Perception & Psychophysics
1997, 59 (6), 839-849

Perceiving the sex and identity of a talker without natural vocal timbre

JENNIFER M. FELLOWES and ROBERT E. REMEZ
Barnard College, New York, New York

and

PHILIP E. RUBIN
Haskins Laboratories, New Haven, Connecticut
and Yale University School of Medicine, New Haven, Connecticut

The personal attributes of a talker perceived via acoustic properties of speech are commonly considered to be an extralinguistic message of an utterance. Accordingly, accounts of the perception of talker attributes have emphasized a causal role of aspects of the fundamental frequency and coarse-grain acoustic spectra distinct from the detailed acoustic correlates of phonemes. In testing this view, in four experiments, we estimated the ability of listeners to ascertain the sex or the identity of 5 male and 5 female talkers from sinusoidal replicas of natural utterances, which lack fundamental frequency and natural vocal spectra. Given such radically reduced signals, listeners appeared to identify a talker's sex according to the central spectral tendencies of the sinusoidal constituents. Under acoustic conditions that prevented listeners from determining the sex of a talker, individual identification from sine-wave signals was often successful. These results reveal that the perception of a talker's sex and identity are not contingent and that fine-grain aspects of a talker's phonetic production can elicit individual identification under conditions that block the perception of voice quality.

What can a listener perceive in the speech of an unfamiliar talker? Even a brief utterance can convey a linguistic message and something about the talker who produced it. Although the perception of personal attributes has commonly been explained by an account separate from the perception of linguistic properties, a recent study has shown that phonetic details can also be used to identify talkers and to distinguish them from one another (Remez, Fellowes, & Rubin, 1997). Surprisingly, when acoustic test materials forced performance to depend on phonetic attributes, listeners occasionally mistook male talkers for female talkers, and vice versa. The present report describes a series of experiments intended to clarify the interpretation of this counterintuitive finding, posing these questions: (1) Is the sex of a talker identifiable in a sine wave utterance replica? (2) Are differences across talkers in the central spectral tendency of the sinusoidal constituents responsible for differing impressions of the sex of a sine wave talker? (3) Are individuals identifiable under acoustic conditions that preclude the identification of sex?

Many studies of talker recognition by ear, by automatic classification, or by visual inspection of spectrograms have sought to tie variation across individuals to distributions of coarse grain characteristics of speech signals, such as the fundamental frequency of phonation and its range, or the average long-term spectrum (see Bricker & Pruzansky, 1976; Carrell, 1981, 1985; Fant, 1966; Hecker, 1971; Hollien & Klepper, 1984; Jassem, 1971; Joos, 1948; Monsen & Engebretson, 1977; Nearey, 1978). In part, these acoustic characteristics have been understood as functional effects of anatomical differences among talkers in laryngeal mass and vocal tract size. Although the perception of specific vocal qualities stems from the glottal spectrum and its supralaryngeal modification, these timbral attributes are only a portion of the differences available for distinguishing talkers perceptually, as studies have recently shown. When a listener becomes familiar with a new talker, the remembered characteristics are likely to include aspects of the talker's habits of producing speech sounds in addition to voice quality (Goldinger, 1996; Lieberman, 1963; Nolan, 1983; Nygaard, Somers, & Pisoni, 1994; Pickett & Pollack, 1963; Sheffert & Fowler, 1995). The idiolectal properties that distinguish talkers were also implicated perceptually in a recent study in which listeners identified familiar talkers from sine wave signals that lacked the pitch and timbre of natural speech (Remez et al., 1997).

In this prior study, Remez et al. (1997) aimed to answer a specific question: Can listeners identify talkers without relying on voice quality? To compose the relevant test, a simple problem was set for the listener: to report on each trial which one of a pair of sentences had been produced by the familiar man or woman whose printed name was

This research was supported by Grants DC00308 (to R.E.R.) and HD01994 (to Haskins Laboratories) from the National Institutes of Health.The authors gratefully acknowledge the meticulous assistance and trenchant advice of Chris Darwin, Steve Goldinger, Harry Levitt, Jennifer Lipton, Larry Rosenblum, Jim Sawusch, Dalia Shoretz, Saskia Smith, Steve Stroessner, Doug Whalen, and Fay Xing. Correspondence should be addressed to R. E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027-6598 (e-mail: remez@paradise.barnard.columbia.edu).

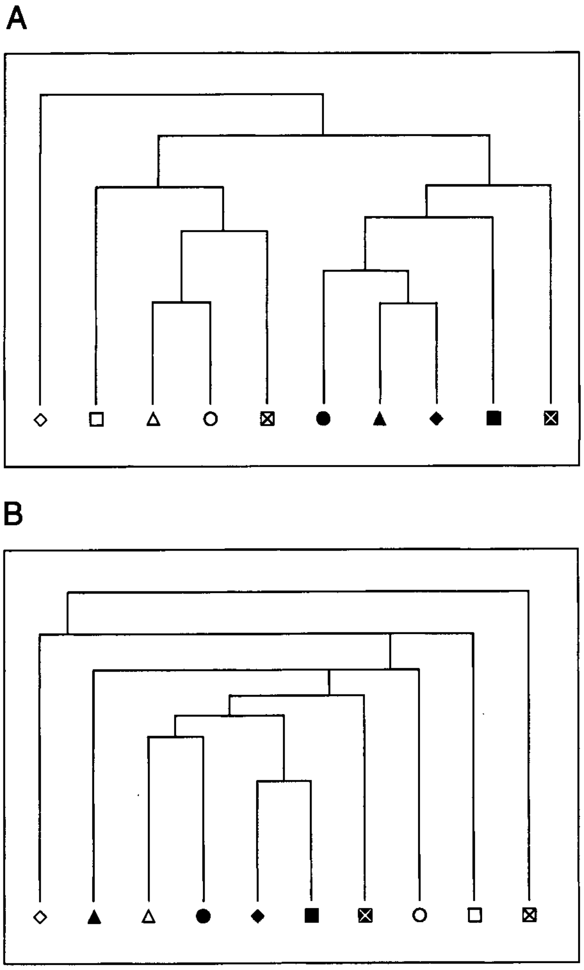

Figure 1. Hierarchical cluster analysis of 10 talkers in Remez et al. (1997). (A) Physical similarities, derived from distances between talker means in the plane of F1 by F2. (B) Perceived similarities, derived from errors of identification by listeners familiar with the talkers in the test set. Male talkers are represented as solid bullets; females are represented as open bullets. Note the segregation of talkers by sex in panel A in contrast to mixed-sex clusters in panel B.

presented at the beginning of the trial. However, the acoustic signals were sine-wave replicas of natural utterances, in which three time-varying sinusoids manifested the frequency and amplitude changes of the three lowest formants of a speech sample from each talker. Such sine-wave sentences exhibit unnatural timbre (Remez & Rubin, in press) and intonation (Remez & Rubin, 1984, 1993) despite their intelligibility. A listener who succeeded in the task therefore perceived the fine phonetic grain of each sine-wave signal in order to recognize the phonetic habits of the individuals in the talker set. This conclusion was bolstered by two converging kinds of evidence.

First, acoustic similarity of the sine-wave signals did not predict perceptual similarity. This is readily discerned in the two panels of Figure 1, which show hierarchical cluster analyses for acoustic and perceptual similarities of the 10 talkers identified by listeners in this test. Each talker's acoustic properties were represented by the average frequency of the first and second formant (cf. Nearey, 1978), from which the acoustic differences of each to each were readily calculated (see Figure 1a). Errors of identification in a perceptual test were used to represent perceptual similarity (Figure 1b). The topology of the clusters differs, of course; however, the more telling difference occurs in the clear separation of talkers by sex in the analysis of acoustic differences and the mingling of the sexes in the analysis of the perceptual errors. The pattern of perceptual errors suggested that the listeners disregarded the auditory form of sine waves—the impressions of several whistles varying simultaneously in pitch and loudness—and instead compared fine-grain phonetic properties of each talker's vowels and consonants conveyed by the tone patterns.

A second kind of evidence encouraged the hypothesis that individual identification was based on phonetic properties in the absence of typical voice quality. In this procedure (Remez et al., 1997), naive listeners were asked on each trial to choose which one of a pair of sine-wave sentences was spoken by a specific talker. Because a naive listener began the test without any familiarity with the 10 talkers in the talker set, the sine-wave talker to be identified was designated by presenting a natural sample of 1 of the 10 at the beginning of each trial. This task allowed analytic listening to the natural sample and implicitly invited exact comparisons of the spectra of natural and sine-wave sentences. In analyzing the performance, the error distribution was treated as an index of perceptual similarity and was submitted to a multidimensional scaling analysis. This again revealed the listener's disregard of auditory qualities of the sine-wave voice in favor of phonetic attributes. Apparently, a male talker was misidentified occasionally as a female talker, and vice versa, when the two talkers had produced phonetically similar realizations of the test sentence. Moreover, the scaling analysis of the performance of naive listeners who matched a natural sample on each trial was highly correlated with the analysis of the performance of listeners who identified sine-wave talkers by long-established familiarity. In neither condition did subjects apparently attend to the sex of the talker.

The outcomes reported by Remez et al. (1997) were explained by supposing that a tonal analog of speech conserves the phonetic grain of the natural utterance from which it derives. This conservative property of a sine-wave sentence can be attributed to its time-varying properties, which are common to a sample of natural speech and its tonal imitation (Remez, Rubin, Pisoni, & Carrell, 1981). The unnatural timbre of sine-wave sentences is accordingly attributable to the tonal carrier, which lacks the broadband resonances, aperiodicities, and harmonic relations of natural speech. Under such circumstances, it is reasonable to expect listeners to form distinct impressions of phonetic properties and only weak impressions of voice quality.

In fact, sine-wave utterance replicas specifically exclude acoustic properties that are thought to underlie the perception of maleness and femaleness of a voice. For instance, neither glottal periodicity nor the spectrum attributed to the shape of the glottal pulse nor perturbations in perfect periodicity (see Carrell, 1985; Klatt & Klatt, 1990; Monsen & Engebretson, 1977) are preserved in a sine-wave replica. Differences between the sexes are manifest in distinct ranges of these three attributes, though none of these acoustic properties, all originating in the glottal excitation, is available in a sine-wave replica of speech, which represents a speech signal merely as the time-varying center frequencies and amplitudes of the formants. Furthermore, the presence of a noise band, of spectral zeroes, and of a greater bandwidth in the lowest resonance in the vocal spectrum, which are held to characterize the distinctive female voice quality relative to male voices (Klatt & Klatt, 1990), are likewise eliminated in sine-wave signals modeled on the formant structure of speech. Although some studies appear to support a claim that the vocal spectrum alone can distinguish males and females perceptually, the use of natural whispered excitation (e.g., Schwartz & Rine, 1968) makes the interpretation uncertain. Because whispered speech often can contain acoustic effects of a periodic source component and surely contains broadband resonances, it is likely that sine-wave replicas of utterances alone lack the very acoustic attributes held to evoke impressions of maleness or femaleness. Overall, it is tempting to conclude that listeners in the study by Remez et al. (1997) confused males and females due to the drastically limited acoustic properties of sine-wave signals.

The present series of four experiments addressed this conjecture with new evidence. In Experiment 1, we found that a sine-wave replica of a natural utterance exhibits sufficient acoustic structure to allow a listener to determine the sex of the talker. In Experiment 2, we determined that frequency transpositions of a sine-wave sentence that altered the central spectral tendency of the variation of each tone can be responsible for impressions of the sex of the talker or for indeterminacy in those impressions. In Experiments 3 and 4, we studied the relationship between individual identification of a sine-wave talker and identification of a talker's sex, and we found that individual identification is not contingent on the identification of sex nor on the acoustic conditions that allow the perceptual determination of the sex of the talker.

EXPERIMENT 1
Identifying the Sex of a Sine-Wave Talker

In Experiment 1, listeners were asked simply to report the apparent sex, male or female, of the talker producing a sine-wave sentence on every trial. Some listeners were tested with sine-wave sentences that replicated the natural formant frequency values—in which case, the 10 sentences preserved the differences in average formant frequency exhibited by the natural utterances of the 10 talkers. To determine whether fine-grain spectral variation conveys segmental or other information sufficient to identify the talker's sex (see Byrd, 1994; McDonough, Ladefoged, & George, 1993; Trudgill, 1974), other listeners were tested with sine-wave sentences that were transposed to eliminate differences between talkers in the central spectral tendency of each tone component. Frequency transpositions preserving pitch differences were imposed in three instances to give every talker's sine-wave replica the same average formant frequencies: (1) the average values of the 5 male talkers were used, (2) the average values of the 5 female talkers were used, and (3) the average values of all 10 talkers were used.

Method

The talkers. Ten sine-wave sentences, each a replica of a natural utterance spoken by a different talker, formed the core of the materials used in the experiments of this report. They were derived from the natural utterances of 5 male and 5 female talkers, each of whom spoke the sentence, "The drowning man let out a yell," amid a list of sentences that were read aloud twice. The natural readings were fluent and neither normative nor vernacular, according to each talker's habit. The talkers were not informed about the purpose of the experiment. The individuals contributing speech samples to this project had been chosen because of their likely familiarity to listeners in one specific test reported by Remez et al. (1997) and without regard to the perceptual differentiability of their speech.

Acoustic test materials. This test employed four sets of 10 sine-wave sentences: (1) *natural* frequency values, (2) *male-transposed* frequency values, (3) *female-transposed* frequency values, and (4) *neutral-transposed* frequency values. The natural frequency sine-wave sentences were created first, according to this procedure: Natural speech was sampled from each subject, who read a list of sentences aloud twice in a sound-attenuating chamber. These utterances were recorded on audiotape with a high-quality voice microphone, filtered (4.5 kHz low-pass, −40 dB/octave rolloff) and sampled (at 10 kHz), using a PCM system implemented on a VAXstation II/GPX. To estimate the center frequency and amplitude of the three lowest frequency formants throughout each utterance, samples were analyzed in two ways: (1) the peak-picking method of linear prediction, and (2) the spectral analysis method of discrete Fourier transforms. By inspecting both representations of the spectrum, a table of sinusoidal synthesis specifications was composed interactively for each utterance consisting of formant frequencies and amplitudes derived at 5-msec intervals. A sine-wave synthesizer (Rubin, 1980) calculated the waveforms with a 10-kHz temporal resolution. These waveforms were stored on the VAX as digital records.

A male, a female, and a neutral set of the 10 sine-wave sentences were created by editing the sine-wave synthesis parameters for each sentence. The sine-wave frequency values for each subject were rescaled such that, in the male set, the tone pattern of all 10 talkers exhibited the same average formant frequencies. In the female set, all 10 tone patterns exhibited the female values. In the neutral set, the tone pattern of every talker exhibited the average values of the entire talker set. The frequency values employed in each transposition condition are shown in Table 1.

Sinusoidal patterns were sequenced and converted from digital records to analog signals, recorded on half-track 0.25-in. audiotape, and presented to the listeners via tape playback. The listeners sat in carrels in a sound-shielded room, and signals were presented binaurally at an approximate level of 65 dB SPL over matched and calibrated headsets.

Procedure. Four separate tests employing different acoustic materials were used in this experiment. Each test comprised 100 trials, in which 10 sine-wave sentences, each produced by a different talker, were presented 10 times in a random order. There were 3 sec

Table 1
Average Formant Values (in Hertz) Used to Transpose Sine-Wave Replicas

Formant	Male	Female	Neutral
First	480.3	548.2	514.2
Second	1555.8	1883.3	1719.6
Third	2466.2	2909.2	2687.7

of silence between trials, with the exception of every 10th trial—after which there were 6 sec of silence.

Listeners. Forty-two students at Barnard College were assigned randomly to one of four test conditions: (1) natural frequency, (2) male-transposed frequency, (3) female-transposed frequency, and (4) neutral-transposed frequency. They were tested in groups of 6 or fewer. All were native speakers of English, and all reported no history of disorder of speech or hearing. None had participated in any other experiment that used sinusoidal signals. The listeners were not familiar with the talkers whose speech samples were used in the acoustic test materials. The listeners were drawn from introductory psychology classes, and they received course credit for their participation. They were asked to decide whether, on each trial, the sex of the sine-wave talker was male or female. Two listeners were eliminated from the data set for failing to follow instructions.

Results and Discussion

Each subject contributed two values (average percent correct identification of talker sex for male and for female talkers) to the data set. An analysis of variance (ANOVA) of the group data was performed to test for the factor of spectrum between subjects, for the factor of sex within subjects, and for their interaction, revealing a significant effect of spectrum [$F(3,36) = 18.1, p < .001$]. This analysis shows that the listeners could determine the sex of a talker only when the tone analogs preserved the natural formant frequency values. The group averages are plotted in Figure 2, in which the height of each bar corresponds to the mean performance of 10 listeners in identifying the sex of the 10 sine-wave talkers, and the error bars represent the critical interval for a post hoc means test according to the method of Tukey. Note that performance did not differ from guessing in the three conditions employing frequency-transposed sine-wave signals.

A significant statistical effect was also observed for the interaction of the factors of spectrum and sex [$F(3,36) = 3.55, p < .025$]. This outcome reflects the differential effects across frequency transpositions on the perceived sex. When each sine-wave sentence was transposed to match the average male values, the listeners were more likely to report that the talker was male; when the test items were transposed to match the average female frequencies, the listeners were more likely to report that the talker was female. All other things being equal, the listeners were more likely to identify the males correctly and the females incorrectly when the sine-wave spectrum was shifted downward in frequency. Conversely, the listeners were more likely to identify the females correctly and the males incorrectly when the spectrum was shifted upward in frequency.

Overall, the outcome of this test provides clear evidence that sine-wave replicas of natural utterances can evoke veridical impressions of the sex of a talker when the natural frequency values are preserved in the time-varying tones. The listeners were accurate in reporting the sex of the talkers when tone complexes replicated the exact frequency variation of each talker's natural speech spectrum. This finding supplements perceptual accounts of the qualitative differences between male and female speech, which have emphasized precise glottal periodicity and spectrum. In addition to those demonstrations of the perceptual effectiveness of the acoustic correlates of vocal pitch and timbre (e.g., Klatt & Klatt, 1990), our results show that listeners can be accurate in their assessments of a talker's sex without the availability of natural vocal acoustic products.

It is also useful to note that the segmental phonetic constituents in the speech samples that we employed were not sufficient to permit the listeners to gauge the sex of the talkers in the frequency-transposed test conditions. In the three conditions that eliminated intertalker differences in central spectral tendency, identification performance did not differ from chance; however, in the male-transposed conditions, the listeners reported hearing males more often, and, in the female-transposed conditions, the listeners reported hearing females more often. Although some research has characterized differential phonetic habits of male and female talkers (see Byrd, 1994), the listeners in our tests heard samples that were evidently too spare or too unusual to evoke reliable impressions of the sex of talkers.

If differences between talkers in the spectral band within which tone analogs vary are effective in eliciting impressions of different sexes, a stricter test is required, nonetheless, to determine whether changes in central

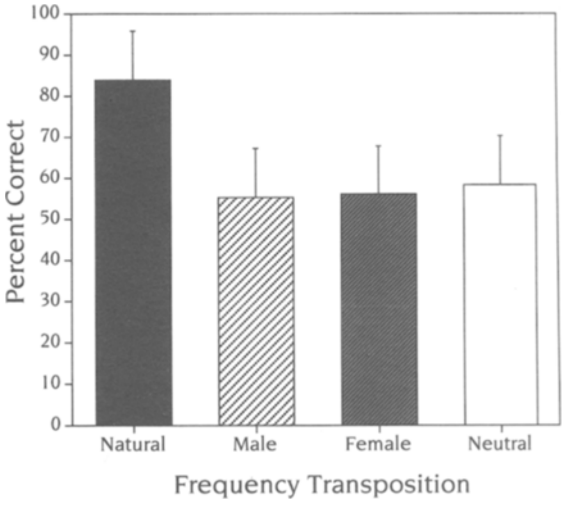

Figure 2. Results of Experiment 1, in which the listeners were asked to report the apparent sex of 10 talkers in four conditions of acoustic variation. Error bars represent the critical interval for a Tukey post hoc means test ($\alpha = .05$).

spectral tendency of each of the tone components alone can evoke different impressions of the sex of a talker. This conclusion follows from the fact that, in Experiment 1, we blocked conditions by the spectrum factor, and the listeners in the natural condition were the only ones to hear intertalker variation in the spectrum during the test. Accordingly, in Experiment 2, we used a single talker—a highly identifiable female—in three frequency transpositions: male, female, and neutral. Because the natural values were not employed in this test, and because the same formant pattern was used in every test item, the test provided a clear measure of the effect on impressions of sex of central spectral tendency while controlling phonetic variation.

EXPERIMENT 2
The Effects of Central Spectral Tendency of Tone Analogs of Formants

In Experiment 2, the sine-wave synthesis parameters of a single talker were transposed to exhibit male, female, or neutral average spectra. By composing the test materials solely of frequency transpositions of a single sinusoidal pattern, we attempted to provide a more controlled assessment of the effect of central spectral tendency of a tonal analog of speech in evoking an impression of a talker's sex.

Method

Acoustic test materials. This test employed three sine-wave sentences, each deriving from the highly identifiable sine-wave replica of a female talker employed in the test set of Experiment 1. Three frequency-transposed versions—a male, a female, and a neutral variant—were prepared, with average values noted in Table 1.

The test items were prepared by digital synthesis and were presented over matched and calibrated headsets by audiotape playback to the listeners seated in a sound-shielded room. Signals were presented binaurally at an approximate level of 65 dB SPL.

Procedure. A single test of 30 trials was used in which three sine-wave sentences, each with a different central spectral tendency achieved by frequency transposition of the tonal components, were presented 10 times in a random order. There were 3 sec of silence between trials, with the exception of every 10th trial—after which there were 6 sec of silence. The listeners were asked to decide on each trial whether the sex of the sine-wave talker was male or female.

Listeners. Eight students at Barnard College, all native speakers of English, participated in this test. None reported a history of disorder of speech or hearing, nor had any participated in an experiment that used sinusoidal signals. No listener was familiar with the talker whose speech had been modeled to produce the test items.

Results and Discussion

Each subject contributed three values to the data set, the percent of all trials on which the male, female, and neutral sentences were identified as "male". A one-way repeated measures ANOVA of the group data was performed to test for an effect of the factor of frequency transposition, revealing a significant difference among the three conditions [$F(2,14) = 128.0, p < .001$]. The group averages are plotted in Figure 3, in which the height of each bar corresponds to the mean percent of trials on which the listeners reported that the talker was male or female for the three spectral variants of the sine-wave pattern; the error bars represent the critical interval for a Tukey post hoc means test. The results of this test show clearly that a sine-wave sentence exhibiting the average formant values of the male talkers in Experiment 1 was so identified in Experiment 2. Similarly, transposing the pattern to exhibit the average female tone frequency values led the listeners to identify the talker as female. When each tonal constituent manifested the central spectral tendency averaged over the entire talker set—incorporating 5 males and 5 females—no consistent impression of the sex of the talker was elicited.

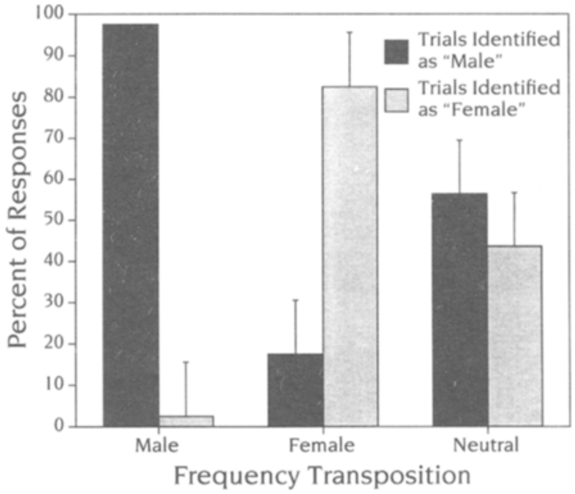

Figure 3. Results of Experiment 2, in which the listeners were asked to report the sex of a sine-wave pattern presented in three frequency-transposed variants.

The outcome of this test indicates that changes in the central spectral tendency of the tonal analogs of formants in a sinusoidal sentence reliably influence listeners' reports of maleness or femaleness. When the spectrum was characteristic of neither males nor females, the listeners' reports also coincided. Overall, this pattern of results indicates that variation in this acoustic property can readily be analyzed as the personal attribute of talker sex even when concurrent variation in phonetic properties is controlled. The results of Experiment 2 corroborate the conjecture, based on the outcome of Experiment 1, that the perception of the sex of a talker can be evoked in the absence of a signal component attributable to the differential glottal pulse rate of males and females. Also, acoustic correlates of differential glottal spectra are not specifically required for listeners to report a definite impression of the sex of the talker of the sentence.

The question prompting these two experiments originated in a puzzling finding by Remez et al. (1997)—namely, that listeners in a test of talker identification using sine-wave signals appeared to confuse males and females for each other, as if the listeners were often indifferent to the sex of a sine-wave talker. As a consequence of Experiments 1 and 2, we can conclude that, in fact, listeners are quite able to identify the sex of a sine-wave talker when asked to do so, even under test conditions that are thoroughly constraining. These findings are difficult to reconcile with the confusion of males and females that was observed in a test of talker identification. Listeners would greatly simplify the task of identifying a sine-wave talker from a set of known size by taking a preliminary step of identifying the sex of the talker. Furthermore, at first approximation, it is reasonable to suppose that the listeners would exploit every shred of information available from an apparently impoverished signal. Why then did listeners fail to simplify the task that was set for them by Remez et al. (1997) by identifying the sex of the sine-wave talkers? In Experiments 3 and 4, we pursued this question by attempting to calibrate the contribution of sex identification to individual identification.

EXPERIMENT 3
Individual Recognition Without Sex Identification

The problem motivating Experiment 3 was the apparent indifference of listeners to acoustic information about the sex of the talker, which would otherwise seem to promote the task of individual identification. An appeal to common sense suggests that the radically simplified signal presented by a sine-wave replica would challenge the perceiver to use all the information that can be extracted from the tone pattern (see Reynolds, 1961; Wilkie & Masson, 1976). The outcome reported by Remez et al. (1997) implies, on the contrary, that the listeners may not have used a source of information that was readily available, as Experiments 1 and 2 of this report demonstrate. A likely reason for the listeners' indifference to information about sex in an individual recognition test was evaluated in Experiment 3. Here, we tested whether individuals can be recognized adequately on the basis of phonetic attributes alone, relieving the necessity of resorting to supplementary information about the talker's sex. The empirical setting of Experiment 3 was a replication of the test of individual identification of Remez et al. (1997), although the sine-wave replicas of the 10 talkers were transposed in Experiment 3 to the neutral frequency values in order to block veridical impressions of talker sex.

Method

Acoustic test materials. Natural and sine-wave sentences were used in this test—natural items as samples to be matched on each trial, and sine-wave items as candidates to be matched to the natural versions. This set of sine-wave sentences had been created for the neutral frequency transposition condition of Experiment 1, in which the listeners were unable to identify the sex of the talkers (see Figure 2). Recall that each sine-wave sentence was derived from the natural speech of a different talker and was transposed in frequency to exhibit the sex-neutral average formant values given in Table 1. Because a sine-wave sentence preserved the abstract pattern of formant frequency variation of a specific talker and, therefore, preserved the acoustic correlates of fine-grain phonetic attributes, this set of acoustic materials was used here to test individual identification from phonetic attributes when the sex of a talker was perceptually indeterminate.

A set of 10 samples of natural speech was used in this test to represent the 10 talkers to be identified. On each trial, the listener was asked to choose the sine-wave sentence spoken by the natural talker. Here, the procedure used the specific natural utterances on which the sine-wave replicas were based. A prior test had shown that listeners succeed in matching sine-wave and natural sentences well (for 8 of the 10 talkers) when the sine-wave items exhibit natural (untransposed) formant frequency values (Remez et al., 1997).

The natural sentences were stored as digital records, sequenced with the synthetic sentences and presented on line by a VAXstation 4000 Model 90 and a Gradient DeskLab audio output system over matched and calibrated headsets to the listeners seated in a sound-shielded room. Signals were presented at an approximate level of 65 dB SPL.

Procedure. Each trial of this test was composed of three events: a natural sentence followed by two sine-wave sentences. One of the pair of sine-wave patterns always had been derived from the natural utterance presented on that trial, whereas the other sine-wave pattern had been derived from a natural utterance produced by 1 of the other 9 talkers. A listener was asked on each trial to report which of the two sine-wave sentences was produced by the talker who spoke the natural utterance.

With 10 different talkers, there were nine comparisons of each sine-wave sentence with each other, making a minimum of 90 trials; counterbalancing for order of presentation of the alternatives resulted in a test of 180 trials. The natural sentence and the first sine-wave sentence were separated by 750 msec of silence on each trial, and the first and second sine-wave sentences were also separated by 750 msec of silence. Between each trial, there were 3 sec of silence, with 6 sec of silence after every 10th trial.

Listeners. Eleven students at Barnard College, all native speakers of English, participated in this test. None had participated in an experiment that used sinusoidal signals, nor did any report a history of disorder of speech or hearing. The listeners were not familiar with the individuals whose speech samples and sine-wave analogs composed the acoustic test materials. In exchange for their participation, the listeners received credit in introductory psychology.

Results and Discussion

A subject contributed 10 values to the data set, the percent of possible trials on which each of the 10 natural samples had been matched to its sine-wave replica. A one-way repeated measures ANOVA of the group data was performed to test for an effect of the factor of talker, revealing a significant difference [$F(9,90) = 11.4, p < .001$]. This reflected the different hit rates observed across the set of 10 talkers. The group averages are plotted in Figure 4, in which the height of each bar corresponds to the mean percent of trials on which the listeners chose the correct alternative to match the natural talker. The gray horizontal region flanking 50% shows the confidence interval for significant difference from chance as established by a post hoc means test (Tukey). The results of this test show clearly that the listeners recognized 8 of the 10 talkers under acoustic conditions that precluded identification of the sex of the talker.

The same talkers who were difficult for the listeners to identify in this condition were also difficult for the listeners to identify when the tonal analogs reproduced the natural formant frequencies (in Remez et al., 1997, Experiment 1). Indeed, this pattern of results shows that lis-

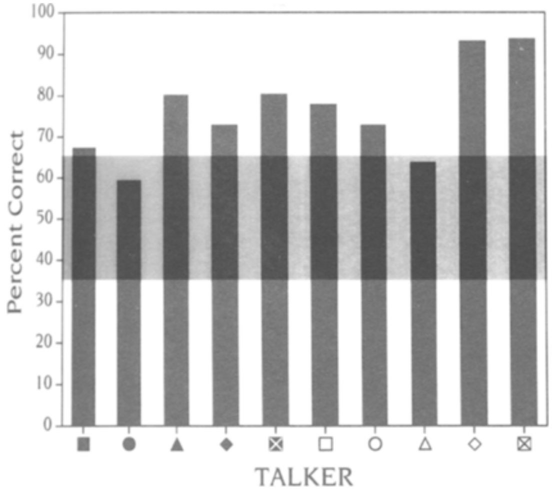

Figure 4. Results of Experiment 3, in which the listeners were asked to match the talker of a natural utterance to candidate sine-wave sentences. The height of each bar corresponds to the group mean percent correct; the gray region around 50% indicates the confidence interval for statistical difference from guessing, as established by a post hoc means test (Tukey, $\alpha = .05$). Eight of 10 talkers were identified by the listeners under acoustic conditions precluding identification of the talker's sex.

teners in a test of individual identification from tonal analogs of speech would be able to disregard the acoustic information about a talker's sex without impairing performance. How then do listeners tell which sine-wave sentence was produced by the talker who spoke the natural sample? The present finding is consistent with the claim that the phonetic attributes evoked by the coherent modulation of sinusoidal carriers provide a basis for perceiving individuals (Remez et al., 1997) and that acoustic properties indicating the sex of a talker must contribute very little information to this perceptual effect.

Despite the appeal of this argument, it is conceivable that performance in this test did not reflect a phonetic comparison of natural and sine-wave sentences at all but was a consequence of a far more superficial perceptual analysis. Specifically, our listeners may have ignored phonetic attributes and simply attempted to analyze the duration of the sentences, thereby treating the procedure as a test of duration matching. Because one of the pair of sine-wave sentences was derived from the natural model presented on each trial, the listeners who evaluated the acoustic materials in this alternative fashion would presumably have achieved good performance without ever attending to phonetic attributes. Moreover, a comparison of duration is the most likely superficial way to solve this matching task, since the auditory form of a sine-wave sentence—the changing pitch and loudness impressions evoked by each of the tonal analogs of the formants—does not readily correspond to impressions of natural speech and is unique to sine-wave vehicles. In contrast, the frequency and amplitude variation of the three or four perceptually important formants in a speech signal cannot be separately resolved (see Julesz & Hirsh, 1972).

To evaluate a duration-based alternative to the phonetic comparison hypothesis, we sought to predict the outcome of the matching test if the listeners had compared the relative duration of natural and sine-wave items. To accomplish this, we reasoned that listeners' ability to choose between sine-wave alternatives in matching the natural sample presented on each trial ought to vary in consistency as a function of the duration difference of the sine-wave candidates. Where the difference in duration of the pair of sine-wave alternatives was large, the listeners should have readily distinguished between them and should have chosen the better match to the natural sentence. Where the duration difference was small, the listeners should not have chosen a match for the natural sentence as well. In the limiting case, if two sine-wave sentences had the identical duration, the listeners should not have been able to select between them consistently to match the natural sample. The sample of 10 sine-wave items had an average duration of 2,055 msec, with a standard deviation of 361 msec and high and low values of 2,904 and 1,458 msec, respectively.

This alternative hypothesis was tested statistically by deriving a hierarchical cluster analysis for the 10 sine-wave sentences based solely on the duration of each item, thereby ranking the 10 items by similarity in duration. This predicted similarity ranking was then compared with a similarity ranking based on the actual responses of the listeners—specifically, on the pattern of errors. A test for correlation of the predicted and observed similarity rankings was definitive, revealing that the prediction based on duration contrast did not match the observations (Spearman's $r_s = .07$, $p > .05$; corrected for tied ranks by the method of Siegel & Castellan, 1988). In short, the performance of the subjects in Experiment 3 does not offer evidence that the listening task is solved by means of a perceptual analysis of duration contrast.

One additional control is warranted to justify this conclusion. In Experiment 3, the listeners were asked to compare sine-wave sentences with natural samples from which they were derived. The fine spectrotemporal structure of the sentences was therefore identical in cognate natural–sine-wave pairs. Although the preponderance of evidence shows that listeners are unable to attend to the auditory form of the acoustic constituents of natural speech nor to identify the precise auditory form of tone patterns corresponding to the phonetic properties of sine-wave replicas (Remez, Rubin, Berns, Pardo, & Lang, 1994), the acoustic materials used in this test allow a slim possibility that the listeners made exact matches on the basis of auditory similarity of portions of natural and sine-wave patterns independent of phonetic variety. In Experiment 4, we imposed a control for this prospect by using different natural samples spoken by the same talkers as those who spoke the 10 sine-wave sentences. In this test, listeners could not detect a correspondence of natural and sine-wave sentences by listening for an exact match.

EXPERIMENT 4
A More Stringent Test of Individual Recognition Without Sex Identification

The findings of Experiment 3 have shown that individual identification can occur even when acoustic conditions do not foster veridical impressions of a talker's sex. While this result encouraged the conclusion that phonetic properties conveyed by sine-wave sentences are sufficient to evoke an impression of a particular talker, there was one alternative interpretation to consider immediately. In Experiment 4, we employed test items designed to prevent the listeners from identifying the sine-wave talkers by listening for an exact spectrotemporal match between natural and sine-wave items. This was achieved by using an alternate utterance of the test sentence as the natural sample of the talker to be matched to sinusoidal candidates that exhibited the sex-neutral central spectral tendency.

Method

Acoustic test materials. As in Experiment 3, both natural and sine-wave sentences were used in this test: natural sentences representing talkers to be matched on each trial, and sine-wave sentences as candidates to be matched to natural versions. Ten samples of natural speech were used in this test to represent the 10 talkers to be identified. These utterances were alternate readings of the sentence, "The drowning man let out a yell," collected at the same session as were the utterances that had served as models for sine-wave replication. The average difference in duration between the natural models for the sine-wave items and the natural samples used in Experiment 4 was approximately 200 msec. The 10 different sine-wave sentences used in this test were the neutral frequency transposition items used in Experiment 3. Accordingly, listeners in this test were not able to exploit a strategy of listening analytically for exact spectrotemporal matches between the auditory form of natural and sine-wave items.

The natural sentences were sequenced with the synthetic sentences, and all were converted with temporal resolution of 10 kHz to analog voltages. The test was presented binaurally at an approximate level of 65 dB SPL over matched and calibrated headsets by a VAXstation 4000 Model 90 and a Gradient DeskLab audio output system.

Procedure. Three acoustic events occurred on each trial of this test: a natural sentence followed by two sine-wave sentences. One of the pair of sine-wave patterns always had been derived from a sample spoken by the talker whose natural utterance was presented on that trial, whereas the other sine-wave pattern had been derived from a sample produced by one of the other 9 talkers. As in Experiment 3, listeners were asked on each trial to report which of the two sine-wave sentences was produced by the talker who spoke the natural utterance.

With 10 different talkers, there were nine comparisons of each sine-wave sentence with each other, making a minimum of 90 trials; counterbalancing for order of presentation of the alternatives resulted in a test of 180 trials. The natural sentence and the first sine-wave sentence were separated by 750 msec of silence on each trial, and the first and second sine-wave sentences were also separated by 750 msec of silence. Between each trial, there were 3 sec of silence, with 6 sec of silence after every 10th trial.

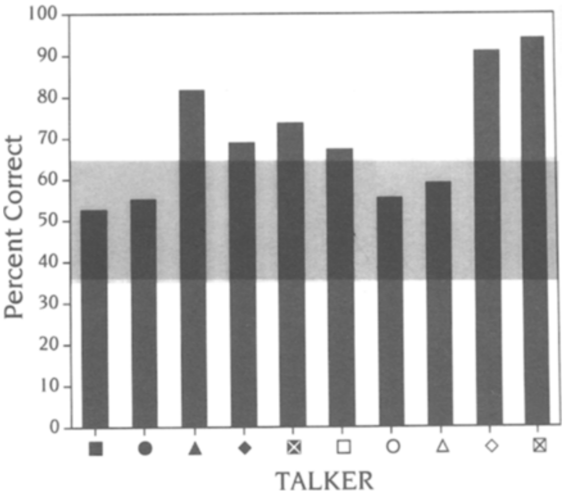

Figure 5. Results of Experiment 4, in which the listeners were asked to match the talker of a natural utterance to candidate sine-wave sentences. The height of each bar corresponds to the group mean percent correct; the gray region around 50% indicates the confidence interval for statistical difference from guessing, as established by a post hoc means test (Tukey, $\alpha = .05$). Six of 10 talkers were identified by the listeners under acoustic conditions precluding an exact spectrotemporal match between natural and sine-wave items, as well as preventing identification of the talker's sex.

Listeners. Eighteen students at Barnard College participated in this test in exchange for credit in introductory psychology classes. All were native speakers of English. None had participated in an experiment that used sinusoidal signals, nor did any report a history of disorder of speech or hearing. The listeners were not acquainted with the talkers whose speech samples were used in this test.

Results and Discussion

Each subject contributed 10 values to the data set, the percent of possible trials on which the sine-wave mate to each of the 10 natural samples had been identified correctly. A one-way repeated measures ANOVA of the group data was performed to test for an effect of the factor of talker, revealing a significant difference [$F(9,153) = 21.4, p < .001$]. Again, this finding reflected the different hit rates observed across the set of 10 talkers. The group means are plotted in Figure 5, in which the height of each bar corresponds to the percent of trials on which the listeners chose the correct alternative to match the natural talker. A gray horizontal region flanking 50% is used to mark the confidence interval for significant difference from chance as established by a post hoc means test (Tukey). The results show clearly that the listeners recognized 6 of the 10 talkers under acoustic conditions that (1) precluded an exact spectro-temporal match and (2) precluded the identification of the sex of the talker.

The outcome of Experiment 4 again indicated that listeners can employ remarkably abstract perceptual criteria for identifying talkers. For 6 of the 10 talkers in the target set, the data reveal the sufficiency of perceptual criteria based on phonetic attributes, as the results of Experiment 3 and our prior work had also indicated (Remez et al., 1997). The listeners here could not have performed the task by listening for exact matches, neither in central spectral tendency of tone variation nor in the precise spectrotemporal form of natural and sine-wave items. Moreover, no simplification in the choice on each trial could be gained by identifying the sex of the sine-wave talkers, because the results of Experiments 1 and 2 showed that the frequency transpositions employed here abolish veridical impressions of a talker's sex.

By employing natural samples in Experiment 4 that differed from the models for the sine-wave items, we had aimed to render useless a strategy of exact comparison of natural and sine-wave sentences. Nonetheless, the listeners may have persisted in treating this test of talker identification as a test of duration matching, inasmuch as this superficial attribute of the test items was readily available, and variation in the duration of sine-wave and natural sentences was well within the listeners' resolving power. The test of this alternative hypothesis in Experiment 3 showed that the listeners had not matched duration in circumstances that permitted exact matches. We had found that the perceptual similarity of the 10 talkers estimated from erroneous identification performance was not correlated with an index of similarity in duration. In Experiment 4, we also found evidence against an explanation of the pattern of results by appealing to the listeners' hypothetical facility in matching duration.

Adopting the method of Experiment 3, we predicted the outcome of the test according to similarity in duration of the sine-wave items, using a hierarchical cluster analysis. We then ranked the predicted similarity of items and calculated the correlation of this ranking with the similarity ranking calculated from the actual identification performance. Again, we found that the observations were not correlated with the prediction based on duration contrast [Spearman's $r_s = -.23, p > .05$]. This finding corroborates the claim that the listeners in Experiment 4 identified individuals by detecting similarities in fine-grain phonetic realization of the items, such as the precise advancement, height and diphthongization of vowels, and the reduction of consonants. But, how good is a performance in which 6 of the 10 talkers were identified?

This outcome was expected, on the basis of the outcome in the precedent study (Remez et al., 1997). However, that prior test had not eliminated intertalker differences in central spectral tendency of the individual tone components of the sine-wave replicas—in which case, the listeners had matched 8 of the 10 natural samples to the sine-wave mates. This finding, in which sine-wave replicas retained the natural frequency values, serves as a benchmark for evaluating performance here. Indeed, the procedure of Experiment 4 shows that something is lost in the transposition to sex-neutral values and in the elimination of naturally occurring intertalker differences in central spectral tendency. Nevertheless, the decrement in performance was not large, and 6 of the 8 talkers who were identifiable when acoustic variation was more ample were still identified here. To assess the results in absolute terms will require more extensive and varied test conditions; especially useful in this regard would be a training study in which both talkers and listeners are selected for their phonetic attributes, inasmuch as these covary and have been known to influence phonetic and lexical perception (cf. Peterson & Barney, 1952). For now, it is defensible to conclude that the recognition of individuals in this small talker set often was not contingent on acoustic properties that otherwise allow identification of a talker's sex. It is likely that this effect was due to the relatively greater relevance and salience of phonetic attributes in recognizing individuals rather than to subtle aspects of the anomalous voice quality of tone analogs of speech.

GENERAL DISCUSSION

Overall, this series of experiments has elaborated the perceptual criteria that listeners use for identifying individual talkers from sine-wave replicas of their utterances. At first, it seemed counterintuitive for the listeners to have identified talkers without exploiting a simplification that the task allowed: categorization of the sex of a talker. These subjects apparently disregarded acoustic attributes that would have facilitated performance. By discovering the acoustic conditions in which listeners can and cannot determine the sex of a sine-wave talker, we now understand that intertalker differences in central spectral tendency between males and females can evoke veridical impressions of a talker's sex (Experiment 1) and that, under artificial conditions, a tone pattern transposed to male, female, and neutral values elicits the corresponding identification of the talker's sex (Experiment 2).

Although the results of Experiments 1 and 2 showed that listeners in the study by Remez et al. (1997) had disregarded the acoustic attributes promoting the perception of a talker's sex when identifying individuals in the test set, the results of Experiments 3 and 4 revealed that there was little reason for listeners to pay particular attention to this acoustic difference among talkers. The basis for identifying talkers was the phonetic form of the utterances, which is apparently perceived without an explicit match to more superficial details of the acoustic vehicle. In other words, when a sine-wave signal is registered, individual identification and sex identification are not contingent. This is surely counterintuitive, weighed against the experience of everyday conversations, in which the pitch and quality of a voice seem to be the personal attributes by which we recognize each other. However, it is unlikely that the listeners here would have succeeded so readily in identifying talkers without a natural sensitivity to events at the fine phonetic grain. Evidently, perceivers can differentiate talkers as well as words from the phonetic properties of speech. Perhaps the ordinary qualitative impressions of talker differences that drive intuition are based, in actuality, on such phonetic properties.

Furthermore, it is surprising to note that the listeners in our test were not opportunistic, using every shred of acoustic structure that can provide information. We would have expected otherwise, on the assumption that a radically reduced signal presses a perceiver to make maximal use of the remaining acoustic structure. In stark contrast to this characterization of perception, it appears as though our listeners gravitated toward reliance on the sinusoidal correlates of phonetic properties when the auditory form of vocal pitch and timbre was unavailable, due to the sine-wave transform. In fact, the listeners succeeded by relying on phonetic variety across the talkers, ignoring other information that was nevertheless available; when listening for a talker's identity, the listeners would have gained very little by attending specifically to determine the talker's sex. This indifference has no term in perceptual accounts, and we note it here for its role in the informational dynamics of the perception of speech.

While the experiments of this report have accounted specifically for a curious aspect of the identification of talkers from tone analogs of speech, the results also contribute to characterizing the perception of talkers more generally. Chiefly, these experiments warrant broadening the conventional description. In the preponderance of cases, the personal information in an utterance has been described as an extralinguistic component of the message (Bricker & Pruzansky, 1976). This description fits, because a linguistic message is rarely specific to the talker who uttered it. However, our experiments focused on two kinds of personal information that are found in the relationship of acoustics to phonetics. One is anatomical, and it pertains to the acoustic consequences of differences in scale of the laryngeal and supralaryngeal articulators. Anatomical differences in scale among talkers are responsible for differences in the spectra of their speech signals and, accordingly, in the perceived quality of their voices. Though a precise account has proven to be elusive (see Kreiman, 1997), the relationship of anatomical scale to signal properties to perceived qualities has defined the problem of intertalker differences (Pisoni, 1997). Nonetheless, our experiments on talker identification from sine-wave analogs of speech show that personal information is not exhausted by considering the pitch and compass of the speaking voice, nor the shape of the spectrum envelope, nor the central tendencies of formant frequency. From the perspective of the perceiver, the differences that remain in a signal after eliminating the acoustic effects of the laryngeal source and the scale differences of the supralaryngeal resonators are sufficient to allow the identification of individual talkers. Personal information is available in an aspect of the signal that does not arise through anatomical variation alone.

The results of our experiments suggest that the segmental phonetic form of an utterance that is rudimentary to lexical identification also furnishes personal information about the talker. This research therefore coincides with a defining assumption of the literature on talker identification—that personal information in an utterance is an extra message—while it reveals the perceptual effect of subphonemic variation in the identification of individuals. In this regard, the results of these experiments corroborate a recent provocative finding that the establishment of familiarity with a talker includes a substantial phonetic component (Nygaard et al., 1994). Last, our findings can inform and constrain speculation about functional and neural architecture for encoding (see Church & Schacter, 1994) and remembering (Van Lancker, Cummings, Kreiman, & Dobkin, 1988) the speech of individuals.

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.

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, **15**, 39-54.

Carrell, T. D. (1981). Effects of glottal waveform on the perception of talker sex. *Journal of the Acoustical Society of America*, **70**, S97.

Carrell, T. D. (1985). *Contributions of fundamental frequency, formant spacing & glottal waveform to talker identification*. Unpublished doctoral dissertation, Indiana University.

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, & Cognition*, **20**, 521-533.

Fant, G. (1966). A note on vocal tract size factors and nonuniform F-pattern scalings. *Speech Transmission Laboratory: Quarterly Progress & Status Report*, **4**, 22-30.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1166-1183.

Hecker, M. H. L. (1971). *Speaker recognition: An interpretive survey of the literature* (ASHA Monographs, No. 16). Washington, DC: American Speech and Hearing Association.

Hollien, H., & Klepper, B. (1984). The speaker identification problem. *Advances in Forensic Psychology & Psychiatry*, **1**, 87-111.

Jassem, W. (1971). Pitch and compass of the speaking voice. *Journal of the International Phonetic Association*, **1**, 59-68.

Joos, M. (1948). Acoustic phonetics. *Language*, **24**(Suppl.), 1-137.

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

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.

Kreiman, J. (1997). Listening to voice: Theory and practice in voice perception research. In K. Johnson & J. Mullenix (Eds.), *Talker variability in speech recognition* (pp. 85-108). San Diego: Academic Press.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language & Speech*, **6**, 172-187.

McDonough, J., Ladefoged, P., & George, H. (1993). Navajo vowels and phonetic universal tendencies. *UCLA Working Papers in Phonetics*, **84**, 143-150.

Monsen, R. B., & Engebretson, A. M. (1977). Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, **62**, 981-993.

Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Bloomington: Indiana University Linguistics Club.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, **5**, 42-46.

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

Pickett, J. M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language & Speech*, **6**, 151-164.

PISONI, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. Mullenix (Eds.), *Talker variability in speech recognition* (pp. 9-32). San Diego: Academic Press.

REMEZ, R. E., FELLOWES, J. M., & RUBIN, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 651-666.

REMEZ, R. E., & RUBIN, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception & Psychophysics*, **35**, 429-440.

REMEZ, R. E., & RUBIN, P. E. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *Journal of the Acoustical Society of America*, **94**, 1983-1988.

REMEZ, R. E., & RUBIN, P. E. (in press). Acoustic shards, perceptual glue. In J. Charles-Luce, P. A. Luce, & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production, & development*. Norwood, NJ: Ablex.

REMEZ, R. E., RUBIN, P. E., BERNS, S. M., PARDO, J. S., & LANG, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**, 129-156.

REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

REYNOLDS, G. S. (1961). Attention in the pigeon. *Journal of the Experimental Analysis of Behavior*, **4**, 203-208.

RUBIN, P. E. (1980). *Sinewave synthesis* [Internal memorandum]. New Haven, CT: Haskins Laboratories.

SCHWARTZ, M. F., & RINE, H. E. (1968). Identification of speaker sex from isolated whispered vowels. *Journal of the Acoustical Society of America*, **44**, 1736-1737.

SHEFFERT, S. M., & FOWLER, C. A. (1995). The effect of voice and visible speaker change on memory for spoken words. *Journal of Memory & Language*, **34**, 665-685.

SIEGEL, S., & CASTELLAN, N. J., JR. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

TRUDGILL, P. (1974). *Sociolinguistics: An introduction*. Harmondsworth, U.K.: Penguin.

VAN LANCKER, D. R., CUMMINGS, J. L., KREIMAN, J., & DOBKIN, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex*, **24**, 195-209.

WILKIE, D. M., & MASSON, M. E. (1976). Attention in the pigeon: A reevaluation. *Journal of the Experimental Analysis of Behavior*, **26**, 207-212.

(Manuscript received April 19, 1996;
revision accepted for publication October 21, 1996.)