentence would be repeated five times in a row. Although subjects were required to write down the entire sentence after the last repetition, they were allowed to write down words or parts of words at anytime during the five repetitions. Listeners were given as much time as needed to listen to and write down the sentences. The transcription tasks took 20 to 30 minutes.

Familiarization Phase

Each of the generalization tests was preceded with a brief familiarization task designed to remind subjects of the correspondence between the training tokens and the names of the speakers. The task was simply an abbreviated version of a training session in which subjects listened and responded to one instance of each training sentence from each talker (30 items total). As in the training phase, 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 natural speech 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 subjects. Half the subjects received the natural generalization test before the sinewave generalization test, whereas the other half received the tests in the opposite order. Each generalization test presented five repetitions of each of the three sentences from each talker 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 talker 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. Each generalization test lasted approximately 30 minutes.

Results

Training Performance

Examination of the training data revealed that listeners had little difficulty learning to identify the speakers from the natural sentences. All listeners reached criteria by the end of the third session. Seven subjects reached criterion after only a single training session. The mean number of training days was 1.75. Speaker identification performance on the last day of training averaged 79% for the subjects in the random presentation group, and 80% for subjects in the blocked presentation group. These values do not differ statistically from one another nor do they differ from the training data reported in Sheffert et al. (1996). The training data show that talker-specific aspects of the speech signal are easily attended to and that voices can be learned easily from sentence length materials.

Since there were no differences in the training performance across the two groups of subjects, all subsequent analyses of the training data were conducted on the combined scores from both groups of participants (N=16). Figure 1 displays identification performance on the last day of training as a function of talker. Overall, female talkers were identified better than male talkers (89% vs. 70% correct for female and male speakers, respectively). An ANOVA comparing talker identification performance on the last day of training revealed a significant effect of speaker sex, $F(1, 158) = 46.98$, $p < .0001$, confirming that the female speakers were more accurately identified than the male speakers.

Insert Figure 1 about here

The training data also revealed variability in the identifiability of different speakers within each sex. An ANOVA was conducted on the training scores for each sex. Reliable differences were found among the female speakers, $F(4, 60) = 4.44$, $p < .003$, and the male speakers, $F(4, 60) = 7.82$, $p < .0001$. Taken together, the natural speech training data show variability in the ease with which certain voices are identified. The patterning of the data across talkers is almost identical to the previous results of Sheffert et al. (1996).

Sinewave Transcription Performance

Performance on the sinewave transcription test was based on the mean number of syllables correctly identified in each sentence. Mean transcription performance for subjects in random presentation group was 49% compared to 42% for subjects in the blocked presentation group. This difference was not statistically significant.

To assess the effects of the different transcription methods on speaker recognition, we compared the data from the present experiment with the data from Experiment 2 in Sheffert et al. (1996). This earlier experiment did not include a sinewave transcription task. Table 1 displays the data from these three conditions. The table shows that the data from present experiment replicates the findings of Sheffert et al. Positive transfer was observed to the novel natural speech utterances, whereas very little transfer was observed to novel sinewave utterances. In addition, there was no benefit from the transcription task.

Table 1.

Proportion correct transcription and speaker recognition performance as a function of transcription presentation method.

	Transcription Presentation Method		
	Random	Blocked	None*
Transcription	.49	.42	--
Training	.79	.80	.78
Natural Test	.82	.87	.88
Sinewave Test	.26	.28	.27

Note. * Data from Experiment 2 of Sheffert et al. (1996).

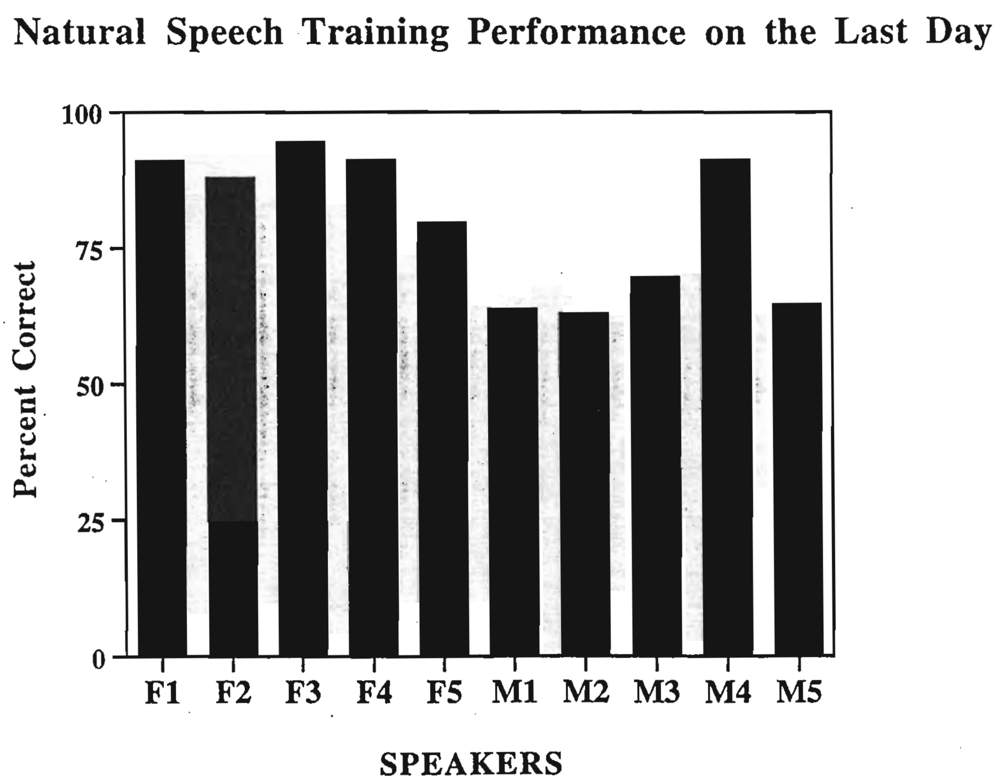

Figure 1. Mean speaker identification performance on natural speech for the last day of training as a function of speaker. F1 though F5 refer to the female speakers; M1 through M5 refer to the male speakers.

Separate ANOVA's were performed on the sinewave transcription scores, natural speech training performance, natural speech generalization and sinewave replica generalization. The statistical analysis confirms what is apparent in Table 1. For all measures, we found no differences in the mean test performance as a function of transcription method. Consequently, this variable was ignored in all further analyses and the data from the random and blocked transcription conditions were combined into a single group.

Generalization Performance

Because of 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. The generalization scores were obtained by dividing the talker identification accuracy on the sinewave and natural speech generalization tests 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. 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 group. The statistical analysis of the generalization tests following training on natural speech was conducted using generalization scores.

Figure 2 displays the generalization scores for the natural speech and sinewave replica generalization tests for each speaker. As in the previous perceptual training study, performance on the two generalization tests were very different. Specifically, speaker-specific knowledge acquired during the perceptual learning phase generalized to novel natural speech sentences, but not to novel sinewave replicas. Listeners' ability to recognize individuals from natural speech samples was 85%, as compared to only 27% for the sinewave samples. An ANOVA comparing the overall means from each of the three conditions (training, natural and sinewave) revealed a significant effect, $F(2, 45) = 179.88$, $p < .0001$. Planned comparisons revealed that training performance was significantly higher than performance on the sinewave replica generalization test [$t(15) = 19.82$, $p < .0001$], but was not reliably different than performance on the natural speech test. In addition, the two generalization test differed reliably from each other [$t(15) = 24.91$, $p < .0001$].

Insert Figure 2 about here

Male and female voices were recognized equally in both generalization tests, as shown in Figure 3. 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.

Insert Figure 3 about here

Similarly, 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

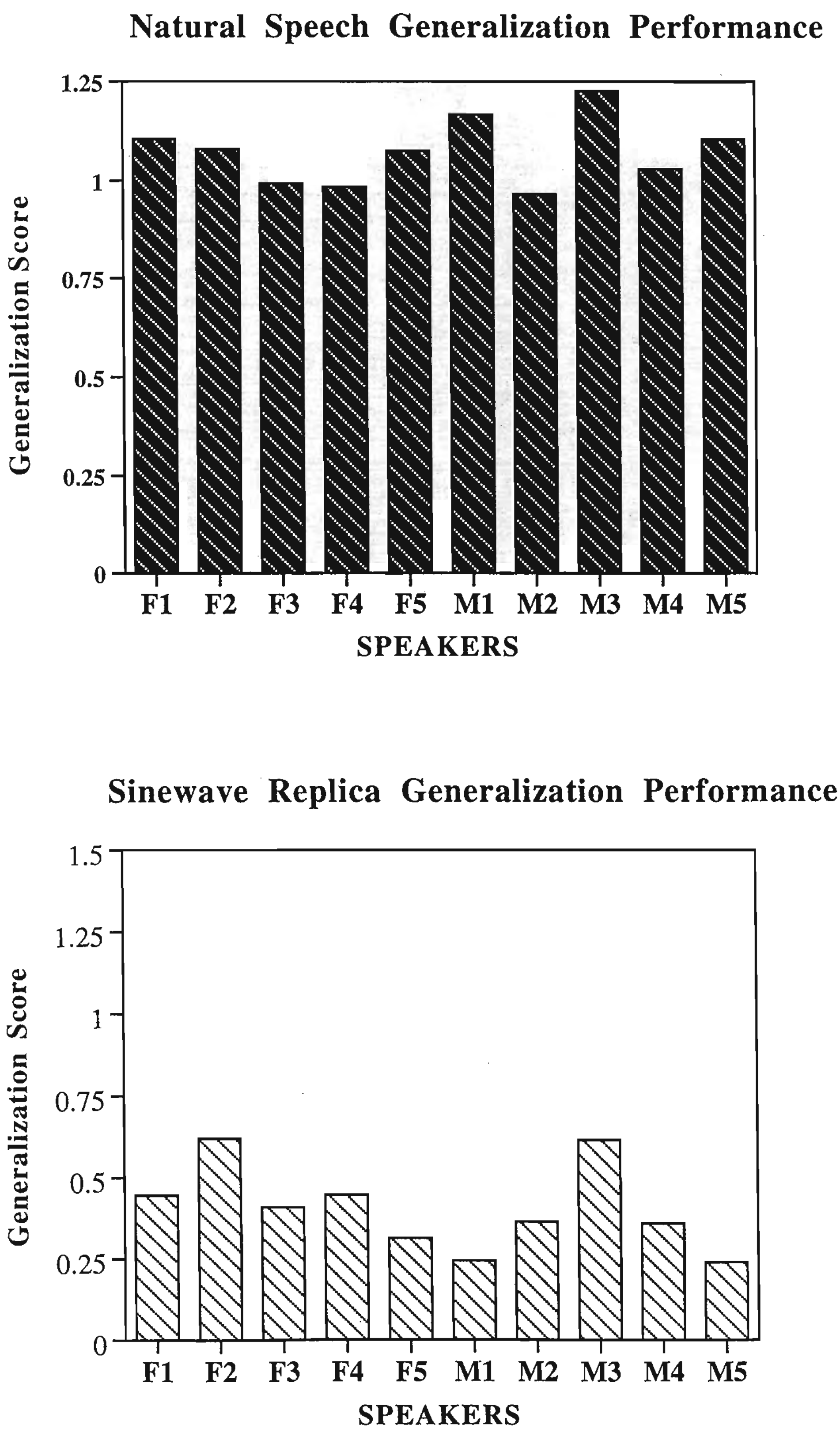

Figure 2. Mean speaker identification performance on the natural speech generalization (top panel) and sinewave replica generalization (bottom panel) as a function of training days and speaker sex. F1 through F5 refer to the female speakers; M1 through M5 refer to the male speakers.

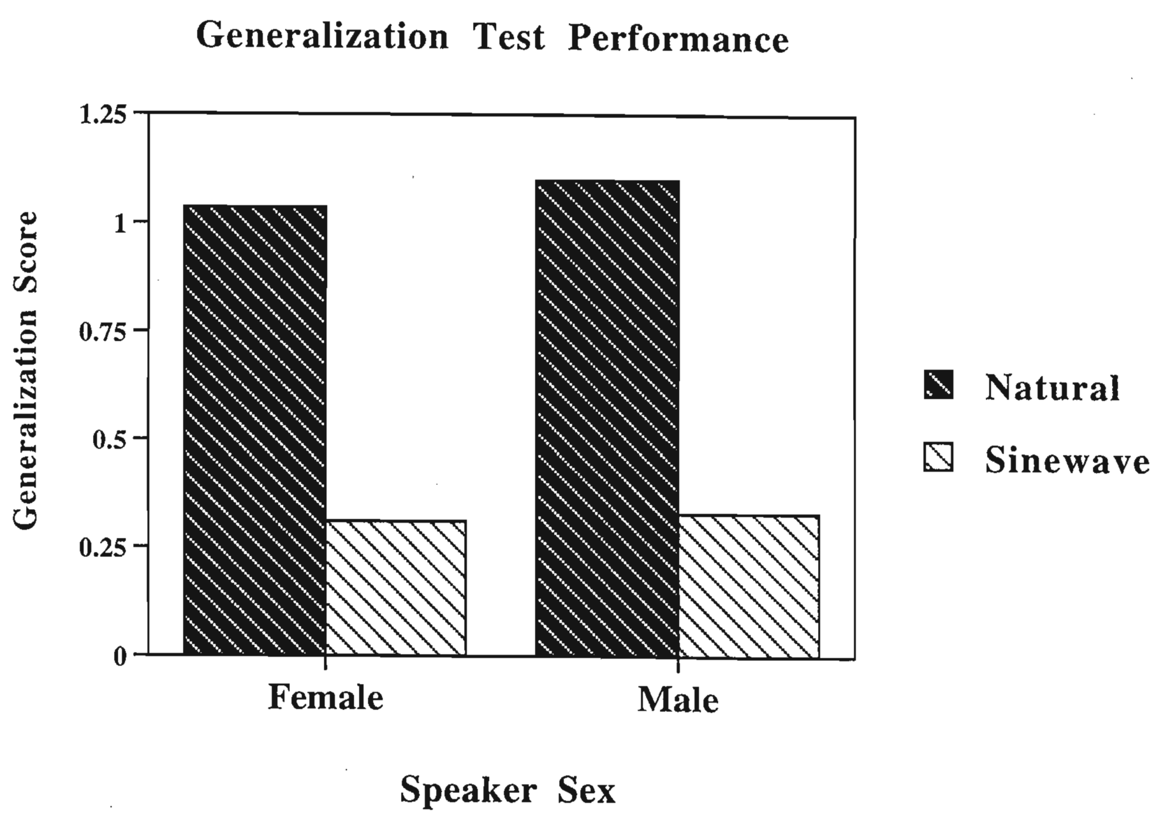

Figure 3. Mean generalization scores on the natural speech and sinewave replica generalization tests as a function of speaker sex.

was conducted on the natural speech generalization scores for each sex. The effect of speaker did not approach significance for either the female or the male speakers. No significant differences were found among the male speakers. For the sinewave replica generalization task, the effect of speaker was significant both for the female speakers $F(4, 60) = 4.66$, $p < .002$, and for the male speakers $F(4, 60) = 4.18$, $p < .005$.

The relationship between perceptual learning and generalization performance was assessed by comparing the relative ranking of the speakers across each condition. A Spearman's rho correlation was conducted between the training and generalization conditions. The analysis indicated that the speakers that were most easily identified during training were also the same speakers most easily identified in the natural test conditions (Spearman's rho = .891). The identifiability of talkers was less similar across the two generalization tests, as indicated by the modest correlation between the two condition (Spearman's rho = .552). The identifiability of talkers was least similar between training and the sinewave test (Spearman's rho = .345). This is another indication that the talker-specific knowledge acquired at training had little impact on performance in the sinewave generalization test. Note, also, that this correlation is somewhat lower than correlation of .515 reported in Sheffert et al. Although one must be cautious comparing measures across experiments, the lower correlation in the present study is consistent with the idea that familiarizing listeners with sinewave does not increase the relationship between training and sinewave test performance.

Discussion

The most important finding from the present study is that prior exposure and familiarity with sinusoidal speech did not improve subjects' ability to utilize talker-specific information contained in sinewave signals. The transcription task presented numerous samples of sinewave speech, with the expectation that subjects would become accustomed to the unusual sound of sinewaves and would become attuned to the phonetic properties of the signals, which would then facilitate sinewave voice recognition. This outcome did not occur. Generalization performance in this experiment was statistically indistinguishable from generalization performance from Sheffert et al. (1996), indicating that experience with sinewaves does not improve sinewave voice recognition.

It is possible that our subjects needed more substantial phonetic experience with sinewave utterances than our transcription task provided. Given that transcription performance was not particularly high (46% overall correct word identification), it may be that transcription was only beneficial to those subjects who could reliably extract phonetic information from the sinewaves. That is, listeners who had difficulty hearing the phonetic information contained in the sinewave replicas may also have had difficulty perceiving the voice information in the signals. If this is so, one would expect transcription performance to be related to sinewave voice recognition performance.

Examination of the individual subject data argues against this possibility. One example is a subject who was a good sinewave transcriber (56% correct word identification), yet had one of the lowest voice recognition scores on the sinewave generalization test (17% correct voice recognition). In general, the best transcribers (i.e., above 70% correct) were no more likely to outperform the worst transcribers (i.e., below 30% correct) on the sinewave generalization test. Thus, the degree to which a subject was able to perceive sinewaves as speech proved to be a poor predictor of their ability to extract talker-specific phonetic information and recognize voices from these patterns.

It is also possible is that the null effect of sinewave transcription occurred because the requirements of the task (word identification) did not overlap sufficiently with the requirements of the generalization test (voice recognition). If generalization performance is determined, in part, by the extent to which retrieval conditions match learning conditions, than perhaps subjects need to become familiar with the acoustic media during the course of voice learning, rather than word identification.

An alternative to the hypothesis tested in this report is that the poor voice transfer from natural speech to sinewave replicas may be due to differences in attention. Listeners may have focused on talker-specific information that was not present in the sinewave signals. Specifically, during natural speech learning, subjects may have focused their attention on the most obvious distinctive properties of speech that cue speaker identity - pitch, timbre or other suprasegmental characteristics. This information was absent from the sinewave test sentences. Poor generalization performance would be expected if listeners learned to distinguish talkers along perceptual dimensions in the natural speech that did not map onto the perceptual dimensions preserved in the sinewave signals.

In contrast, listeners trained from sinewave utterances were forced to rely on phonetic or segmental information to distinguish talkers, since other cues to talker identity were obliterated during sinusoidal synthesis. To a lesser extent, the same may be true of the listeners in Remez et al. (1997). Given that all these listeners were trained speech scientists or linguists and used to listening "phonetically", they may have focused more attention towards the properties of the natural speech that sinewaves possess. Consequently, they were often able to recognize their colleagues from a sinewave replica.

One way to determine if the pattern of results from the present experiment as well as those obtained in Experiment 2 of Sheffert et al. (1996) were due to differences in attention would be to assess transfer to a media that preserves many of the same suprasegmental cues to voice identity as natural speech while also being acoustically unusual and impoverished, like sinewave speech. Backwards speech fulfills these requirements. Backwards speech distorts temporally based fine-grained segmental information, such as information about consonants and diphthongs. Although some phonetic information is preserved, such as vowel quality, it is not enough to support word identification. As a result, backward speech is completely unintelligible. However, temporal reversal does not distort many suprasegmental aspects of speech. Characteristics based on long-term spectra such as fundamental frequency, F0 contour, speaking rate and formant relations are largely intact. Because these cues are important for speaker identification, it is relatively easy to recognize many familiar voices from reversed speech (Van Lancker, Kreiman, & Emmorey, 1995). If subjects are primarily exploiting suprasegmental information during natural speech training, than transfer from natural to reversed speech should be better than transfer to sinewaves. In addition, reversed speech sounds very strange. If transfer is positive, we will have further evidence that the peculiarity of the acoustic signal is not be the critical determinate of transfer in voice learning. This experiment, now currently underway in our lab, should provide additional insight into the learning and generalization of talker information.

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.

Laver, J. (1991). *The gift of speech.* Edinburgh, Scotland: Edinburgh University Press.

Remez, R.E., Fellowes, J.M., & Rubin, P.E. (1996). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, **23**, 651-666.

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

Rubin, P.E. (1980) *Sinewave Synthesis.* Internal Memorandum, Haskins Laboratories, New Haven CT.

Sheffert, S. M., Pisoni, D. P., Remez, R. E. & Fellowes, J. M. (1996). Perceptual learning of natural and sinewave voices. In *Research on Spoken Language Processing Progress Report No. 20* (pp. 275-296). Bloomington, IN: Speech Research Laboratory, Indiana University

Van Lancker, D., Kreiman, J. & Emmorey, K. (1995). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics,* **13**, 19-38.