Published in final edited form as:
J Exp Psychol Hum Percept Perform. 2002 December ; 28(6): 1447–1469.

Learning to Recognize Talkers From Natural, Sinewave, and Reversed Speech Samples

Sonya M. Sheffert,
Psychology Department, Central Michigan University

David B. Pisoni,
Department of Psychology, Indiana University Bloomington

Jennifer M. Fellowes, and
Department of Psychiatry, Columbia University

Robert E. Remez
Department of Psychology, Barnard College

Abstract

In 5 experiments, the authors investigated how listeners learn to recognize unfamiliar talkers and how experience with specific utterances generalizes to novel instances. Listeners were trained over several days to identify 10 talkers from natural, sinewave, or reversed speech sentences. The sinewave signals preserved phonetic and some suprasegmental properties while eliminating natural vocal quality. In contrast, the reversed speech signals preserved vocal quality while distorting temporally based phonetic properties. The training results indicate that listeners learned to identify talkers even from acoustic signals lacking natural vocal quality. Generalization performance varied across the different signals and depended on the salience of phonetic information. The results suggest similarities in the phonetic attributes underlying talker recognition and phonetic perception.

When a talker produces an utterance, the listener simultaneously apprehends the linguistic form of the message as well as the nonlinguistic attributes of the talker's unique vocal anatomy and pronunciation habits. Anatomical and stylistic differences in articulation convey an array of personal or indexical qualities, such as personal identity, sex, approximate age, ethnicity, personality, intentions or emotional state, level of alcohol intoxication, and facial expression (see Bricker & Pruzansky, 1976; Chin & Pisoni, 1997; Cook & Wilding, 1997; Doddington, 1985; Kreiman, 1997; Scherer, 1986; Tartter, 1980; Walton & Orlikoff, 1994).

Personal characteristics play an important role in communicative interactions. This is especially true for listeners who are unable to use indexical attributes available in other modalities as a result of neurological impairments in face recognition (*prosopagnosia*: Benton & Van Allen, 1968; Bodamer, 1947; Damasio, Damasio, & Van Hoesen; 1982) or visual impairments (Bull, Rathborn, & Clifford, 1983; Yarmey, 1986). Over the course of a

Correspondence concerning this article should be addressed to Sonya M. Sheffert, Psychology Department, Central Michigan University, Sloan Hall 214, Mount Pleasant, Michigan 48859. sonya.sheffert@cmich.edu.

A portion of these findings was presented at the 134th Annual Meeting of the Acoustical Society of America, December 1997, San Diego, California.

Editor's Note. Carol Fowler served as the action editor for this article.—DAR

lifetime, listeners acquire very detailed and enduring knowledge about many different talkers. The ability to recognize a talker begins in utero (Hepper, Scott, & Shahidullah, 1993) and develops rapidly throughout infancy and childhood (DeCasper & Fifer, 1980; Jusczyk, Hohne, Jusczyk, & Redanz, 1993; Mandel, Jusczyk, & Pisoni, 1995), reaching adult levels of proficiency by age 10 (Mann, Diamond, & Carey, 1979).

An extensive literature on human talker recognition dates to the work of McGehee (1937), who examined the reliability of ear-witness testimony, and the studies by Peters (1955) and Pollack, Pickett, and Sumby (1954), who examined laboratory effects of linguistic content on talker recognition. This literature describes the effects of acoustic, procedural, and individual attributes that affect the recognition and discrimination of unfamiliar talkers (see Bricker & Pruzansky, 1976; Clifford, 1980; Hecker, 1971; Kreiman, 1997; and Read & Craik, 1995, for reviews).

In contrast, much less is known about how a listener recognizes a familiar talker beyond the benchmarks that reveal perceptual, cognitive, and neural differences in the classification of familiar and unfamiliar talkers (Papçun, Kreiman, & Davis, 1989; Schmidt-Nielsen & Stern, 1985; Schweinberger, Herholz, & Sommer, 1997; Van Lancker & Canter, 1982; Van Lancker & Kreiman, 1987; Van Lancker, Kreiman, & Cummings, 1989). Moreover, few studies have examined how a listener becomes familiar with a talker (Legge, Grosmann, & Pieper, 1984; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). These studies show that repeated or extended exposure to a talker's speech increases a listener's sensitivity to talker-specific attributes, improving the ability to differentiate familiar from unfamiliar talkers. Left unspecified, however, are the properties of the speech signal that are most relevant for learning and recognizing familiar talkers from novel utterances.

The research described in this article investigated the recognition of familiar talkers, examining the contribution of different talker-specific properties of a speech signal to perceptual learning. To set the task in this experimental design, we trained our listeners to identify different talkers using signals that were acoustically modified to preserve different properties that were arguably talker-specific. Listeners heard sentence-length natural, sinewave, or reversed speech samples. Their knowledge of the talker was then assessed using generalization tests in which a novel set of natural, sinewave, or reversed speech samples were used and listeners were asked again to identify the talkers. Our intention was to permit a comparison of the attributes available in the learning conditions and in the generalization tests with those proposed in several classic and recent accounts of individual identification. This comparison allowed us to assess the extent to which talker identification exploits segmental phonetic attributes and to evaluate evidence favoring a dissociation between indexical and phonetic processing in speech perception.

A Traditional View of Word and Talker Recognition

Historically, theoretical treatments of speech perception have separated the features and processes used to perceive and to represent the linguistic content of an utterance from the features and processes used to encode talker attributes (Halle, 1985; Laver & Trudgill, 1979). Accordingly, many accounts have asserted that a word in lexical memory contains abstract phonological forms, which preserve only commonalties across spoken instances (see Goldinger, 1998; Goldinger, Pisoni & Luce, 1996; and Klatt, 1989, for reviews). An emphasis on abstract representations aims to accommodate the variation in the acoustic manifestations of words resulting from differences among talkers (e.g., Peterson & Barney, 1952). For an acoustic signal to be recognized as a particular word, a listener must first convert an idiosyncratic pattern of vocal tract resonances produced in a specific phonetic, social, and affective context to a form approximating a canonical, talker-neutral, linguistic

representation in lexical memory. To accomplish this, talker-specific attributes are separated from phonetic attributes during early perceptual processing, thereby normalizing the speech signal with respect to differences across talkers (Halle, 1985; Joos, 1948; Pisoni, 1997; Studdert-Kennedy, 1974, 1976).

A Traditional View of Talker Recognition

A complementary set of processes is often asserted to underlie the recognition of individuals. The vocal quality of an individual talker is represented using features that are linguistically irrelevant but that characterize the talker across utterances. This dissociation of linguistic and indexical properties in speech perception warrants a listener to separate the perceptual analysis of consonants and vowels from the analysis of vocal characteristics, such as melodic pattern, the roughness or smoothness of the vocal quality, or the speech rate. The qualitative vocal attributes can then be compared with the remembered qualities of familiar individuals in determining that a particular sample is a known individual, thereby accessing other remembered aspects of the talker.

A fundamental goal for researchers has been to discover a pattern of acoustic features that varies in parallel with the differences among a set of talkers. Several acoustic properties have figured prominently in the literature on talker recognition (Bricker & Pruzansky, 1976; Laver & Trudgill, 1979). These include fundamental frequency of phonation, the typical frequencies of the vocal tract resonances, the structure of glottal harmonics, and the fine-grained power spectra of nasals and vowels.

Fundamental frequency, the principal basis for impressions of vocal pitch, is the rate at which the vocal folds open and close in voiced speech, and it is independent of the vocal resonances that principally signal the linguistic content. Differences in fundamental frequency across talkers, or variation in fundamental frequency within an utterance, do not substantially alter the consonants and vowels (although producing certain phonemes can affect fundamental frequency; see Silverman, 1986). It is not surprising, then, that vocal pitch is an extremely salient component of vocal quality and accounts for most of the variance in multidimensional scaling studies of talker recognition (Carterette & Barnebey, 1975; Gelfer, 1988; Matsumoto, Hiki, Sone, & Nimura, 1973; Voiers, 1964; Walden, Montgomery, Gibeily, & Prosek, 1978). Other impressions of vocal quality, such as the degree of breathiness, hoarseness, or creakiness, arise from spectral effects of different modes of laryngeal vibration or from morphological variations in the vocal folds. These glottal features can also be distinguished from those that carry linguistic form. However, factors ascribed to laryngeal quality in one language may be linguistically contrastive in another language (Ladefoged & Ladefoged, 1980).

In addition to anatomical differences, talkers also differ in speaking style. Style can be a reflection of dialect, colloquially termed *accent*, which is described by linguists as a socially and culturally determined variant of expression shared by members of a linguistic community (Trudgill, 1974). Arguably, dialect is also differentiated sexually in American English (Byrd, 1994). In fact, in reassignment of sex from male to female, neither surgery nor administration of hormones affects the vocal anatomy. Yet an individual who retains the phonatory frequency and spectrum of the presurgical sex can be perceived as the surgically assigned sex by altering articulatory style (Gunzburger, 1995; Spencer, 1988). Speaking style also reflects one's idiolect, or idiosyncratic pronunciation habits (Labov, 1986; Laver, 1980, 1991); speakers of a dialect therefore differ in their production of phonetic segments, making their phonetic inventories similar but not identical. The utility of dialectal and idiolectal variation in the perceptual identification and differentiation of individuals has received less attention relative to physiological and anatomical factors (e.g., fundamental

frequency and vocal tract scale), as a result of the widespread assertion that qualitative aspects of speech convey a talker's identity independent of the linguistic phonetic aspects.

Evidence Favoring a Traditional View

There are empirical grounds for arguing that linguistic and talker-specific properties are perceived and remembered independently. Several acoustic manipulations affect the two types of attributes differently. For example, temporal reversal of natural speech distorts temporally based segmental linguistic attributes, affecting consonants, diphthongs, formant transitions, syllable shape, and the relative duration of segments at word initial or word final position. It is impossible to perceive the lexical content of an utterance played backward. The speech seems to be composed of an unknown foreign language, yet the natural vocal timbre is preserved.

Familiar and unfamiliar talkers can be reliably identified from reversed speech samples (Bartholomeus, 1973; Bricker & Pruzansky, 1966; Clarke, Becker, & Nixon, 1966; Van Lancker, Kreiman, & Emmorey, 1985). For example, Van Lancker et al. (1985) tested the recognition of 45 familiar, famous talkers (e.g., John F. Kennedy or Johnny Carson) from veridical and reversed speech samples. They found that in the reversed speech condition, a listener's ability to select the correct talker from among six response alternatives was reduced by only 13% relative to the veridical speech condition. In fact, recognition of several talkers was not impaired at all by temporal reversal, and one talker (Richard M. Nixon) was actually identified better in reversed speech. The authors concluded that talkers can be recognized from "pitch, pitch range, rate, vocal quality, and vowel quality, but without benefit of acoustic detail reflecting specific articulatory and phonetic patterns, and orderly temporal structure" (Van Lancker et al., 1985, p. 30).

Similarly, listeners are able to recognize talkers from filtered speech (where low or high frequencies are selectively attenuated), even though filtering hampers the perception of linguistic properties yet spares the meter and melody and some qualitative aspects of speech (Compton, 1963; Lass, Phillips, & Bruchey, 1980; Pollack et al., 1954). In contrast, listeners have difficulty recognizing talkers when acoustic correlates of laryngeal vibration are removed from the signal, despite the availability of linguistic properties. For example, whispering an utterance substantially reduces talker identification without gravely impairing intelligibility (Meyer-Eppler, 1957; Tartter, 1991; Tartter & Braun, 1994; see also Coleman, 1973). Reversed speech, filtered speech, and whispered speech exhibit a wide range of phonetic attributes. Consequently, these acoustic manipulations do not prevent a listener from using variation in each talker's pronunciation to differentiate among individuals.

Functional data from neuropsychological studies of brain-damaged patients are also often cited as support for the view that linguistic and indexical processing are independent. Consider, for example, recent discussions of phonoagnosia, a deficit in recognizing familiar talkers, associated with parietal lobe damage in the right hemisphere. Within their diagnostic class, patients exhibit normal speech perception and speech production abilities (Van Lancker, Cummings, Kreiman, & Dobkin, 1988). In contrast, patients exhibiting aphasia, a deficit in one or more aspects of language processing associated with temporal lobe damage in the left hemisphere, are able to recognize talkers. The fact that damage to different neuroanatomical substrates can lead to dissociable impairments in either indexical or linguistic processing supports an inference that these dimensions are mediated by independent anatomic structures. However, a closer examination of these neuropsychological case studies reveals combinations of impairments. For example, two of the three phonoagnosic patients described by Van Lanker et al. (1988) also exhibited language disorders (e.g., Wernicke's aphasia, anomia). It remains to be seen how these

mixed cases can be reconciled within a framework that relies on pure case dissociations for evidence of independent indexical and linguistic processing systems (but see Van Orden, Pennington, & Stone, 2001).

An Alternative to a Traditional View

Recent investigations have suggested a different perspective on the relation between linguistic and indexical perception of speech, namely, one in which these two sets of attributes combine during language processing (see Pisoni, 1997, for a review). For example, several researchers have shown that listeners adjust to trial-to-trial variation in talker identity. This evidence comes from experiments showing that identification of phonemes and words is faster and more accurate when items are spoken by a single talker as compared with several different talkers (Cole, Coltheart, & Allard, 1974; Creelman, 1957; Martin, Mullennix, Pisoni, & Summers, 1989; Nearey, 1989; Rand, 1971; Verbrugge, Strange, Shank-weiler, & Edman, 1976). Similarly, the perception of a target word can be shifted systematically by talker-specific formant patterns in a precursor sentence (Broadbent & Ladefoged, 1960; Remez, Rubin, Nygaard, & Howell, 1987). Other experiments show that listeners cannot ignore talker attributes during phonetic processing, even when explicitly instructed to do so (Green, Tomiak, & Kuhl, 1997; Mullennix & Pisoni, 1990). That is, the perception of the talker and the linguistic form of an utterance occur in an integral or contingent fashion.

These findings suggest that a listener calibrates the standards of phonetic perception by attending to indexical characteristics. Other experiments exploring the effects of talker variation on memory show that linguistic perception preserves talker-specific attributes, contrary to a traditional description of spoken language processing (Halle, 1985; Joos, 1948; Ladefoged & Broadbent, 1957; Summerfield & Haggard, 1973). For example, subjects are more accurate at implicitly identifying perceptually degraded words (Church & Schacter, 1994; Goldinger, 1996; Schacter & Church, 1992; Sheffert, 1998b) or explicitly recognizing a repeated word (Craik & Kirsner, 1974; Geiselman & Bellezza, 1977; Luce & Lyons, 1998; Palmeri, Goldinger, & Pisoni, 1993; Sheffert, 1998a; Sheffert & Fowler, 1995) when the same talker, rather than a different talker, speaks the word on the first and second occasions. Apparently, some attributes of the specific instances in which a listener recognizes a word become a part of the long-term memory of that lexical item. In fact, the extra processing time and encoding resources used to perceive and to encode talker-specific properties may actually result in richer, more distinctive episodic representations (Goldinger, Pisoni, & Logan, 1991).

Theoretically, some findings (Church & Schacter, 1994; Schacter & Church, 1992) appear to be consistent with the view that a phonetic segmental sequence is represented in a form that is separate and incompatible with the features used to represent talker qualities. For example, Schacter and Church (1992) found that study-to-test changes in the talker reduced priming on implicit memory tasks but had no effect on explicit memory tasks. The authors interpreted this dissociation as evidence for separate memory systems: an implicit system that reflects memory for instance-specific acoustic features of a word (e.g., the specific talker who spoke it) and abstract word form (e.g., phonological features) and an explicit system that reflects word meaning and other associative properties.

However, data from Church and Schacter (1994) show that the acoustic changes that affect memory performance are limited to those that interact with the identification of individual phonetic segments. For example, word repetitions that differ in talker, fundamental frequency, or intonation contour reduce implicit memory performance. Importantly, each of these manipulations involves a change in pronunciation of phonetic segments. In contrast,

phonetically irrelevant acoustic variations, such as changes in signal amplitude, have no effect on implicit memory performance. Thus, the extent to which an ostensibly nonlinguistic acoustic attribute affects word memory may depend on the degree to which its features are relevant to, or overlap with, those that determine phonetic perception.

In addition, the view that memory of talkers and words are retained in separate systems that are accessed by different tasks does not account for numerous other reports of talker-specificity effects in explicit recognition, such as those obtained by Craik and Kirsner (1974), Palmeri et al. (1993), Sheffert and Fowler (1995), Sheffert (1998b), and others. In fact, when the data are examined across many experiments, the patterning of the findings is more consistent with the view that talker properties and linguistic form are associated within a single memory system (Goldinger, 1996; Sheffert, 1998a). Dissociations may result from functional differences among various types of information and fluctuations in performance resulting from changes in task demands (including reliance on phonetic information).

A series of perceptual learning studies conducted by Nygaard and colleagues (Nygaard & Pisoni, 1998, Nygaard, Sommers, & Pisoni, 1994) provides clear evidence of a link between linguistic and indexical identification. In these experiments (Nygaard & Pisoni, 1998), listeners were trained over several days to identify a set of talkers from sentence-length utterances. After training, they were asked to transcribe new sentences presented in noise. Surprisingly, listeners who were familiar with the talkers transcribed test sentences more accurately than listeners who were unfamiliar with the talkers. That is, phonetic perception was facilitated by experience with the talkers. This finding is strong evidence that linguistic and nonlinguistic properties of speech are concurrent. Sheffert and Olson (2000) recently extended these findings by showing that novel utterances produced by familiar talkers are more likely to be retrieved from long-term episodic memory. These data generalize prior results and indicate that the effects of familiarity are durable.

The studies reviewed so far indicate that experience with a talker, whether it is derived from a single trial or from several days of training, affects the linguistic analysis of speech. These results suggest the possibility that linguistic and indexical attributes are transmitted in parallel, using some of the same features. A wide range of effects on word processing brought about by talker variation and talker familiarity would be explained by assuming that words and talkers share a common representational code. Unfortunately, the complexity of natural speech makes it difficult to determine whether a common representation might be based in acoustic attributes, phonetic segmental features, global suprasegmental features, or a combination of these.

A Common Code

In a recent experiment, Remez, Fellowes, and Rubin (1997) tested the hypothesis that a common code for representing both talkers and words is phonetic in nature. Remez et al. reduced the acoustic dimensionality of speech by applying the technique of sinusoidal replication. Sinusoidal synthesis (Rubin, 1980) generates a nonspeech sinusoidal (pure tone) pattern that tracks the changing formant center frequencies of a naturally produced utterance. Sinewave speech can be thought of as an acoustic caricature of the original utterance, lacking fine-grained acoustic details of natural speech and which evoke impressions of natural vocal quality, such as a fundamental frequency, broadband resonances, harmonic structure, and various aperiodic elements. Sinewave signals evoke impressions of the segmental phonetic attributes of speech despite unspeechlike timbre and anomalous intonation, and, consequently, most listeners are able to perceive the linguistic content of a sinewave utterance (Remez, Rubin, Pisoni, & Carrell, 1981).

The study by Remez et al. (1997) further showed that these phonetic properties of the sinewaves also preserve talker-specific aspects of speech. In this experiment, sinewave utterances modeled from the natural speech of 10 talkers were presented to listeners in a test of individual recognition. The listeners had become highly familiar with the talkers over many years of informal social contact. Listeners were quite successful in identifying their colleagues from the sinewave presentation of utterances. This result arguably demonstrates that the perception of a talker's characteristics is prominently available in a kind of talker-specific aggregation of phonetic properties (Remez et al., 1997). The authors suggested that word and talker recognition can therefore use a common code composed of segmental phonetic attributes.

The results also showed considerable individual variation in the recognizability of different talkers, as evidenced by the fact that in some conditions recognition accuracy was below chance for 2 of the 10 talkers in the set. Because Remez et al. (1997) did not manipulate or control a listener's familiarity with the talkers in the test set, it is impossible to know whether the observed variation was due to differences in the degree or quality of familiarity or to other factors in the talker ensemble, such as perceptual distinctiveness or discriminability of the specific samples.

To address this issue, we used a laboratory-based training procedure to control for the familiarity of talkers in tests composed of materials from Remez et al. (1997). The training task allowed us to pose a critical question: Is it possible to learn to recognize a talker from an acoustically unnatural signal? A perceptual learning paradigm enabled us to examine in detail how variation in the rate and degree of perceptual learning affects the identifiability of different talkers.

The Present Study

Our investigation had several objectives. First, we sought to establish whether perceptual learning of talkers can take place in the absence of the acoustic correlates of vocal quality typically assumed to underlie talker identification. This finding offers evidence that phonetic segmental properties alone can specify individuals and support talker learning. Second, because listeners were able to learn to identify the talkers from sinewaves, we aimed to assess the flexibility, abstractness, and feature structure of the talker categories that develop during perceptual training by measuring generalization using novel sinewave and natural speech samples. A final objective was to assess the extent to which the attributes of an individual talker transfer to a different type of signal. We accomplished this by testing whether a talker who is easy to identify from a sinewave token is also easy to identify from a natural speech token.

In subsequent experiments, we used the same paradigm to examine talker learning from natural speech and reversed speech and generalization to natural, sinewave, and reversed speech samples. (See Table 1 and the *Method* section from Experiment 1.) In this series of tests, we aimed to characterize the relation between identifying individual talkers and apprehending the linguistic form of the message.

Experiment 1: Learning to Identify Talkers From Sinewave Sentences

To determine the ability of listeners to identify a talker without relying on natural vocal qualities, we conducted a test in which the sentences used were sinewave replicas of natural utterances. Subjects at the outset were completely unfamiliar with individuals in the talker ensemble. We trained the subjects to a criterion of 70% accuracy. Knowledge of the talkers was then assessed using two generalization tasks in which listeners heard a novel set of sentences and were asked to identify the talker on each trial. In one generalization test, the

sentences were sinewave replicas, and in the other generalization test, the sentences were natural samples. In both cases, the generalization tests used novel utterances that the subjects had not heard during training.

On the basis of the findings of Remez et al. (1997), which demonstrated talker identification from sinewave utterances, we predicted that with sufficient training, the phonetic attributes available in sinewave replicas would support the perceptual learning of individuals. We also expected that this knowledge would incorporate durable features of a talker's articulatory habits and would not simply manifest rote memorization of the specific acoustic items. However, we expected generalization performance to reflect the known examples and, consequently, to be best in the condition in which the acoustic form of the samples was the same in training and test phases. Specifically, we expected performance to be better on a sinewave generalization test than on a natural speech generalization test.

The method used in Experiment 1 is very similar to the method used throughout the experiments reported in this article. Therefore, our procedure is described in detail in Experiment 1, and we note only departures from the general method in subsequent experiments.

Method

Listeners—Nineteen adults were recruited from the Bloomington, Indiana, community using an advertisement in the local newspaper. Of these, 5 did not complete the study for personal reasons, and 6 were excused because of slow progress during initial training.[1] The remaining 8 subjects completed the sinewave training phase and the two generalization tests. All subjects in this experiment and subsequent experiments were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. None of the listeners were familiar with the test materials. They were paid $5 per hour for their participation.

Test Materials—The natural and sinewave sentences used in the present experiments were used in Remez et al. (1997). The test items consisted of two sets of sentences. The first set contained nine natural utterances produced by each of five male and five female talkers (see the Appendix). The talker ensemble was relatively heterogeneous, representing different American English and British English dialects. (Talkers F3 and M4 were the British speakers.) Each talker read the nine sentences aloud in a natural manner. The sentences were recorded on audiotape in a soundproof booth and were low-pass filtered at 4.5 kHz, digitally sampled at 10 kHz, equated for root-mean-square (RMS) amplitude and stored as sampled data with 12-bit amplitude resolution.

The second set of sentences consisted of sinewave replicas of the original natural speech tokens. To create these items, the frequencies and amplitudes of the three oral formants and the intermittent nasal and fricative formants were measured at 5-ms intervals, relying interactively on two representations of the spectrum: linear predictive coding and discrete Fourier transform. Three time-varying sinusoids were then synthesized to replicate the oral and nasal formant pattern, and a fourth sinusoid was synthesized to replicate the fricative pattern, based on the center frequencies and amplitudes obtained in the acoustic analysis (Rubin, 1980). The sinewave synthesis procedure preserved patterns of spectrotemporal

[1]The performance of these subjects began to asymptote after several sessions, and they were excused after expressing frustration with their lack of progress. During debriefing, the subjects indicated that they found learning sinewave voices to be inordinately difficult and had trouble remaining vigilant and motivated during the training sessions. We felt it was prudent to excuse these individuals rather than allow them to become increasingly frustrated or bored. Similar difficulties were experienced by subjects in a later experiment using reversed speech (see Experiment 5).

change of the vocal resonances while eliminating the fundamental frequency, harmonic relations, and fine-grained spectral details of natural speech. Subjectively, the sentences were unnatural in vocal quality.

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 specific sentences or to the order of presentation of specific items.

Procedure

Training phase: Listeners were trained over several days to name the 10 talkers of the sinewave utterances. They were tested in groups of three or fewer in a quiet listening room. During each training session, each subject heard a random ordering of five repetitions of three sentences from each talker (150 items total). There was no blocking by talker or sentence. The same three sentences were used for each talker in each training session. Following Legge et al. (1984), we encouraged attention to talker attributes rather than to the content of the message by telling subjects beforehand which sentences they would be hearing and by posting a printed list of the sentences next to the CRT display.

The sinewave training sentences were presented binaurally to subjects at 75 dB over matched and calibrated stereophonic headphones (Model DT100; Beyerdynamic, Farmingdale, NY). Each subject was asked to listen carefully to each sentence and to pay close attention to the qualities that seemed to distinguish individual talkers. Each time a sentence was presented, the subject was asked to press 1 of 10 buttons labeled with each talker's name on a computer keyboard. Keys 1–5 were labeled with female names and keys 6–10 with male names. All the names were common, monosyllabic names, such as "Ann," "Mike," or "Bob." After each response, the accuracy of the response and the name of the correct talker were displayed on the computer screen in front of the subject and recorded in the computer. Each training session lasted approximately 30 min. Training continued until each subject achieved an average of 70% correct talker recognition performance.

Familiarization phase: Before beginning each of the generalization tests, all subjects completed a brief familiarization task to reinstate the correspondence between the sinewave tokens and the talker's names. The familiarization task was simply an abbreviated version of a training session in which subjects listened and responded to one instance of each sentence produced by each talker (30 items total). The items were presented in a random order, and accuracy feedback was given after each response. The familiarization task lasted approximately 8 min.

Generalization tests: After reaching a 70% correct criterion in the sinewave training phase, each subject completed two generalization tests. One generalization test presented three unfamiliar sinewave sentences, whereas a second test presented three unfamiliar naturally produced sentences. 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 test presented five repetitions of each of the three sentences in a random order (150 items total). Once again, subjects were provided with a transcription of the sentences they would be hearing. Their responses were not corrected during either of the two generalization tests.

Results and Discussion

Training Performance—Analysis of the training data revealed that listeners were, in fact, able to learn to identify the 10 talkers from sinewave signals. Training performance showed continuous improvement (at least for the two thirds of the original sample who completed the study). After the first training session, talker identification performance was above chance and steadily increased by an average of 5% each day. By the last day of training, listeners were able to identify the talkers with a mean accuracy of 76%. Figure 1 displays the learning data from all the subjects. Each subject's talker identification performance is displayed as a function of training days and talker sex. Figure 1 illustrates that perceptual learning progressed at different rates for different subjects. Some subjects became attuned to the talker-specific attributes relatively quickly, whereas others displayed extremely slow progress. The number of days needed to reach the 70% criterion varied from 9 days to 16 days.

The data from the last day of training showed variation in the identifiability of talkers within the training set (see Figure 2). A one-way repeated measures analysis of variance (ANOVA) revealed a significant effect of talker identity on the recognition performance, $F(9, 63) = 9.71$, $p < .0001$, $MSE = 154.52$. This indicates that sinewave samples differ in distinctiveness and identifiability. To determine which talker differences were significant, we calculated the critical mean difference using the Scheffé procedure. The overall mean difference between male and female talkers (15.5%) was not significant, although the best recognized talker was a female (F2) and the worst recognized talker was a male (M2). Moreover, some of the differences between individual talkers within each sex were significant (i.e., F2 vs. F4, M2 vs. M4). This suggests that the extent of the variation between female and male talkers as a group was not greater than the difference among the talkers within each sex.

Generalization Performance—To accommodate the large differences in the identifiability of talkers in the initial training test, we normalized the generalization performance relative to initial learning. This was accomplished by dividing the talker identification accuracy on the generalization test by talker identification accuracy on the training task (using the last day of training).

At the time of the generalization testing, half of the subjects received the natural speech test before the sinewave generalization test, and the other half received the opposite order. To assess whether the order of the tests affected talker recognition, we conducted a repeated measures ANOVA on the generalization data, with generalization test and talker identity as within-subject factors and test order as a between-subjects factor. The main effect of test order was not significant ($p > .10$), nor did it interact with any variable. Consequently, we pooled the data from the two test order groups, and subsequent analyses are based on the combined data from both groups.[2]

Figure 3 displays the generalization scores for the natural speech test (top panel) and sinewave generalization test (bottom panel) for each talker. The generalization data from both tests show that talker-specific knowledge acquired during perceptual learning of sinewaves generalized to novel natural and sinewave sentences and was not dependent on the specific samples used during training. This indicates that listeners learned something general or abstract about a talker's speech. Moreover, the same level of generalization occurred in both tests, even though the natural speech condition used sentences that differed

[2]Possible effects of test order were assessed in all subsequent experiments. In each case, the effect of test order was not reliable and did not interact with generalization test type, talker sex, or talker identity. Consequently, data from both test order groups were pooled to form a single composite test group.

substantially in content and acoustic form from the training items. Specifically, talker recognition decreased from 76% correct at the end of training to 46% correct for the natural test and 44% correct for the sinewave test (with chance approximating 10% correct).

An ANOVA comparing the overall means from each of the three conditions revealed a significant effect, $F(2, 7) = 532.23$, $p < .0001$, $MSE = 0.01$. Talker recognition was different from training performance for the natural speech trials, $t(7) = 7.34$, $p < .001$, root-mean-square error (RMSE) = 0.04, and for the sine-wave replica trials, $t(7) = 11.18$, $p < .0001$, RMSE = 0.03. However, the difference in performance between the two generalization tests was not significant. We also found that the talker differences we observed in the initial training data were present in the data from each generalization test. Separate one-way ANOVAs showed that the factor talker identity affected performance on the natural speech test, $F(9, 63) = 2.53$, $p < .02$, $MSE = 0.18$, and the sinewave test $F(9, 63) = 3.38$, $p < .002$, $MSE = 0.05$.

Another way to examine the relationship between perceptual learning during training and generalization performance is to consider the pattern of talker identification across each condition. Individual talkers differed in identifiability at training, with some talkers easier to recognize than others. To what extent did this relative ranking of talkers correlate across training and test conditions?

To assess this, we correlated the proportion of correct identification for each talker across the training and test conditions. As expected, the data showed that individual sinewave talker identification at training was highly correlated with sinewave talker identification at test, $r(8) = .81$, $p < .004$. Sinewave talker identification was also highly correlated with natural speech talker identification, $r(8) = .85$, $p < .002$. This finding can be accounted for if we assume that listeners were able to resolve each talker's articulatory habits from natural speech and from sinewave replicas of speech.

In summary, the data from Experiment 1 provide evidence that naive listeners can learn to identify different talkers solely from the phonetic attributes preserved in sinewave signals, in the absence of the traditional qualitative attributes of vocal sound production. We also observed striking differences in the identifiability of talkers within our training set. One motivation for training all our listeners to a specific criterion was to determine whether the variation in talker identification observed by Remez et al. (1997) arose from a priori differences in the familiarity of listeners with each talker. Our data suggest that perceptual distinctiveness or discriminability of the talkers in the set is the primary source of the differences in identification performance, and not the result of differences in familiarity.

Our training method also allowed us to track the development of each listener's perceptual learning over time. For instance, the training data show that individual listeners differed in their ability to learn to identify the talkers. We considered several listener-specific variables, such as age, sex, prior musical training, and bilingualism, to determine whether these factors interacted with talker learning (Cook & Wilding, 1997; McGehee, 1937; Thompson, 1985; Van Wallendael, Surface, Parsons, & Brown, 1994). These characteristics did not reliably differentiate fast and slow learners, learners who identified same-sex talkers more accurately than opposite-sex talkers, or listeners who were generally more accurate across all talkers. Exactly what makes some people excel in recognizing talkers is not known and is difficult to determine because the characteristics most relevant for talker recognition may differ from talker to talker, from listener to listener, and perhaps from occasion to occasion (see Kreiman, Gerratt, Kempster, Erman, & Berke, 1993).

Another provocative finding of our study is that perceptual learning from the sinewave training task generalized to novel natural and sinewave sentences. Apparently, subjects were

not simply memorizing a set of auditory forms and associating a proper name with it but were abstracting specific attributes of a talker's speech, which could then be used to recognize the same talker producing another utterance. The fact that generalization was similar across the two different tests suggests that the same acoustic–phonetic correlates of talkers were used in both instances. This conclusion is bolstered by the fact that generalization performance with both sinewave speech and natural speech were highly correlated with sinewave training performance. Taken together, these findings are consistent with the proposal that individual attributes of a talker are carried by segmental phonetic properties in addition to vocal timbre and that the apprehension of linguistic and individual attributes can converge in a common representational code.

Experiment 2: Learning to Identify Talkers From Natural Speech

The purpose of Experiment 2 was to determine whether the symmetry in generalization performance that occurred following training on the sinewave utterances is particular to that learning condition, or whether a similar pattern of generalization can be found when listeners learn to identify talkers from natural speech. The design and method of Experiment 2 were identical to those used in Experiment 1, except that the training sentences were natural speech samples. Generalization was assessed using natural speech and sinewave tests. We expected that the perceptual learning of talkers from natural speech would proceed very rapidly and would readily generalize to novel natural speech sentences.

It was more difficult, however, to predict the generalization of talker-specific knowledge to sinewave samples. One likely outcome was that training with the natural speech would facilitate talker identification from sinewaves. This prediction is based on the data from Experiment 1 and Remez et al. (1997), which show that phonetic characteristics are available from sinewave replicas of their natural utterances. An alternative outcome, no less likely, was that training with natural speech samples would not transfer to the sinewave replicas. This prediction is based on the findings of Nygaard and Pisoni (1998). They found that learning to identify talkers from sentence-length materials did not generalize to isolated words and did not improve the intelligibility of word-length materials. Nygaard and Pisoni explained this by suggesting that speech samples presented in sentence form draw a listener's attention to global suprasegmental properties of utterances, such as the characteristic pitch and compass of intonation, and syllable meter. To identify a talker from an isolated word, a listener must attend to more fine-grained segmental properties of a talker's articulatory habits. If this account applies to the present circumstances, it means that coarse-grained qualitative features available from natural sentences may not match the segmental phonetic features available to identify talkers from sinewaves. Moreover, in natural speech samples the differences in vocal quality across a set of talkers is prominent, perhaps more salient than the subtle distinctions of idiolect. If learning favors these qualitative attributes, to the detriment of segmental properties, then we should find poor transfer to sinewave sentences because they lack these qualitative attributes.

Method

Listeners—Eight adults participated in Experiment 2 in exchange for payment ($5 per hour).

Test Materials and Procedure—The materials used in this experiment were identical to the sentences used in Experiment 1. Three sentences were randomly selected without replacement for the natural speech training, natural speech generalization, and sinewave speech generalization tasks. All sentences were rotated through all conditions for each subject to prevent effects of specific items or test orders.

The general procedure for Experiment 2 was consistent with the previous experiment. However, in this experiment, listeners were trained to name the 10 individuals from samples of natural speech. Natural speech tokens were also used in the familiarization task that preceded the generalization tests. The two generalization tests were identical to those used in Experiment 1.

Results and Discussion

Training Performance—As we expected, listeners learned to identify the 10 talkers from the naturally produced sentences very rapidly, usually within 1 day. Five subjects reached criterion after only a single training session. The remaining 3 listeners reached the criterion by the end of the second session. The rapid learning indicates that the natural sentences provided listeners with a salient sample of each talker's indexical attributes.

Figure 4 displays identification performance on the last day of training as a function of talker. A one-way repeated measures ANOVA revealed an effect of talker identity, $F(9, 63) = 5.46$, $p < .0001$, $MSE = 212.10$. Post hoc comparisons revealed significant differences in identification performance between talker M5 and talkers F1, F3, F4, and M4 (all four mean differences exceed the Scheffé critical mean difference of 31%). The average performance of the male talkers did not differ significantly from the female talkers.

Generalization Performance—Figure 5 displays the generalization scores for the natural speech test and sinewave replica test for each talker. Because accuracy on the natural speech test exceeded performance on the sinewave test for each talker, the generalization data are displayed together, with the natural test data represented by the height of each bar and the sinewave replica data as a portion of each bar (see the darkened segments). The data for the two generalization tests differed markedly. Listeners' ability to recognize individuals was 88% for the natural speech generalization test but only 27% for the sinewave generalization test.

An ANOVA comparing the overall means from each of the three conditions (training, natural, and sinewave) revealed a highly significant effect, $F(2, 7) = 293.59$, $p < .0001$, $MSE = 0.01$. Surprisingly, performance on the training task (78% correct) was reliably lower than performance on the natural speech test, $t(7) = 3.17$, $p < .05$, RMSE = 0.03. This effect may be the result of the familiarization task preceding the generalization tests. The purpose of the familiarization task was to remind subjects of the correspondence between a particular name and specific talker, and the task itself is an abbreviated training task. Although the familiarization task presented only 30 items, it apparently improved a listener's talker knowledge. Training performance was, however, significantly higher than performance on the sinewave replica generalization test, $t(7) = 21.5$, $p < .0001$, RMSE = 0.02. The two generalization tests differed reliably from each other, $t(7) = 22.89$, $p < .0001$, RMSE = 0.03, and performance on the sinewave test exceeded chance (10%), $t(7) = 8.38$, $p < .0001$, RMSE = 0.03.

A significant effect of talker identity was found in the natural speech condition only, $F(9, 63) = 7.08$, $p < .0001$, $MSE = 0.017$. Post hoc Scheffé tests confirmed that two of the male talkers (M2 and M5) were identified significantly less accurately than most of the female talkers. Nonetheless, the overall average for the male talkers did not differ significantly from the female talkers. There were no talker effects in the sinewave condition.

The relationship between perceptual learning during training and the pattern of talker generalization on each test was also assessed. The analysis revealed that proficiency in identifying a talker during the training protocol was highly correlated with performance on the generalization test with natural samples, $r(8) = .86$, $p < .002$, and was moderately

correlated with performance on the generalization test with sinewave samples, $r(8) = .64$, $p < .05$. The pattern of correlation shows that the relative ease of identifying the 10 talkers from natural samples was preserved only roughly with sinewave replicas.

In summary, Experiment 2 showed that listeners easily learned to identify a talker from naturally produced sentences. Moreover, talker identification at training was similar in magnitude and in pattern to the natural speech generalization test.

A clue about the function used to perceive and to remember the attributes of a talker is provided in the results of the sinewave speech generalization test. In Experiment 2, indexical knowledge acquired during training with natural speech did not generalize well to sinewave utterances. In contrast, sinewave generalization was much better following training with sinewave utterances (Experiment 1). One interpretation of these findings is that segmental phonetic features are not immediately exploited when a listener is first learning to identify a talker. Perhaps an ordinary form of attention during the encoding of a new talker's vocal characteristics focuses on vocal quality and global suprasegmental attributes such as pitch height and range (see Nygaard & Pisoni, 1998; Van Lancker et al., 1985).

However plausible this conclusion is, the correlation analyses do not corroborate it. Specifically, the correlation between the natural training and sinewave test conditions was significant, albeit smaller than the correlation between the training and natural test conditions. This finding raises a possibility that the talker-specific representations of natural speech include much more phonetic detail than the generalization test suggests. This was certainly true in the experiments of Remez et al. (1997), in which listeners had learned to identify acquaintances informally from natural samples and generalized readily to sinewave instances.

An alternative interpretation of the sinewave generalization results is that our listeners were indifferent to the segmental phonetic attributes conveyed by the sinewaves, favoring instead a more superficial analysis aimed at first deriving each talker's sex and then merely guessing the identity of a talker from among the five male or five female names. If this strategy were used, chance performance would be 20% rather than 10%, a value not very different from the performance levels obtained on the sinewave test in this experiment (27%) and subsequent tests reported in this article that assessed sinewave talker identification after training on natural speech utterances (control experiment = 26% and 28%; Experiment 4 = 22%). This view also predicts the following: (a) Within-sex talker identification should be consistently poor across the five talkers and (b) male sinewave talkers should not be misidentified as female talkers or vice versa.

We evaluated these predictions and found no evidence to support them. For instance, within each group of male and female talkers, identification accuracy was much more variable than would be expected if performance were based primarily on identification of talker sex, followed by a random choice within sex. Listener responses (both correct and incorrect) were not evenly distributed across same-sex talkers, nor was there any evidence that just one or two responses were used by a given listener for all talkers (e.g., biased guessing). Further examination of each subject's errors revealed many instances in which a male was mis-identified as a female and a female misidentified as a male, and such errors occurred for all talkers and were made by every listener. Specifically, 28.5% of the listener errors (averaged across all subjects) were cross-sex misidentifications. Overall, female talkers were less likely to be confused with male talkers (17% vs. 39% cross-sex error rate for female and male talkers, respectively). For example, talker F3 was almost never confused with a male (1% cross-sex error rate), whereas talkers M1 and M5 were often misidentified as a female (43% cross-sex error rate for both talkers).

We also found that a listener's tendency to misidentify a talker's sex was unrelated to overall talker recognition accuracy or to accuracy at recognizing a particular individual. To illustrate this, consider the results from 1 subject who showed good recognition of some of the sinewave talkers (e.g., 73% correct on talker "M4") while also showing that M3 was misidentified as a female talker on 50% of the incorrect trials. In contrast, this subject's accuracy on talker M1 was low (7%), yet also showed a similar degree of cross-sex misidentification, with M1 misidentified as a female on 43% of the incorrect trials. As a whole, the pattern of perceptual errors in the data from this experiment is representative of the error data from other sinewave tests we report in this article and counts as evidence against the conjecture that our listeners were simply categorizing sinewave talkers by sex.

An explanation of the relatively poor sinewave generalization performance observed in Experiment 2 may lie in the method we used in our experiment. In the present study, a listener's ability to recognize individuals was only 27% in the sinewave generalization test but was approximately 55% in the talker identification test used by Remez et al. (1997; see their Figure 4). Three differences in method may account for the difference in sinewave talker identification performance across these two experiments. First, in our experiment, familiarity was acquired through training with a small number of sentences over a few days. In contrast, the listeners in the report of Remez et al. had become familiar with the talkers over several decades. Second, our generalization test task presented three sentences three times, in random order, whereas Remez et al. presented the same sentence six times. In this case, the use of the same linguistic content of the utterance from trial to trial may have improved a listener's ability to discriminate among individuals (Read & Craik, 1995). Third, it is also possible that the novelty of the sinewaves led our subjects to focus on the unusual auditory impressions that these tonal signals evoke, rather than on the transfer-relevant phonetic properties. We conducted a control experiment to address this issue.

The procedure replicated the sequence used in Experiment 2, with the exception of an additional transcription task interpolated between the training from natural speech items and of testing generalization with sinewave and natural speech items. The transcription task offered 29 samples of sinewave speech, modeled on the speech of a male talker who was not among the talkers used in the training set, permitting subjects to become accustomed to the unusual sound and linguistic properties of sinewaves.

We compared two transcription presentation methods in which subjects heard several repetitions of each sentence either in a random order or blocked by sentence. In both conditions, subjects transcribed the entire sentence after its last repetition. We found that transcription accuracy did not differ across groups (random = 49%, blocked = 42%); neither did the transcription task improve sinewave generalization (27% correct). In addition, the extent to which a subject was able to transcribe the words conveyed by sinewaves proved to be a modest predictor of the ability to recognize talkers from these patterns, $r(16) = .68$, $p < .002$. The null effects of the transcription task may mean that our subjects needed more substantial experience with sinewave utterances or that the processing requirements of the transcription task, word identification, did not overlap sufficiently with the requirements of the generalization test, talker recognition.

Alternative Conceptualizations—An alternative to the sinewave familiarity hypothesis tested in the control experiment is that the poor transfer from natural speech to sinewave replicas results from a mismatch between the attributes readily available and attended to during perceptual learning and those that are prominent during generalization. Under ordinary circumstances, many features are available to index a specific talker. Facing a demand to quickly learn about new individuals, a perceiver hypothetically can rely on a small sample of a talker's indexical attributes, biased toward the most prominent properties:

qualitative characteristics such as pitch, meter, or timbre, to name a few. These attentional biases can arise in the moment, from the perceptual salience of certain indexical attributes, or can stem from a lifetime of experience identifying talkers. Of course, this set of properties is absent from sinewave sentences, which exhibit other, arguably less prominent indexical attributes: idiosyncratic vowel expression, spirantized stop releases, or consonant assimilation, to name a few.

We aimed therefore to devise a test to determine whether the pattern of results obtained in Experiment 2 was due to a shift in attention between qualitative and segmental phonetic characteristics of a talker (e.g., shifting to the latter following sinewave training). One method was to assess generalization to a spectral form that preserves the qualitative properties prominent in natural speech. To establish a parallel to the sinewave conditions, the spectral form would also have to be acoustically unusual and unfamiliar.

Reversed speech met this designation. Temporal reversal of natural speech disrupts some of the attributes of the signal, leaving other short- and long-term properties relatively unaltered. Time-critical and ordinal segmental properties are distorted. Some consonants are distorted because they exhibit a patterned sequence of the acoustic correlates of an articulatory hold-and-release; likewise, the more slowly changing diphthongs ordinarily manifest a gradual ordered change in resonant frequency, which is distorted by a reversal. Reversal also disrupts the metrical and melodic contour of a syllable train. The sustained aperiodicity of fricatives is unaltered in reversal, although the transition to voicing at release of the articulation is distorted, making this class of consonants partly impaired by reversal. Slowly changing vowel nuclei are the least distorted by temporal reversal. Accordingly, it is impossible to perceive much of the specific segmental or lexical content of an utterance played backward, and a speech sample of English prepared in this manner sounds like an unfamiliar language.

Temporal reversal does not eliminate all phonetic attributes. Nevertheless, the phonetic properties that remain in these signals are insufficient to allow lexical access of the original articulated sequence of words. Also, reversal does not hamper acoustic transmission of the long-term spectrum of a talker's speech, nor does it alter the frequency or frequency range of glottal pulsing. Consequently, this acoustic transformation provides reliable acoustic correlates of vocal pitch height and variation and of speaking rate; of the central tendency and range of formant frequency variation; and, at slowly changing syllable nuclei, even of talker-specific vowel quality.

Accordingly, reversed speech is rich in indexical qualities and is relatively poorer in conveying the segmental grain in which idiolect is defined. Consequently, listeners are usually able to recognize talkers from reversed speech (Bartholomeus, 1973; Bricker & Pruzansky, 1966; Clarke et al., 1966; Van Lancker et al., 1985), and we can be confident that the basis of this ability rests largely on attention to qualitative as opposed to idiolectal attributes.

In the first two experiments we report in this article, we investigated whether a listener would learn to recognize a talker when the samples featured phonetic segmental properties to the detriment of qualitative aspects of the talker's speech. In the next three experiments, we explored how listeners learn to categorize unfamiliar talkers when qualitative attributes are featured to the detriment of segmental phonetic and lexical properties.

Experiment 3: Recognizing an Unintelligible Talker

Experiment 3 was designed to test the hypothesis that subjects primarily exploit glottal source quality rather than fine-grained phonetic properties during natural speech training.

Testing this hypothesis required a comparison of generalization to reversed and sinewave speech following training with natural speech. Because reversed speech preserves the glottal source whereas sinewave replicas do not, the comparison is a measure of the importance of qualitative aspects of the glottal spectrum for talker identification. Moreover, because reversed speech seems unusual and is unintelligible, this test also provided a control for the possibility that the mere peculiarity of an acoustic signal is a critical determinant of generalization performance.

Method

Listeners—Nine adult listeners volunteered to participate in exchange for payment ($5 per hour). One was excused from the study for not attending a training session.

Test Materials and Procedure—The training materials were natural speech sentences. The generalization test materials were sinewave replicas and temporally reversed natural sentences. The reversed speech was created by inverting the series of sampled values of the natural items using a signal processing program (Cool Edit 96; Syntrillium Software Corporation, 1996). Other aspects of the training, familiarization, and testing were consistent with the previous experiments.

Results and Discussion

Training Performance—The results of the natural speech training are nearly indistinguishable from the previous instances of this condition. There was significant variation in the identifiability of the 10 talkers, $F(9, 63) = 10.57$, $p < .0001$, $MSE = 167.65$. There was no difference in the average identifiability of male and female talkers.

Generalization Performance—The generalization data are of primary interest in this experiment (see Figure 6). Listeners' ability to recognize talkers declined from 79% correct at the end of training to 53% correct for the reversed speech test (white portion) and 22% correct for the sinewave test (dark portion). The overall means from the three conditions differed, $F(2, 14) = 76.04$, $p < .0001$. Performance differed significantly between training and the reversed speech trials, $t(7) = 5.65$, $p < .001$, RMSE = 4.66, and the sinewave replica trials, $t(7) = 10.49$, $p < .0001$, RMSE = 5.48, and between each generalization test, $t(7) = 8.46$, $p < .0001$, RMSE = 3.68. Performance in the sinewave trials exceeded chance, $t(7) = 2.91$, $p < .02$, RMSE = 7.39.

To resolve the pattern of results more sharply, we assessed the similarity between the patterns of talker identification during training and generalization. Talker identification from natural speech training was significantly correlated both with reversed speech identification, $r(8) = .71$, $p < .02$, and sinewave talker identification, $r(8) = .77$, $p < .009$. These values indicate a good match between the features learned during training and the parameters used to identify a talker at generalization.

In summary, the data of Experiment 3 showed that subjects were able to identify a familiar talker from reversed speech after training on natural speech. In contrast, talker recognition from sinewave signals was far poorer, approaching chance. The fact that reversed speech generalization was fairly accurate suggests that the qualitative attributes of a talker's speech are perceptually prominent during learning from natural samples. Moreover, although performance was poor on the sinewave generalization test, the correlation analyses show that listeners also appear to encode some fine-grained phonetic attributes well enough to preserve the pattern of relative memorability of the individuals despite an acoustic transform that eliminates the qualitative aspects of vocal sound. This may indicate that under normal circumstances, listeners naturally encode talkers using a mixture of different properties,

perhaps relying most on qualitative and suprasegmental characteristics, while also using phonetic details.

Experiment 4: Getting to Know an Unintelligible Talker

In Experiment 4, we assessed listeners' ability to learn to identify talkers from reversed speech. Time reversal eliminates or disrupts many of the fine-grained acoustic patterns necessary for phonetic perception and, accordingly, for word recognition. This experiment tests the sufficiency of qualitative aspects of speech in the relative absence of many phonetic and all lexical impressions. To date, no researchers have examined listeners' ability to learn about talkers from reversed speech under controlled conditions. (All prior reports examined identification of famous talkers learned through incidental exposure from the media [cf. Van Lancker et al., 1985].) We expected that learning to identify a talker from reversed speech would be easier than learning about a talker from sinewaves because reversed speech preserves a wider variety of attributes, many of which are associated with natural vocal quality, such as timbre and pitch. If listeners learn first to distinguish talkers along qualitative perceptual dimensions, rather than along idiolectal dimensions, we should find rapid learning from samples of reversed speech. We also sought to determine whether reversed speech samples are informative about the natural attributes of a talker. We expected to find reasonably good generalization to novel reversed speech tokens but poor generalization to sinewave replicas, on the principle that qualitative attributes learned from reversed speech are incommensurate with the grain of segmental phonetic attributes preserved without natural vocal quality in sinewave replicas.

Method

Listeners—Eight adults were recruited from the Bloomington community and were paid $5 per hour for their participation.

Test Materials and Procedure—The training materials were reversed speech sentences. The generalization materials were reversed speech and sinewave speech sentences. Other aspects of the training, familiarization, and testing procedures were identical to the previous experiments.

Results and Discussion

Training Performance—Listeners learned to identify individuals from reversed speech samples at a rate intermediate between training with natural speech and training with sinewave replicas. The majority of listeners (five of the eight) reached criterion after three training sessions, and the remaining listeners required between 6 and 11 sessions to reach criterion. Figure 7 displays the talker identification performance as a function of training days and talker sex. For all listeners, performance was above chance after the first training session and increased by an average of 14% each day for the five fast learners and by an average of 5% each day for the three slower learners.

Figure 8 shows talker recognition performance on the last day of training as a function of talker. There were no significant differences in the identifiability of talkers.

Generalization Performance—Figure 9 displays the generalization scores for the reversed speech and sinewave generalization tests for each talker. A listener's ability to recognize talkers decreased from 72% correct at the end of training to 59% correct for the reversed test and 16% correct for the sinewave test. The overall means from each of the three conditions were significantly different, $F(2, 14) = 138.69$, $p < .0001$, $MSE = 48.60$. Training differed from the reversed speech generalization trials, $t(7) = 2.76$, $p < .05$, RMSE

= 4.23, and from the sinewave replica trials, $t(7) = 32.41$, $p < .0001$, RMSE = 1.70. The generalization tests differed significantly from one another, $t(7) = 10.98$, $p < .0001$, RMSE = 3.95, and performance on the sinewave test exceeded chance, $t(7) = 3.98$, $p < .005$, RMSE = 4.02. Additional analyses indicated that there were no differences in recognition among the talkers.

The correlations comparing the relative identifiability of talkers across training and generalization conditions show that talker identification from reversed speech training was highly correlated with reversed speech talker identification, $r(8) = .81$, $p < .004$. In contrast, there was no correlation between training and sinewave talker identification, $r(8) = .33$, $p < .35$. This latter finding is strong evidence that the attributes used to learn the talkers were unavailable in sinewave replicas. That is, the features with which listeners became familiar during training did not correspond to the indexical attributes available in sinewave replicas.

Experiment 5: The Robustness of Qualitative Indexical Attributes

In Experiment 5, we trained listeners to recognize talkers using reversed speech. However, the generalization of that knowledge was assessed using natural speech samples rather than reversed speech. This condition is warranted to test whether the qualitative attributes available in reversed speech samples match those of natural speech. The study also measures sinewave generalization to test the reliability of the sinewave results obtained in the previous experiments.

Method

Listeners—Ten adults were recruited from the Bloomington community and were paid $5 per hour for their participation. Two were excused because of extremely slow progress during training. Eight completed the experiment.

Test Materials and Procedure—The training materials were reversed speech samples, and the generalization test materials were natural speech and sinewave replicas. Other aspects of the training, familiarization, and testing procedures were consistent with the previous experiments.

Results

Training Performance—The training data from this experiment are almost indistinguishable from the training data obtained in Experiment 4. The number of days needed to reach criterion ranged from 3 to 9. A one-way ANOVA revealed a significant effect for talker identify, $F(9, 63) = 4.99$, $p < .0001$, $MSE = 273.36$, with reliable differences in identification among individual talkers but not between male and female talkers as a group.

Generalization Performance—Figure 10 displays the generalization scores for the natural and sinewave generalization tests. The data show that listeners' ability to recognize talkers decreased from 74% correct at the end of training to 50% correct for the natural test and 23% correct for the sinewave test. An ANOVA comparing the overall means from each condition revealed a significant effect, $F(2, 14) = 41.47$, $p < .0001$, $MSE = 120.54$. Recognition was significantly different from training for the natural speech trials, $t(7) = 2.43$, $p < .05$, RMSE = 6.73; for the sinewave replica trials, $t(7) = 19.03$, $p < .0001$, RMSE = 2.58; and between the two generalization tests, $t(7) = 5.28$, $p < .001$, RMSE = 6.21. Talker recognition from sinewaves exceeded chance, $t(7) = 5.39$, $p < .001$, RMSE = 4.23.

Natural speech generalization was affected by variation among the 10 talkers, $F(9, 63) = 3.29$, $p < .002$, $MSE = 0.13$, but not by variation across talker sex. The effect of talker identity was not significant in the sinewave generalization data.

Finally, individual talker identification from reversed speech training was not significantly correlated with either natural speech talker identification, $r(8) = .57$, $p < .08$, or sinewave talker identification, $r(8) = .59$, $p < .08$. This latter finding is consistent with the previous experiment, which found a weak correlation between reversed speech training and sinewave generalization.

Discussion

In Experiments 3–5, we observed that listeners who were trained to recognize talkers from natural or reversed speech samples could identify familiar talkers in either form. This suggests that when a listener learns a talker's indexical attributes from a medium rich in qualitative attributes (natural speech or reversed speech), the resulting talker-specific knowledge is sufficiently robust to permit talker recognition from novel utterances, whether natural or reversed. We also found that listeners trained on reversed speech were not very successful at identifying familiar talkers from sine-wave replicas, which is consistent with the natural speech training data from Experiment 2.

Although it is tempting to propose that listeners relied primarily on qualitative and suprasegmental attributes rather than segmental phonetic attributes (idiolect) in learning to identify natural talkers, closer analysis of Experiments 3–5 argues against this conclusion. In particular, the correlation analyses, which assessed the similarity of the attributes used to recognize talkers across the different acoustic forms, showed that attributes yielded by natural and sinewave talkers were more similar than was evidenced in contrasting natural and reversed talkers. In fact, the relative identifiability of talkers from reversed speech training only correlated with the reversed speech tests. One certainly would not expect to find a relationship between reversed speech training and sinewave speech tests, given that the former preserves qualitative attributes to the detriment of linguistic attributes, whereas the latter preserves phonetic features without natural vocal quality. However, the fact that there was only a weak correlation between reversed speech training and the natural speech tests is instructive. It indicates that although listeners attend to suprasegmental, qualitative, and phonetic attributes when they learn to identify individuals from natural samples, the resulting knowledge of a talker's attributes may include substantially more phonetic detail than heretofore suspected.

Further examination of the data from Experiments 3–5 suggests that the perceptual learning paradigm used in these experiments produced results that are consistent with naturalistic studies of talker learning. For example, in our study, overall talker identification from reversed speech was remarkably similar to the findings from Bartholomeus (1973), who found that talker identification accuracy from reversed speech was equal to 73% of forward talker identification accuracy. In our study, talker identification from reversed speech (averaged across Experiments 3 and 5) was 75% of forward talker identification. Our correlation analysis also yielded results that are very similar to results from Van Lancker et al. (1985), who reported that the correlation between familiar talker recognition from natural speech and reversed speech was .55, which is similar to the correlations obtained in our study (.71 and .57 in Experiments 3 and 5, respectively). More generally, the similarity of the data of our studies and those of Bartholomeus (1973) and Van Lancker et al. (1985) provides preliminary evidence that a laboratory-based perceptual learning procedure can approximate learning in natural settings. If this similarity is substantiated, then the controls achieved through this form of design and testing permit a direct empirical approach to these questions without sacrificing validity.

General Discussion

Traditional accounts of talker recognition propose that an individual talker is identified by virtue of qualitative characteristics that are nondistinctive linguistically (e.g., Bricker & Pruzansky, 1976; Halle, 1985). Accordingly, segmental phonetic attributes are assumed to be used only for word recognition. The present set of five experiments encourages a revision in the traditional account. Experiment 1 showed that listeners could learn to identify talkers in the absence of the qualitative attributes of vocal sound production that have been assumed, typically, to be indispensable for such ability. The experiment revealed that a listener's knowledge of talker-specific attributes acquired from sinewave sentence replicas generalized to novel sinewave and natural speech sentences. This outcome was consistent with the prior claim (Remez et al., 1997) that familiarity with a talker's indexical attributes includes the phonetic segmental properties composing an individual's idiolect and, therefore, that idiolectal variation can drive both the learning and the recognition of specific individuals. Experiment 2 calibrated the relative ease with which listeners learned to identify talkers in the laboratory from samples of natural speech. However, the form of this knowledge was limited and did not generalize to novel sinewave utterances. This was true even when listeners acclimated to the peculiar auditory qualities of sinewave replicas. Using a test reciprocal to the method of sinewave replication, Experiments 3, 4, and 5 showed that listeners can also recognize talkers largely on the basis of qualitative attributes, from incomprehensible reversed speech samples. The outcomes of the reversed speech experiments indicate that vocal quality and pitch range (both well preserved in these signals) can be used by listeners to identify talkers.

Taken together, the studies (summarized in Tables 2 and 3) show that listeners can exploit a wider range of talker-specific properties than is classically assumed (Bricker & Pruzansky, 1976). Under some circumstances, perception of a talker depends more on global qualitative attributes, whereas in other circumstances, phonetic attributes alone can specify a talker. Evidently, there is no single set of features or perceptual processes that can be used to identify both words and talkers. These results have several implications for a general account of the recognition of a familiar talker.

Variety

Two factors appear to govern the kind of indexical attributes that a listener uses when becoming familiar with an individual talker. The first factor is the variety of indexical properties available in a speech sample. Natural speech expresses a great assortment of linguistic, personal, and social attributes, and the presence of multiple acoustic correlates of each attribute in the speech signal contributes both to the robustness of linguistic perception and to talker recognition. The overall performance levels that we observed during training and generalization reflect the relative richness of the supply of indexical properties.

Our data show that talker learning from natural speech was relatively rapid, and the attributes that became familiar during the training procedure generalized very well to novel instances of natural speech. Reversed speech preserves a portion of this assortment of indexical properties, because temporal reversal eliminates some of the segmental and suprasegmental phonetic attributes that otherwise are useful to index a talker. Subjects found it somewhat more difficult to learn a talker's characteristics from reversed speech and were not always able to recognize a known talker from a novel instance of reversed speech. Sinewave replicas offer the least abundant assortment of indexical attributes of the three acoustic signals, preserving only a subset of the features available in natural samples. Learning about a talker from sinewave sentences was slow and difficult for our listeners, and the knowledge acquired during training generalized only moderately well to new sinewave sentences. Apparently, listeners who had become familiar with the talkers from a medium

that preserved many natural properties were able to recognize the talkers from a new set of sentences better than listeners who had become familiar with the talkers from a medium with a less ample supply. Our data indicate that increasing the variation and richness of indexical properties during training produces highly abstract talker knowledge that generalizes readily to new contexts.

Exclusive Attributes

The nature or type of indexical attribute present in each acoustic form of speech counts as a second constraint on the features that are learned when a listener becomes familiar with a talker. For example, talker-specific attributes encountered in sinewave replicas rest primarily on segmental phonetic properties and their aggregation in speech meter, whereas those present in reversed speech are primarily qualitative, exclusive of much of the fine-grained phonetic properties on which idiolectal characterizations are based. Natural samples contain veridical manifestations of qualitative, segmental phonetic and suprasegmental properties, although the extent to which listeners make use of segmental phonetic attributes in remembering and perceiving individuals has so far been unclear.

To appraise the kind of properties that listeners use in identifying individuals, we compared the similarity in the relative ranking of talkers by the identification performance that we observed across different acoustic forms and across these five experiments. With this approach, we sought to reveal the extent to which the particular remembered attributes of an individual talker transfer to a different type of signal, allowing the familiar talker either to be recognized or not. These analyses differ from the analyses of overall talker recognition accuracy because here we focused on the relative identifiability of each individual talker from sinewave, natural, and reversed speech tokens, independent of the average level of recognition across all talkers. For example, if similar indexical features are used to identify talkers from sinewave and natural speech, we should find that the differences in identification accuracy among talkers in the context of one signal are correlated with the accuracy differences obtained in the context of another signal (see Table 3).

The analysis shows that the rank order of talkers in the sinewave training condition was significantly correlated with each of the natural speech training conditions: Experiment 2, $r(8) = .71$, $p < .02$; Experiment 3, $r(8) = .84$, $p < .002$. This indicates that the relative ease of identifying the talkers from natural samples closely matched the relative ease of identifying talkers from sinewave replicas. The high correlations provide evidence that the perceived properties of familiar sinewave and natural speech talkers are similar, and they suggest that listeners were relying on phonetic attributes during both natural and sinewave talker learning. In contrast, there was little similarity between the perceptual spaces of familiar sinewave and reversed speech talkers, as evidenced by the weak and nonsignificant correlations between sinewave speech training and reversed speech training: Experiment 4, $r(8) = .31$, $p < .39$; Experiment 5, $r(8) = .40$, $p < .25$. This is not wholly unexpected, given that reversed speech and sinewave speech preserve different assortments of phonetic details, suprasegmental meter and intonation, and natural vocal quality.

The correlations between the mean identifiability of reversed speech and natural speech samples provide further evidence that listeners rely at least partly on segmental phonetic properties when learning to identify talkers from natural speech. Specifically, we found that the correlations between performance with reversed and natural samples were actually lower than the correlations between sinewave and natural samples. This is surprising, given that reversed speech preserves many of the physical acoustic properties associated with vocal quality, which have been favored in prior discussions of individual talker recognition (Bricker & Pruzansky, 1976; Van Lancker et al., 1985). Moreover, overall accuracy data in this report show better generalization between natural and reversed speech than between

natural and sinewave speech. Extrapolating from precedent, one would plausibly expect listeners to treat reversed and natural signals similarly and to treat sinewave and natural signals as altogether different.

Instead we found the opposite pattern: There was less similarity in the relative identifiability of talkers between reversed and natural speech. The correlations between reversed and natural speech ranged from .63 to .71, lower than those found in comparing natural speech and sinewave signals (range = .71 to .84). The reversed speech training data from Experiment 4 were significantly correlated with each of the natural speech training conditions: Experiment 2, $r(8) = .71$, $p < .02$; Experiment 3, $r(8) = .63$, $p < .05$. The reversed speech training data from Experiment 5 were significantly correlated with one of the two natural speech training conditions: Experiment 2, $r(8) = .64$, $p < .05$; Experiment 3, $r(8) = .58$, $p < .08$. The attributes excluded by reversal of speech are arguably responsible for establishing the order of identifiability of the talkers. In summary, although the correlations between reversed and natural speech are lower and less reliable than the correlations between sinewave and natural speech, the values indicate a solid relationship between the two signals. This suggests that reversed speech and sinewave replicas each preserve attributes of a natural utterance that can be used by listeners to recognize a familiar talker.

Accuracy and Relative Identifiability

If sinewave speech preserves attributes of natural speech, why did our listeners have such difficulty recognizing sinewave talkers following natural speech training? Perhaps the low accuracy levels on the sinewave tests (as well as the moderate accuracy on the reversed speech tests) are explained by the hypothesis that generalization performance is influenced by the perceptual match between the current generalization test token and previously encountered instances of the individual's speech. We assume that memory for a presented talker includes a number of different kinds of features. These include indexical and nonindexical perceptual properties of a talker's gestures and nonindexical features of the acoustic signal. During generalization testing, memory is probed by a new test token containing features of the talker and context. Recognition of a talker depends on the match between memory and the test item; the better the match, the greater likelihood the talker is perceived as familiar and identified by name. Generalization to a different signal type reduces the perceptual match between the current test token and previously encountered instances of the individual's speech. Therefore, a portion of the decrease in accuracy across training and generalization can be attributed to a simple mismatch in nonindexical features of the acoustic signal.

Generalization accuracy is further influenced by the variety and richness of indexical properties available at test. Each signal type preserves a different assortment of talker-specific segmental and suprasegmental features. Sinewave replication drastically reduces the dimensionality of the natural speech signal and to a greater extent than temporal reversal. It is perhaps not surprising that generalization was lowest on the sinewave test following natural speech training, given that only a limited subset of the indexical features available during training were present at the time of testing.

Independence of Talker Identity and Talker Sex

Although the accuracy levels on the sinewave tests were relatively low, performance surpassed chance, indicating a significant contribution of the attributes common to natural and sinewave versions of speech. The results, moreover, are incompatible with an obvious strategic gambit. For example, consider the hypothetical performance if subjects had classified each unidentified sine-wave talker by sex, and only then chosen randomly within sex. This strategy of classification first by sex and then by personal identity has been

proposed for the identification of individuals from visual perception of the face (e.g., Ellis, 1986). Although we cannot rule out this possibility entirely in our findings, several lines of converging evidence argue against this conjecture.

Evidence that the acoustic properties indicating the sex of a talker make a limited contribution to sinewave talker identification was confirmed through a series of hierarchical cluster analyses based on identification errors from Remez et al. (1997; see also Fellowes, Remez, & Rubin, 1997, Figure 1). The resulting topologies indicated that perceived similarity among talkers did not amount to a sort by sex, because male and female sinewave talkers were frequently misidentified for one another. These findings agree with the data in the present report, which showed that our listeners made many cross-sex errors when generalizing from natural to sinewave utterances (see Experiment 2). When identifying a talker from a sinewave utterance, listeners apparently disregarded auditory characteristics of the signal, which could conceivably be useful for sex identification (e.g., the central spectral tendency of formant variation). Instead, subjects arguably mistook talkers for each other when they shared specific pronunciation habits. These results are consistent with the premise that listeners were exploiting consistent segmental phonetic differences among the talkers to identify them.

The extent to which perception of talker identity is contingent on perception of talker sex was addressed directly by Fellowes et al. (1997) using perceptual tests and different types of sinewave signals. The first set of sinewave tokens were the same sinusoidal replicas used in the current training experiments. These tokens faithfully replicate the formant estimates of each talker's natural speech. Fellowes et al. found that listeners could report the sex of the talker from sinewaves, though performance was not free of errors. Overall, this result showed that the sinewave replicas contain sufficient structure to distinguish male from female talkers much of the time.

What do listeners use to determine the sex of a sinewave talker? Is the perception of a talker's sex a prerequisite for individual identification? Fellowes et al. (1997) addressed these questions using a second set of modified sex-neutral sinewave signals. Intertalker differences were completely eliminated from the tokens by rescaling the tone pattern of every talker to match the average formant values of all ten talkers. Listeners found it impossible to determine the sex of a sinewave talker from these signal. This indicated that differences in the central spectral tendency of each tonal analog of formant frequencies can be used to distinguish the sex of a sinewave talker. The subjects were still able to recognize the identity of 6 of the 10 talkers from sex-neutral sinewaves. This latter finding demonstrates that segmental phonetic attributes preserved in sinewaves can be used for individual talker identification even in the absence of acoustic structure that indexes sex.

More generally, talker identification performance is largely unaffected by features of talker sex, even though listeners are capable of assessing talker sex when asked to do so. It is intriguing that the analogous sex-conditional account of visual identification of individuals has received little empirical support (Bruce & Young, 1986). Perhaps the circumstances of identifying a talker represent a stable aspect of sensing, discerning, and remembering individuals regardless of modality.

Interactions Between Indexical and Linguistic Processing Across Modalities

The hypothesis that idiolect, the subtle characteristics of a talker's articulation, can be used to identify individuals predicts interactions between word perception and talker recognition. Indeed, many examples of such interactions can now be found in the technical literature. For example, the perception of talker and phonetic attributes occur in an integral or parallel-contingent fashion (Green et al., 1997; Mullennix & Pisoni, 1990). Numerous studies have

shown that changes in a talker affect phoneme and word identification (Cole et al., 1974; Martin et al., 1989) and memory for spoken words (Craik & Kirsner, 1974; Goldinger, 1996; Schacter & Church, 1992). In addition, the ease with which a talker is identified directly affects intelligibility, with words spoken by familiar talkers being easier to perceive (Nygaard et al., 1994) and to retrieve from long-term memory (Sheffert & Olson, 2000) relative to words spoken by unknown talkers.

Analogous dependencies between the perception of phonetic and talker attributes have also been found in the visual domain. For instance, trial-to-trial variation in the talker reduces lip-reading accuracy (Yakel, Rosenblum, & Fourtier, 2000). Other research shows that remembered attributes of a visual talker interact with lexical and indexical memory (Sheffert & Fowler, 1995). Familiarity with a talker's face has also been shown to improve the accuracy of visual phoneme identification (Schweinberger & Soukup, 1998), of short-term memory for audiovisual words, and of audiovisual speech integration (Walker, Bruce, & O'Malley, 1995).

Recent experiments using point-light displays of a talker's articulatory gestures reveal additional evidence that visible talker properties interact with the visual perception of the linguistic message. Dynamic visual articulation can be isolated from other aspects of the facial topography by placing florescent landmarks on the cheeks, lips, tongue, and teeth. Under the appropriate illumination, viewers see only a pattern of moving dots when the talker articulates. Previous studies have shown that point-light displays of a talker's articulating face convey enough of the phonetic segmental grain to distinguish many phonemes and words and that this sensory stream is readily integrated with auditory speech (see Rosenblum & Saldaña, 1998, for a review). Moreover, these displays also permit the recognition of a familiar talker's face in silence (Rosenblum, Yakel, Baseer, & Panchal, 2002). Rosenblum et al. have argued that subjects were able to recognize familiar faces by using idiolectal properties conveyed visually. They further suggested that contingencies between lip-reading and face recognition might best be explained by assuming that linguistic and indexical properties are transmitted in parallel using phonetic gestures.

Neuroimaging data substantiate the similarities in perceptual effect of visual and auditory samples of phonetic gestures. Using functional magnetic resonance imaging, Campbell (1998) found that silent lip-reading activates areas of the auditory cortex that have classically been associated with auditory speech processing, although neither static poses nor nonspeech facial motion activates language areas.

These findings are consistent with the more general proposal that listeners notice and remember aspects of speech that vary across talkers and are available in multiple sensory modalities. Linguistic and indexical properties originate in the same event, the articulatory movement of a vocal tract. Conceivably, identifying words and talkers could be based on a general capacity to discriminate the subtleties of phonetic expression.

In summary, when a listener learns to identify a new talker, the familiar details include segmental phonetic properties of the talker's speech. The overall performance levels in our report underestimate the contribution of phonetic attributes to identifying and recognizing a familiar talker. The fact that acoustic conditions favoring segmental or qualitative attributes produced different outcomes shows that these indexical properties, which differ in grain from fine to coarse, may be registered in parallel, because listeners did not use them in a contingent fashion. Taken together, these findings indicate that phonetic properties of speech can play a role in perceiving and remembering the properties of a talker no less than qualitative properties of the message itself.

Acknowledgments

This research was supported by National Institute on Deafness and Other Communicative Disorders Grant DC00111 to Indiana University Bloomington and Grant DC00308 to Barnard College.

We thank Luis Hernandez for technical support, Nathan Large for assisting with data collection for Experiments 3–5, and Dalia Shoretz and Rebecca Piorkowski for lending scholarly and technical assistance to the project. We give special thanks to Vivian Tartter and two anonymous reviewers for their valuable comments on earlier versions of this article.

References

Bartholomeus B. Voice identification by nursery school children. Canadian Journal of Psychology. 1973; 27:464–472. [PubMed: 4766153]

Benton AL, Van Allen MW. Impairment in facial recognition in patients with cerebral disease. Cortex. 1968; 4:344–359.

Bodamer J. Die prosop-agnosie. (Die Agnosie des Physiogno-mieerkennens) [Prosopagnosia. (Agnosia for face detection)]. Archhiv für Psychiatrie und Zeitschrift für Nervenkrankheiten. 1947; 179:6–54.

Bricker PD, Pruzansky S. Effects of stimulus content and duration on talker identification. Journal of the Acoustical Society of America. 1966; 40:1441–1449. [PubMed: 5975580]

Bricker, PD.; Pruzansky, S. Speaker recognition. In: Lass, NJ., editor. Contemporary issues in experimental phonetics. New York: Academic Press; 1976. p. 295-326.

Broadbent DE, Ladefoged P. Vowel judgments and adaptation level. Proceedings of the Royal Society of London. 1960; 151:384–399. [PubMed: 13849177]

Bruce V, Young A. Understanding face recognition. British Journal of Psychology. 1986; 77:305–327. [PubMed: 3756376]

Bull R, Rathborn H, Clifford BR. The voice recognition accuracy of blind listeners. Perception. 1983; 12:223–226. [PubMed: 6657428]

Byrd D. Relations of sex and dialect to reduction. Speech Communication. 1994; 15:39–54.

Campbell, R. How brains see speech: The cortical localisation of speechreading in hearing people. In: Campbell, R.; Dodd, B.; Burnham, D., editors. Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech. Hove, England: Psychology Press; 1998. p. 177-193.

Carterette, EC.; Barnebey, A. Recognition memory for voices. In: Cohen, A.; Nooteboom, S., editors. Structure and process in speech perception. Heidelberg, Germany: Springer-Verlag; 1975. p. 246-265.

Chin, SB.; Pisoni, DB. Alcohol and speech. San Diego, CA: Academic Press; 1997.

Church BA, Schacter DL. Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1994; 20:521–533.

Clarke, FR.; Becker, RW.; Nixon, JC. Characteristics that determine speaker recognition (Report No. ESD-TR-66-638). Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command; 1966.

Clifford BR. Voice identification by human listeners. Law and Human Behavior. 1980; 4:373–394.

Cole RA, Coltheart M, Allard F. Memory for speaker's voice: Reaction time to same or different-voiced letters. Quarterly Journal of Experimental Psychology. 1974; 26:1–7. [PubMed: 4814860]

Coleman RO. Speaker identification in the absence of inter-subject differences in glottal source characteristics. Journal of the Acoustical Society of America. 1973; 53:1741–1743. [PubMed: 4719259]

Compton AJ. Effects of filtering and vocal duration upon the identification of speakers, aurally. Journal of the Acoustical Society of America. 1963; 35:1748–1752.

Cook S, Wilding J. Earwitness testimony: Never mind the variety, hear the length. Applied Cognitive Psychology. 1997; 11:95–111.

Craik FIM, Kirsner K. The effect of speaker's voice on word recognition. Quarterly Journal of Experimental Psychology. 1974; 26:274–284.

Creelman CD. Case of the unknown talker. Journal of the Acoustical Society of America. 1957; 29:655.

Damasio AR, Damasio H, Van Hoesen GW. Prosopagnosia: Anatomical basis and neurobehavioral mechanism. Neurology. 1982; 32:331–341. [PubMed: 7199655]

DeCasper AJ, Fifer WP. Of human bonding: Newborns prefer their mothers' voices. Science. 1980 Jun 6.208:1174–1176. [PubMed: 7375928]

Doddington GR. Speaker recognition: Identifying people by their voices. Proceedings of the IEEE. 1985; 73:1651–1664.

Ellis, HD. Processes underlying face recognition. In: Bruyer, R., editor. The neuropsychology of face perception and facial expression. Hillsdale, NJ: Erlbaum; 1986. p. 1-27.

Fellowes JM, Remez RE, Rubin PE. Perceiving the sex and identity of a talker without natural vocal timbre. Perception & Psychophysics. 1997; 59:839–849. [PubMed: 9270359]

Geiselman RE, Bellezza FS. Incidental retention of speaker's voice. Memory & Cognition. 1977; 5:658–665.

Gelfer MP. Perceptual attributes of voice: Development and use of rating scales. Journal of Voice. 1988; 2:320–326.

Goldinger SD. Words and voices: Implicit and explicit memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22:1166–1183.

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Goldinger SD, Pisoni DB, Logan JS. On the nature of talker variability effects on recall of spoken word lists. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1991; 17:152–162.

Goldinger, SD.; Pisoni, DB.; Luce, PA. Speech perception and spoken word recognition: Research and theory. In: Lass, NJ., editor. Principles of experimental phonetics. New York: Academic Press; 1996. p. 277-327.

Green KP, Tomiak GR, Kuhl PK. The encoding of rate and talker information during phonetic perception. Perception & Psychophysics. 1997; 59:675–692. [PubMed: 9259636]

Gunzburger D. Acoustic and perceptual implications of the transsexual voice. Archives of Sexual Behavior. 1995; 24:339–348. [PubMed: 7611850]

Halle, M. Speculations about the representation of words in memory. In: Fromkin, VA., editor. Phonetic linguistics: Essays in honor of Peter Ladefoged. New York: Academic Press; 1985. p. 101-114.

Hecker M. Speaker recognition: An interpretive survey of the literature. ASHA Monographs. 1971; 16:1–103. [PubMed: 4943814]

Hepper P, Scott D, Shahidullah S. Newborn and fetal response to maternal voice. Journal of Reproductive and Infant Psychology. 1993; 11:147–153.

Joos MA. Acoustic phonetics. Language. 1948; 24(Suppl 2):1–136.

Jusczyk PW, Hohne EA, Jusczyk AM, Redanz NJ. Do infants remember voices? Journal of the Acoustical Society of America. 1993; 94:2373.

Klatt, DH. Review of selected models of speech perception. In: Marslen-Wilson, W., editor. Lexical representation and process. Cambridge, MA: MIT Press; 1989. p. 169-226.

Kreiman, J. Listening to voices: Theory and practice in voice perception research. In: Johnson, K.; Mullenix, J., editors. Talker variability in speech processing. San Diego, CA: Academic Press; 1997. p. 85-108.

Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. Journal of Speech and Hearing Research. 1993; 36:21–40. [PubMed: 8450660]

Labov, W. Sources of inherent variation in the speech process. In: Perkell, JS.; Klatt, DH., editors. Invariance and variability in speech processes. Hillsdale, NJ: Erlbaum; 1986. p. 402-425.

Ladefoged P, Broadbent DE. Information conveyed by vowels. Journal of the Acoustical Society of America. 1957; 29:98–104.

Ladefoged P, Ladefoged J. The ability of listeners to identify voices. UCLA Working Papers in Phonetics. 1980; 49:43–51.

Lass NJ, Phillips JK, Bruchey CA. The effect of filtered speech on speaker height and weight identification. Journal of Phonetics. 1980; 8:91–100.

Laver, J. The phonetic description of voice quality. Cambridge, England: Cambridge University Press; 1980.

Laver, J. The gift of speech. Edinburgh, Scotland: Edinburgh University Press; 1991.

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

Legge GE, Grosmann C, Pieper CM. Learning unfamiliar voices. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1984; 10:298–303.

Luce PA, Lyons EA. Specificity of memory representations for spoken words. Memory & Cognition. 1998; 26:708–720.

Mandel DR, Jusczyk PW, Pisoni DB. Infants' recognition of the sound patterns of their own names. Psychological Science. 1995; 6:314–317.

Mann VA, Diamond R, Carey S. Development of voice recognition: Parallels with face recognition. Journal of Experimental Child Psychology. 1979; 27:153–165. [PubMed: 458368]

Martin CS, Mullennix JW, Pisoni DB, Summers WV. Effects of talker variability on recall of spoken word lists. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1989; 17:152–162.

Matsumoto H, Hiki S, Sone T, Nimura T. Multidimensional representation of personal quality and its acoustic correlates. IEEE Transactions on Audio and Electroacoustics. 1973; 21:428–436.

McGehee F. The reliability of the identification of the human voice. Journal of General Psychology. 1937; 17:249–271.

Meyer-Eppler W. Realization of prosodic features of whispered speech. Journal of the Acoustical Society of America. 1957; 29:104–106.

Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics. 1990; 47:379–390. [PubMed: 2345691]

Nearey TM. Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America. 1989; 85:2088–2113. [PubMed: 2659638]

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Perception & Psychophysics. 1998; 60:355–376. [PubMed: 9599989]

Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. Psychological Science. 1994; 5:42–46. [PubMed: 21526138]

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1993; 19:1–20.

Papçun G, Kreiman J, Davis A. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America. 1989; 85:913–925. [PubMed: 2926007]

Peters, RW. The effect of length of exposure to speaker's voice upon listener reception (Joint Project Report No. 44). Pensacola, FL: U.S. Naval School of Aviation Medicine; 1955. p. 1-8.

Peterson GE, Barney HE. Control methods used in the study of the vowels. Journal of the Acoustical Society of America. 1952; 24:175–184.

Pisoni, DB. Some thoughts on "normalization" in speech perception. In: Johnson, K.; Mullennix, JW., editors. Talker variability in speech processing. San Diego, CA: Academic Press; 1997. p. 9-32.

Pollack I, Pickett JM, Sumby WH. On the identification of speakers by voice. Journal of the Acoustical Society of America. 1954; 26:403–406.

Rand TC. Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research. 1971; SR25/26:141–146.

Read D, Craik FJM. Earwitness identification: Some influences on voice recognition. Journal of Experimental Psychology: Applied. 1995; 1:6–18.

Remez RE, Fellowes JM, Rubin PE. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance. 1997; 23:651–666. [PubMed: 9180039]

Remez RE, Rubin PE, Nygaard LC, Howell WA. Perceptual normalization of vowels produced by sinusoidal voices. Journal of Experimental Psychology: Human Perception and Performance. 1987; 13:40–61. [PubMed: 2951488]

Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science. 1981 May 22.212:947–950. [PubMed: 7233191]

Rosenblum, LD.; Saldaña, HM. Time-varying information for speech perception. In: Campbell, R.; Dodd, B.; Burnham, D., editors. Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech. Hove, England: Psychology Press; 1998. p. 61-81.

Rosenblum LD, Yakel DA, Baseer N, Panchal A. Visual speech information for face recognition. Perception & Psychophysics. 2002; 64:220–229. [PubMed: 12013377]

Rubin, PE. Sinewave synthesis [Internal memorandum]. New Haven, CT: Haskins Laboratories; 1980.

Schacter DL, Church BA. Auditory priming: Implicit and explicit memory for words and voices. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1992; 18:915–930.

Scherer KR. Vocal affect expression: A review and a model for future research. Psychological Bulletin. 1986; 99:143–165. [PubMed: 3515381]

Schmidt-Nielsen A, Stern K. Identification of known voices as a function of familiarity and narrow-band coding. Journal of the Acoustical Society of America. 1985; 77:658–663.

Schweinberger S, Herholz A, Sommer W. Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. Journal of Speech, Language, and Hearing Research. 1997; 40:453–463.

Schweinberger S, Soukup G. Asymmetric relationships among perceptions of facial identify, emotion, and facial speech. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1998; 24:1748–1765.

Sheffert SM. Contributions of surface and conceptual information to recognition memory. Perception & Psychophysics. 1998a; 60:1141–1152. [PubMed: 9821776]

Sheffert SM. Format-specificity effects on auditory word priming. Memory & Cognition. 1998b; 26:591–598.

Sheffert SM, Fowler CA. The effects of voice and visible speaker change on memory for spoken words. Journal of Memory and Language. 1995; 34:665–685.

Sheffert, SM.; Olson, E. Audiovisual talker familiarity and long-term memory for spoken words. Paper presented at the 41st Annual Psychonomic Society of America Meeting; New Orleans, LA. 2000 Nov.

Silverman K. F0 segmental cues depend on intonation: The case of the rise after voiced stops. Phonetica. 1986; 43:76–92.

Spencer LE. Speech characteristics of male-to-female transsexuals: A perceptual and acoustic study. Folia Phoniatica. 1988; 40:31–42.

Studdert-Kennedy, M. The perception of speech. In: Sebeok, TA., editor. Current trends in linguistics. The Hague, the Netherlands: Mouton; 1974. p. 2349-2385.

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

Summerfield, Q.; Haggard, MP. Report of speech research in progress. Vol. 2. Belfast, Ireland: Queens University of Belfast; 1973. Vocal tract normalization as demonstrated by reaction times; p. 12-23.

Syntrillium Software Corporation. Cool Edit 96. Phoenix, AZ: Author; 1996.

Tartter VC. Happy talk: Perceptual and acoustic effects of smiling on speech. Perception & Psychophysics. 1980; 27:24–27. [PubMed: 7367197]

Tartter VC. Identifiability of vowels and speakers from whispered syllables. Perception & Psychophysics. 1991; 49:365–372. [PubMed: 2030934]

Tartter VC, Braun D. Hearing smiles and frowns in normal and whisper registers. Journal of the Acoustical Society of America. 1994; 96:2101–2107. [PubMed: 7963024]

Thompson CP. Voice identification: Speaker identifiability and a correction of the record regarding sex effects. Human Learning. 1985; 4:19–27.

Trudgill, P. Sociolinguistics: An introduction. Harmondsworth, England: Penguin; 1974.

Van Lancker D, Canter GJ. Impairment of voice and face recognition in patients with hemispheric damage. Brain and Cognition. 1982; 1:185–195. [PubMed: 6927560]

Van Lancker D, Cummings J, Kreiman J, Dobkin BH. Phonoagnosia: A dissociation between familiar and unfamiliar voices. Cortex. 1988; 9:195–209. [PubMed: 3416603]

Van Lancker D, Kreiman J. Voice discrimination and recognition are separate abilities. Neuropsychologia. 1987; 25:829–834. [PubMed: 3431677]

Van Lancker D, Kreiman J, Cummings J. Voice perception deficits: Neuroanatomical correlates of phonoagnosia. Journal of Clinical and Experimental Neuropsychology. 1989; 11:665–674. [PubMed: 2808656]

Van Lancker D, Kreiman J, Emmorey K. Familiar voice recognition: Patterns and parameters. Part I: Recognition of backwards voices. Journal of Phonetics. 1985; 13:19–38.

Van Orden G, Pennington B, Stone G. What do double dissociations prove? Cognitive Science. 2001; 25:111–172.

Van Wallendael LR, Surface A, Parsons HD, Brown M. “Earwitness” voice recognition: Factors affecting accuracy and impact on jurors. Applied Cognitive Psychology. 1994; 8:661–673.

Verbrugge RR, Strange W, Shankweiler DP, Edman JR. What information enables a listener to map a talker’s vowel space? Journal of the Acoustical Society of America. 1976; 79:1086–1100.

Voiers WD. Perceptual bases of speaker identity. Journal of the Acoustical Society of America. 1964; 36:1065–1073.

Walden B, Montgomery A, Gibeily G, Prosek R. Correlates of psychological dimensions of talker similarity. Journal of Speech and Hearing Research. 1978; 21:265–275. [PubMed: 703276]

Walker S, Bruce V, O’Malley C. Facial identify and facial speech processing familiar faces and the McGurk effect. Perception & Psychophysics. 1995; 24:38–45.

Walton JH, Orlikoff RF. Speaker race identification from acoustic cues in the vocal signal. Journal of Speech and Hearing Research. 1994; 37:738–745. [PubMed: 7967558]

Yakel DA, Rosenblum LD, Fourtier MA. Effects of talker variability on speechreading. Perception & Psychophysics. 2000; 62:1405–1412. [PubMed: 11143452]

Yarmey AD. Verbal, visual, and voice identification of rape suspect under different levels of illumination. Journal of Applied Psychology. 1986; 71:363–370. [PubMed: 3745075]

Appendix Linguistic Test Materials Used in Experiments 1–5

1. A termite looks like an ant.
2. Tighten the belt by a notch.
3. Break the dry bread into crumbs.
4. The bride wore a white gown.
5. Lubricate the can with grease.
6. My jaw aches when I chew gum.
7. The kitten climbed out on a limb.
8. The scarves were made of shiny silk.
9. The drowning man let out a yell.

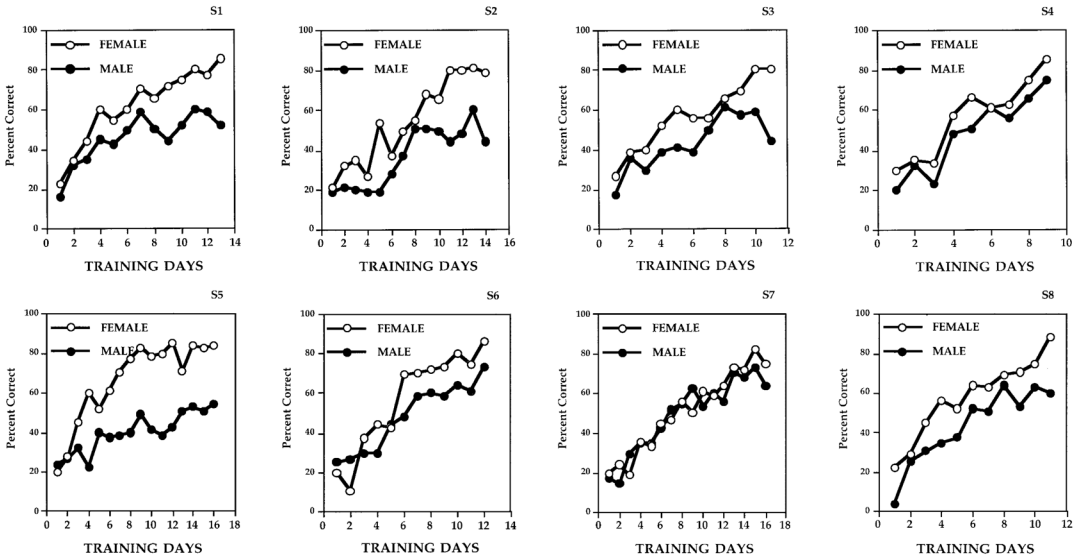

Figure 1.
Mean talker identification performance on the sinewave training for Subjects 1–8 (S1–S8) as a function of training days and talker sex in Experiment 1.

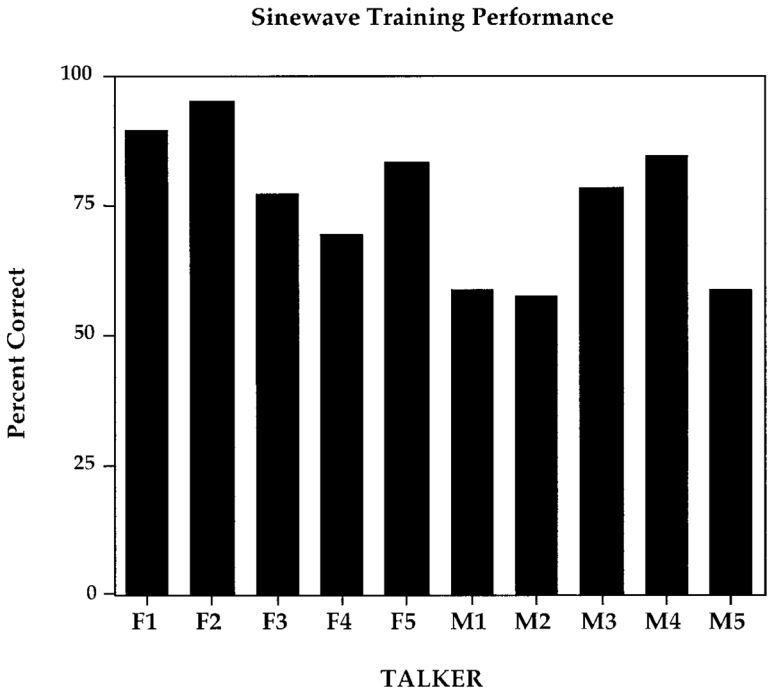

Figure 2.
Talker identification performance on sinewave replicas for the last day of training in Experiment 1. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

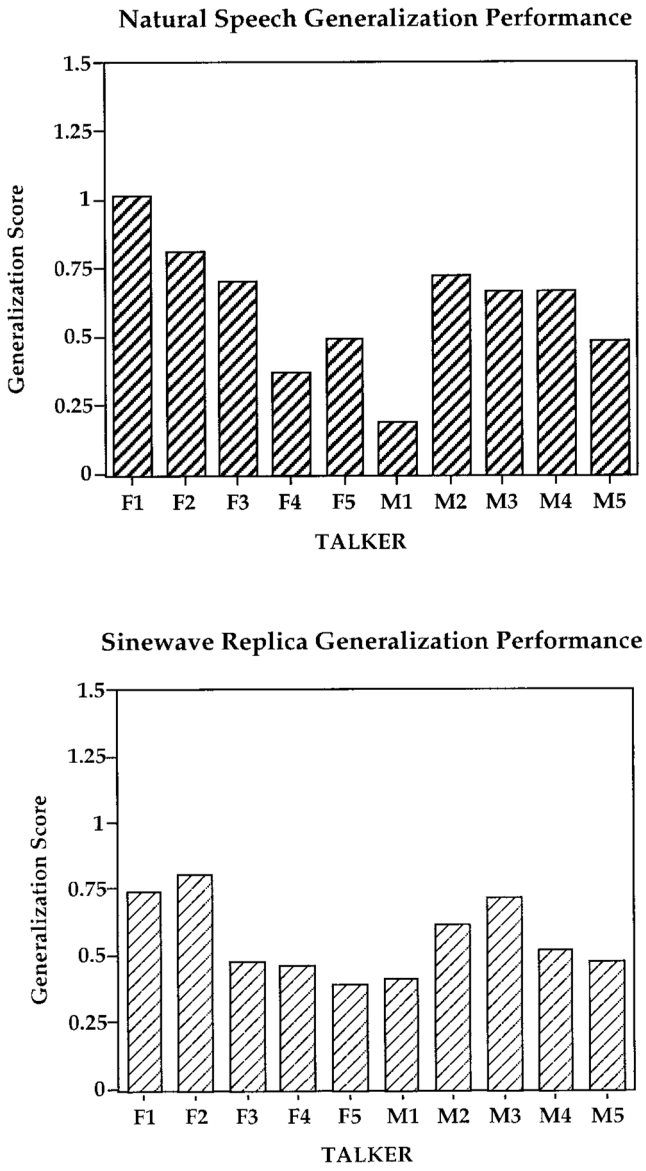

Figure 3.
Mean talker identification performance on the natural speech generalization test (top) and the sinewave replica generalization test (bottom) in Experiment 1. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

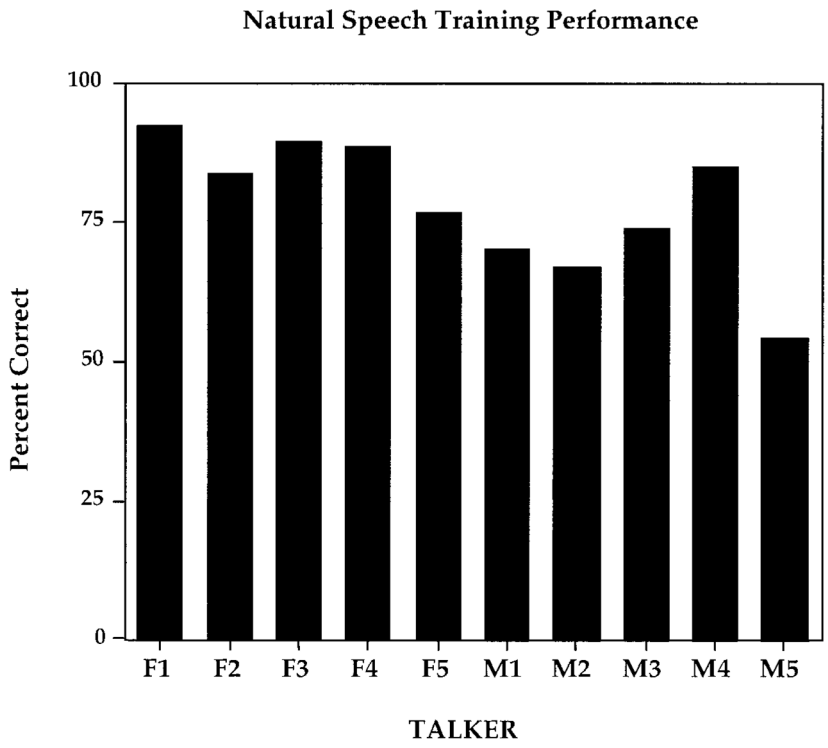

Figure 4.
Speaker identification performance on the natural speech sentences for the last day of training in Experiment 2. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

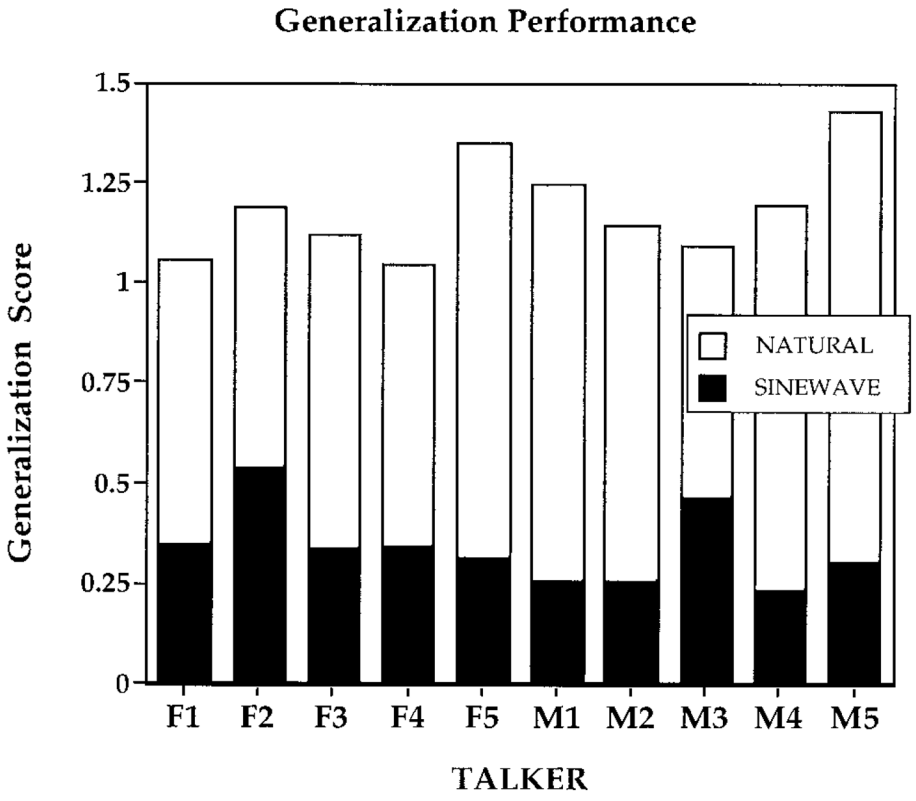

Figure 5.
Mean talker identification performance on the natural speech generalization test and the sinewave replica generalization test in Experiment 2. Performance is displayed using a value-added stacked bar graph. Performance on the natural speech test is represented by the height of the entire bar; performance on the sinewave replica test is represented by the dark section of each bar. Identification performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

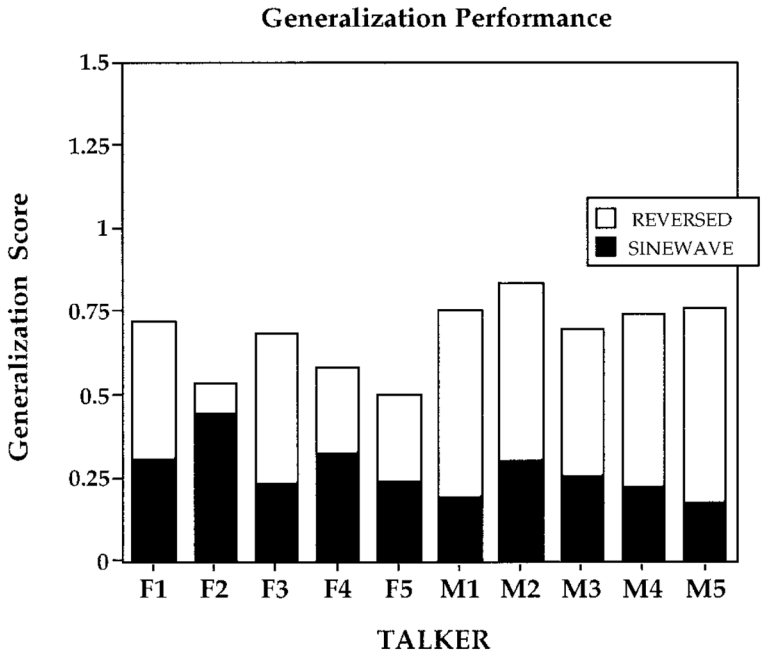

Figure 6.
Mean generalization scores on the reversed speech generalization test and the sinewave replica speech generalization test in Experiment 3. Performance is displayed using a value-added stacked bar graph. Performance on the reversed speech test is represented by the height of the entire bar; performance on the sinewave replica test is represented by the dark section of each bar. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

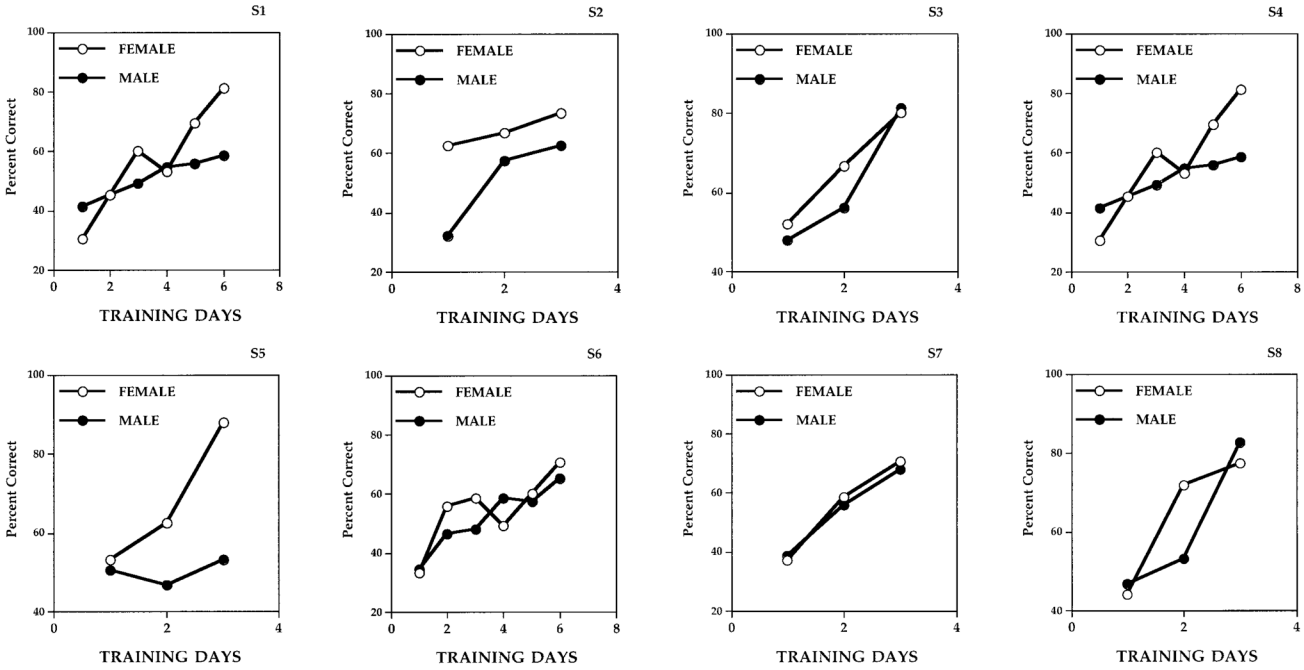

Figure 7.
Mean talker identification performance on the reversed speech training for Subjects 1–8 (S1–S8) as a function of training days and talker sex in Experiment 4.

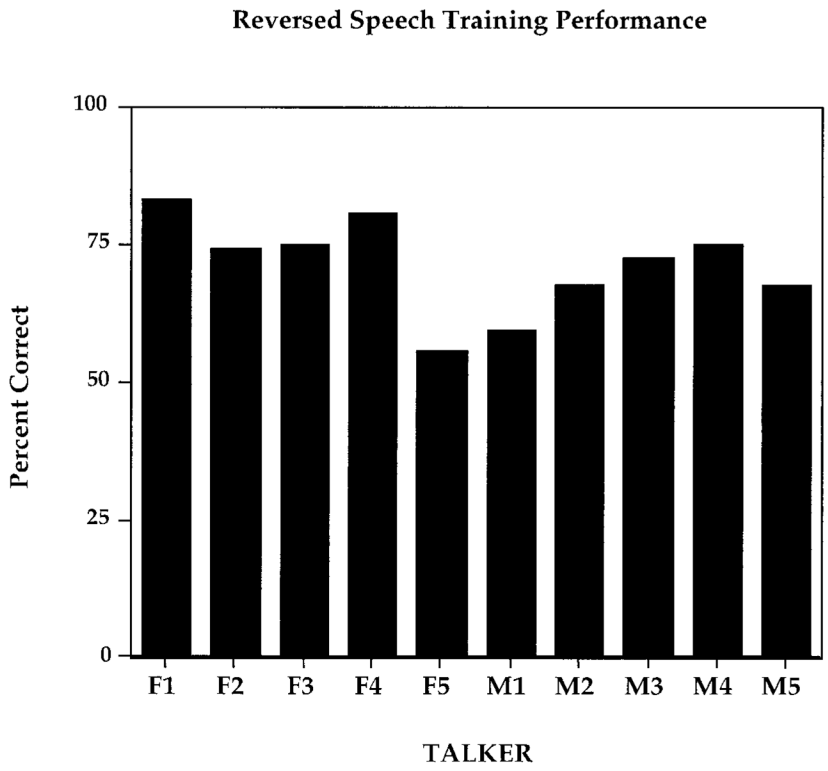

Figure 8.
Mean talker identification performance on reversed speech sentences for the last day of training in Experiment 4. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

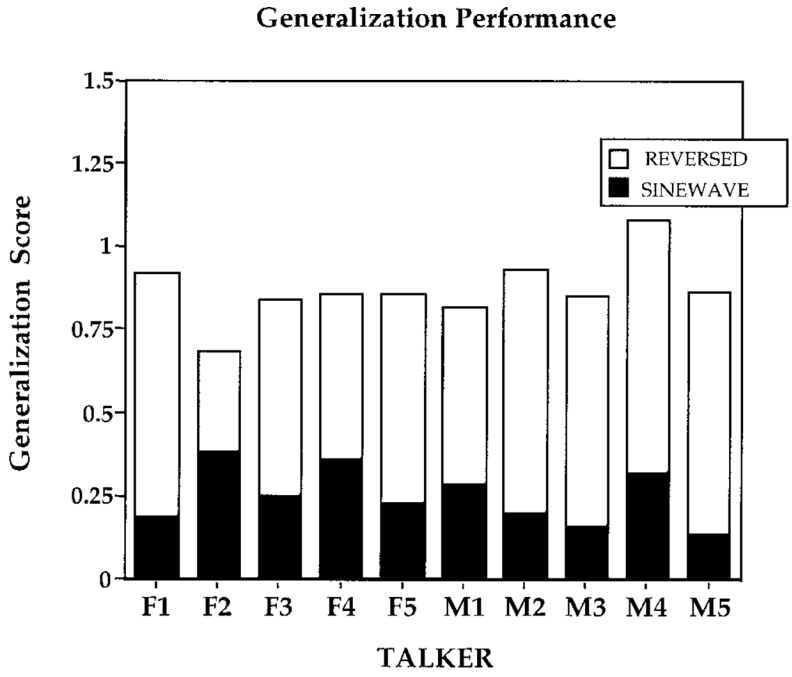

Figure 9.

Mean talker identification performance on the reversed speech generalization test and the sinewave replica generalization test in Experiment 4. Performance is displayed using a value-added stacked bar graph. Performance on the reversed speech test is represented by the height of the entire bar; performance on the sinewave replica test is represented by the dark section of each bar. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

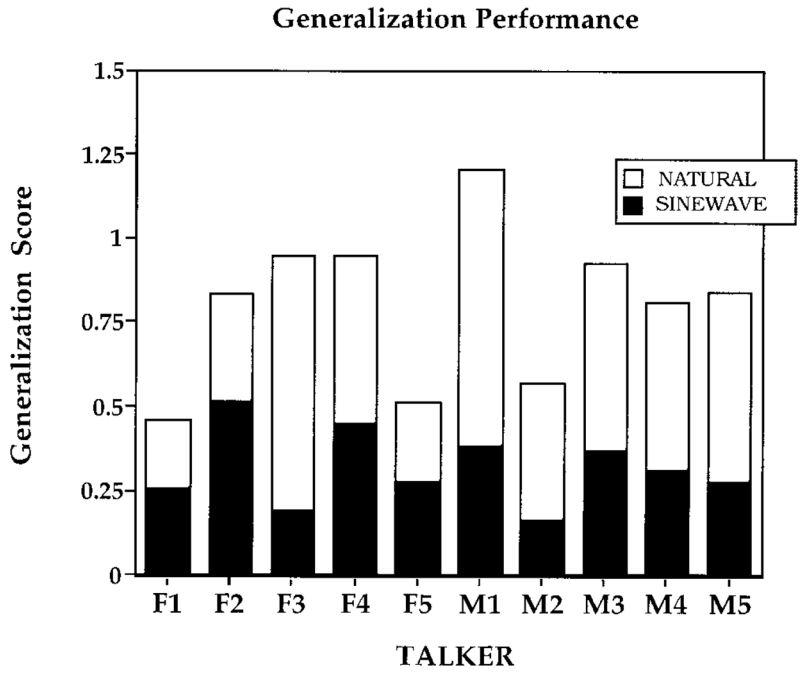

Figure 10.

Mean talker identification performance on the natural speech generalization test and the sinewave replica generalization test in Experiment 5. Performance is displayed using a value-added stacked bar graph. Performance on the natural speech test is represented by the height of the entire bar; performance on the sinewave replica test is represented by the dark section of each bar. Performance is displayed as a function of talker. F1 through F5 refer to the female talkers; M1 through M5 refer to the male talkers.

Table 1

Summary of Talker Training and Generalization Test Conditions in Experiments 1–5

	Condition	
Experiment	**Training**	**Generalization**
1	Sinewave speech	Sinewave, natural
2	Natural speech	Sinewave, natural
3	Natural speech	Sinewave, reversed
4	Reversed speech	Sinewave, reversed
5	Reversed speech	Sinewave, natural

Table 2

Median Number of Training Days and Mean Proportion Correct for Talker Recognition in Experiments 1–5

			Generalization test performance[a]		
Experiment	**Training condition**	**Training days**	**Sinewave**	**Natural**	**Reversed**
1	Sinewave	13	.44	.46	
2	Natural	1	.27	.88	
3	Natural	2	.22		.53
4	Reversed	5	.16		.59
5	Reversed	5	.23	.56	

[a]Estimated proportion correct from guessing alone is approximately .10.

J Exp Psychol Hum Percept Perform. Author manuscript; available in PMC 2012 September 05.

Table 3

Correlations From Experiments 1–5

		Generalization test performance		
Experiment	**Training condition**	**Sinewave**	**Natural**	**Reversed**
1	Sinewave	.81*	.85	
2	Natural	.64*	.86	
3	Natural	.77*		.71
4	Reversed	.33*		.81
5	Reversed	.59*	.57	

*$p \leq .05$.