Nusbaum, H., & Morin, T. (1989, May). *Perceptual normalization of talker differences.* Paper presented at the 117th meeting of the Acoustical Society of America, Syracuse, NY.

Nusbaum, H., & Morin, T. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: IOS Press.

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

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1–20.

Peters, R. W. (1955). *The relative intelligibility of single-voice and multiple-voice messages under various conditions of noise* (Joint Project Report No. 56, pp. 1–9). Pensacola, FL: U.S. Naval School of Aviation Medicine.

Pitt, M. A., & Samuel, A. G. (1990). Attentional allocation during speech perception: How fine is the focus? *Journal of Memory and Language, 29,* 611–632.

Rand, T. (1971). Vocal tract size normalization in the perception of stop consonants. *Haskins Laboratories Status Report on Speech Research, SR-25/26,* 141–146.

Samuel, A. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology, 18,* 452–499.

Schacter, D., & Church, B. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 915–930.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America, 60,* 213–224.

Summerfield, Q., & Haggard, M. (1973). *Vocal tract normalization as demonstrated by reaction times. Report of speech research in progress* (No. 2, pp. 1–12). Belfast, Ireland: Queen's University of Belfast.

Turvey, M. T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review, 80,* 1–52.

Verbrugge, R., Strange, W., Shankweiler, D., & Edman, T. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America, 60,* 198–212.

Weenink, D. J. M. (1986). The identification of vowel stimuli from men, women, and children. *Proceedings 10 from the Institute of Phonetic Sciences of the University of Amsterdam,* pp. 41–54.

Wood, C. C. (1974). Parallel processing of auditory and phonetic information in speech discrimination. *Perception & Psychophysics, 15,* 501–508.

Wood, C. C. (1975). A normative model for redundancy gains in speeded classification: Application to auditory and phonetic dimensions in speech discrimination. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (Vol. 1, p. 55–77). Hillsdale, NJ: Erlbaum.

5

Listening to Voices

Theory and Practice in Voice Perception Research

JODY KREIMAN

The study of voice production and perception has a long history in many disciplines (Laver, 1981). What makes two speech samples sound like the same voice? What must they have in common, and what can vary without altering personal quality? How do listeners discriminate among or recognize voices, and how accurate is "earwitness" testimony? Questions like these have long engaged researchers in disciplines ranging from performing arts to police science, from engineering to biomedicine.

Traditionally, linguists and speech perception researchers have treated speaker information as an "extra message" in the acoustic signal that must be normalized or factored away so that the semantic message may be recovered (Peters, 1954). However, recent research (e.g., Nygaard, Sommers, & Pisoni, 1994, and the references contained therein, and the other chapters in this volume) suggests that speaker-specific aspects of the signal are not "noise," but instead provide information critical to extracting the semantic message. Thus interest in voice perception has increased among speech perception researchers.

A large literature exists describing the study of voice perception, and a number of reviews have appeared (Bricker & Pruzansky, 1976; Bull & Clifford, 1984; Hecker, 1971; Hollien, 1990; Kreiman, 1987; Laver, 1981; Siegman, 1987; Yarmey, 1995). (The general term *voice perception* will be used throughout this chapter to denote any task where a listener hears a voice and makes a response, including recognition, discrimination, and evaluation tasks, as described below.)

In this chapter I review some recent work, and argue that researchers have not in fact answered any of the long-standing fundamental questions about voice perception. Furthermore, I argue that the approaches traditionally used in voice perception research will in fact never provide answers to the fundamental questions of how listeners recognize voices, discriminate among speakers, or even of how accurately listeners can identify speakers.

The point of this discussion is *not* that previous research is invalid or without interest. The focus of this chapter on voice perception also means that several research areas where substantial progress is being made, including laryngeal biomechanics (e.g., Berke & Gerratt, 1993) and voice synthesis (e.g., Childers & Lee, 1991; Granström, 1992; Karlsson, 1992; Klatt & Klatt, 1990) are excluded from discussion.[1] My goal in this chapter is rather to suggest that current approaches to voice perception research have been exhausted, and that a new approach is necessary if the questions of interest are ever to be answered.

5.1 REVIEW OF THE LITERATURE ON VOICE PERCEPTION

5.1.1 Overview of Voice Perception Tasks

Following the classic speech chain model (Denes & Pinson, 1993), voice perception is traditionally viewed as a succession of stages that transmit information from a speaker to a listener (Bricker & Pruzansky, 1976). Perceived vocal quality reflects the contribution of all the stages in Figure 1.

Within a given study, differences among speakers are generally assumed to contribute the majority of variance in voice perception tasks. That is, when recognition accuracy varies from target to target, the differences are assumed to be due mostly to differences in the similarity of the target and test voices. Similarly, differences in ratings in evaluative tasks are assumed to be due largely to differences in the extent to which the voices being rated possess the quality being rated (Kreiman & Gerratt, 1996). Differences among speakers (*interspeaker* variability in quality) may be due to innate factors (anatomy, age, gender, and so on) and/or to learned (dialectal and idiolectal) aspects of speech (Garvin & Ladefoged, 1963; Ladefoged, 1956). Speakers also differ from occasion to occasion in the precise way they produce utterances (*intraspeaker* variability in quality). Thus the task for listeners in discrimination studies is traditionally described as determining whether observed differences in quality are due to within- or between-speaker factors. In recognition tasks, listeners compare a stored representation to a newly presented probe voice and make a similar decision.

[1] Most studies of pathological voice quality, machine recognition, recognition via spectrogram, and voice production have also been excluded. For reviews of these areas, see, for example, Bricker and Pruzansky (1976), Hollien (1990), Kreiman, Gerratt, Kempster, Erman, and Berke (1993), and Titze (1994).

FIGURE 1 Stages of transmitting information from a speaker to a listener.

The amount and kind of information available about who is speaking also depends on the utterance produced. For example, limited information about articulation is available from steady state vowels, and whisper does not provide information about the vocal source. The amount and kind of information available to listeners may also be affected by filtering, masking, or other characteristics of the transmission system used to present signals. Listeners also contribute to voice perception through their experience with different classes of voice, perceptual habits, attention to the task, and so on (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993).

Finally, the output of the voice perception process is a decision, the precise form of which depends on the listener's task. Two broad classes of task are generally distinguished: voice evaluation tasks, and voice identification (discrimination or recognition) tasks (Bricker & Pruzansky, 1976). In evaluation tasks, stimuli are rated on some scale, or are compared with respect to some attribute. For example, listeners may estimate a speaker's age, or rate the breathiness of a voice, or compare two samples to determine which is rougher. Such tasks are particularly common in the literature on perception of pathological voice quality (see Kreiman et al., 1993, for review). Performance may be evaluated by counting the number of correct responses when "correct" is defined (for example, when age is estimated), by correlating responses with measures from some other domain (the correlation between estimates of personality and the results of a psychological inventory, or the correlation between scalar ratings and acoustic measures, for example), or by the extent to which listeners agree in their ratings. Identification tasks involve associating a voice sample with the speaker who produced it by naming the speaker, by selecting another sample (believed to have been) produced by the same speaker, or by determining whether two voice samples were produced by the same or different speakers. Performance in these tasks is usually measured by the number of correct responses.

The majority of voice perception studies have investigated the effects of changes in one or more of the cells in Figure 1 on voice ratings or recognition scores. Studies have examined the perception of speaker characteristics (including the relative contributions of the source and vocal tract to personal quality; disguise; mimicry; perception and recognition of gender, height, weight, and race from voice; and so on), effects of changes in the speech presented (sample duration, content, and different conditions of distortion), effects of changes in tasks

(for example, including a face during learning, or varying retention intervals), effects of altering the transmission system (e.g., presenting voices over the telephone), and finally, the contributions of listeners to voice perception (differences between populations of listeners in perceptual strategy or accuracy, confidence in responses, effects of experience or training, and so on). The review of these areas that appears next focuses on work that has appeared since Bricker and Pruzansky's (1976) review article (see also Hecker, 1971).

5.1.2 Studies Focusing on Speaker Characteristics

Many studies have examined the extent to which listeners can judge a speaker's personal characteristics from voice. Gender is accurately judged, although mistakes do occur (Lass, Almerino, Jordan, & Walsh, 1980; Lass, Hughes, Bowyer, Waters, & Bourne, 1976; see also Mullennix, Johnson, Topcu-Durgun, & Farnsworth, 1995). Estimates of age are generally accurate within a decade (Hartman & Danauer, 1976; Jacques & Rastatter, 1990; Linville & Fisher, 1985; Ptacek & Sander, 1966; Shipp, Qi, Huntley, & Hollien, 1992; see Helfrich, 1979, for review). Listeners apparently can also judge whether a speaker is sober or intoxicated with better than chance accuracy (Pisoni & Martin, 1989). The large literature concerning the perception of stress, mood, and emotional state from voice has been reviewed by Disner (1982), Scherer (1986; Pittam & Scherer, 1993), and Siegman (1987).

Studies of other personal characteristics have produced mixed or negative results. For example, it is unclear whether a speaker's race can be accurately evaluated from voice samples (cf. Lass, Almerino et al., 1980; Lass, Mertz, & Kimmel, 1978; Walton & Orlikoff, 1994). Judgments of a speaker's height and weight have generated particular controversy. Lass and colleagues (e.g., Lass, Barry, Reed, Walsh, & Amuso, 1979; Lass, Beverly, Nicosia, & Simpson, 1978; Lass & Davis, 1976; Lass, Phillips, & Bruchey, 1980) originally reported that listeners were accurate in their assessments of height and weight, under a variety of conditions of signal distortion. Subsequent studies from other laboratories reported negative results (Gunter & Manning, 1982; Kunzel, 1989). Most recently, Van Dommelen (1993) reanalyzed the data from Lass, Beverly, et al. (1978), and found that listeners' judgments, though wrong, were highly consistent. He concluded that significant vocal stereotypes exist for these characteristics, as they do for judgments of personality from speech (e.g., Apple, Streeter, & Krauss, 1979; Aronovitch, 1976; Hunt & Lin, 1967; Kramer, 1963; Scherer, 1979; Taylor, 1934).

Accuracy of identification varies widely by speaker, and some voices (whether familiar or unfamiliar) are more confusable than others (Bricker & Pruzansky, 1966; Nygaard & Pisoni, 1996; Papcun, Kreiman, & Davis, 1989; Thompson, 1985a, 1985b; Williams, 1964; Yarmey & Matthys, 1992; cf. Clifford, Rathborn, & Bull, 1981, who found differences among voices with short, but not long, retention intervals). Thompson (1985a) reported that after 7 days hit rates (the percentage of correct "same" or "heard previously" responses) ranged from

15–90% for a set of 12 "nondistinctive" voices; false alarm rates (the percentage of incorrect "heard previously" responses) ranged from 15–50%. Clifford (1980; Clifford et al., 1981) reported differential effects of voice distinctiveness at short delays (10–130 min) between presentation and test: The number of correct responses declined over time for "hard-to-recognize" voices, but not for easy-to-recognize voices. Papcun et al. (1989) reported no differences in hit rates for easy-to-remember versus hard-to-remember voices at delays of 1, 2, or 4 weeks, but did find increases in the number of false alarms attributed to less distinctive targets as delay increased. Yarmey (1993) found that recognition of stereotypical "bad guy" voices was significantly worse than that of "good guy" voices. Differences in recognizability among speakers may be related in part to the speaker's gender: two studies reported that male voices were recognized at better rates than were female voices, although large differences in recognizability were observed within both groups of speakers (Nygaard & Pisoni, 1996; Thompson, 1985a). However, little else is known about the reasons for the variability observed. The literature provides little insight into the connection between perceptually salient aspects of voices and differences that predict subsequent recognition performance.

5.1.3 Studies Focusing on Stimuli

5.1.3.1 Sample Content and Duration

Hecker (1971) and Bricker and Pruzansky (1976) provide extensive reviews of studies examining how changes in the sample of speech a listener hears (e.g., vowels vs. phrases, phonemic content, duration) affect recognition and discrimination performance. These studies suggest that performance on recognition or discrimination tasks is much better predicted by the size of the sample of the speaker's phonetic repertoire than by the duration of the sample, with performance generally peaking with sample durations of about 1 s (Bricker & Pruzansky, 1966; Pollack, Pickett, & Sumby, 1954). A number of recent forensically oriented studies have further examined the contribution of voice sample duration to speaker recognizability. Clifford (1980) found no improvement in immediate recall for voices when stimuli increased in duration from one to four sentences; a subsequent experiment reported that percent correct recognition was greater when listeners heard two rather than one sentence. Legge, Grosmann, and Pieper (1984) reported that, after a retention interval of 15 min, the percentage of correct recognitions increased as sample duration increased from 6 to 60 s in a two-alternative forced-choice task; recognition of 6-s stimuli did not exceed chance levels. In a series of delayed recognition tasks, with delays ranging from a few minutes to 1 week, Yarmey (1991) and Yarmey and Matthys (1992) reported that both hit and false alarm rates increased at all delays as sample durations increased; stimuli ranged in duration from 18 s to 7.8 min. In contrast, Read and Craik (1995) found good recognition with sample durations of only 4 to 5 s after 17 days, provided identical voice samples were used at learning and recall. Emotionally

laden samples did not produce better recall than did samples with neutral content. Finally, Deffenbacher et al. (1989) suggested that effects of stimulus duration on recognition accuracy may depend on the delay between learning and testing. They reported no effect of stimulus duration for an immediate recall task, but a significant effect as delay increased.

5.1.3.2 Source versus Vocal Tract Contributions to Individual Voice Quality

A number of studies have used whisper, artificial voicing sources, and analysis-by-synthesis techniques to estimate the relative contributions of source and vocal tract information to personal quality. Early studies suggested vocal tract information contributes more to personal quality than source information does; however, eliminating either source of information reduces performance, but not to chance levels (see Bricker & Pruzansky, 1976). Abberton and Fourcin (1978) asked listeners to identify familiar voices from the output of a laryngograph, thus eliminating vocal tract and source–tract interaction information from the signal. They concluded that the source alone provides enough information to identify a familiar speaker from a small set, although they reported large differences among listeners in their performance. Tartter (1991) found that listeners were able to match whispered to phonated vowels produced by the same speaker, suggesting that vocal tract information is also adequate to identify speakers (although listeners again varied widely in performance, from 39–96% correct). Kuwabara and Takagi (1991) used analysis-by-synthesis to manipulate formant frequencies, bandwidths, and fundamental frequency (F0). They concluded that vocal tract information is more important than source information for identifying known speakers: Changes in formant frequencies exceeding 5% destroyed personal quality, whereas changes in F0 and bandwidths had little impact.

5.1.3.3 Signal Distortion

Previous studies (reviewed in Bricker & Pruzansky, 1976) indicate that playing voice samples backward reduces performance on voice perception tasks. Van Lancker and colleagues (Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985) evaluated the effects of rate-altering signals or playing them backwards on recognition of famous voices. Performance in all conditions depended on the voice: Some voices were identified equally well in distorted and normal conditions; some were unrecognizable when played backward, and others when rate-altered. These studies are discussed further below.

5.1.3.4 Disguise and Mimicry

Disguise significantly reduces both discrimination scores and recognition scores, for unfamiliar and familiar voices, and for voices listeners have been trained to identify (Clifford, 1980; Hirson & Duckworth, 1993; Hollien, Majewski, & Doherty, 1982; Reich, 1981; Reich & Duke, 1979). On the other hand, mimics are seldom mistaken for their targets (Hall & Tosi, 1975; Hollien et al., 1982; Rosenberg, 1973).

5.1.3.5 Effects of Stimulus Familiarity on Voice Perception

Listeners recognize familiar voices well, but not perfectly. Ladefoged and Ladefoged (1980) reported 31% of a set of familiar voices were successfully identified from the word *hello;* 66% were correctly recognized from a single sentence, with an overall false identification rate of 11%. Rose and Duncan (1995) reported that hello produced correct recognition rates ranging from 47–60% for a smaller set of voices. Listeners who knew the speakers well failed to identify 25% of samples and produced 18% false identifications. Schmidt-Nielsen and Stern (1985) reported that co-workers recognized each others' voices well (88% correct). The rate of correct identifications increased with rated familiarity; no relationship was found between the rated distinctiveness of a voice and the rate of correct identifications.

Other studies suggest listeners are not particularly good judges of whom they will recognize. Van Lancker, Kreiman, and Emmorey (1985) and Van Lancker, Kreiman, and Wickens (1985) asked listeners to identify 45 famous male speakers by name from 2-s voice samples. For voices that listeners claimed were familiar, naming averaged 27% correct in this unlimited response set task, and 70% correct in a six-alternative multiple-choice task. Read and Craik (1995) also found that a priori familiarity is not a good index of recognizability. Familiarity did not enhance performance on a delayed recognition task if the listener failed to identify the voice in a pretest, even if the listener claimed to know the voice very well.

5.1.3.6 Effects of Native Language, Accent, and Tone of Voice on Voice Perception

Perception of personality from accented speech is subject to the same kinds of stereotyping as other judgments of personality from speech (e.g., Gallois & Callan, 1981; Strongman & Woosley, 1967). A speaker's accent or tone of voice has produced variable effects on recognition scores. In an immediate recall task, Goldstein, Knight, Bailis, and Conover (1981) found effects of foreign accent on recognition scores for short stimuli, but not for longer utterances. Thompson (1987) reported that hit rates for native speakers of English were higher when stimuli were spoken in unaccented English than when they were in Spanish, but that hit rates for accented English did not differ from either English or Spanish speech. (False alarm rates did not differ for any of the conditions.) Hollien et al. (1982) found lower recognition scores when listeners did not know the language spoken than when they did. Spanish/English bilingual listeners recognized voices equally well when either language was spoken; bilingual speakers were better recognized when speaking the listener's native language (Goggin, Thompson, Strube, & Simental, 1991). Changing the emotional tone of utterances appears to have drastic effects on speaker recognizability. Saslove and Yarmey (1980) found that when listeners heard a hostile voice but were asked to recognize it from a

neutral sample, recognition was at chance. Similarly, Read and Craik (1995) found that recognition was at chance when the emotional tone of a voice was different at learning and test.

5.1.4 Studies Focusing on Differences among Tasks

A number of studies have manipulated the conditions under which voices are learned and/or recognized.

5.1.4.1 Retention Set Size

Increasing the number of voices to be remembered decreases the likelihood that a target will be correctly identified after a delay (Carterette & Barnebey, 1975; Legge et al., 1984).

5.1.4.2 Effects of Increasing the Number of Exposures to the Target Voice

Goldstein and Chance (1985, cited in Deffenbacher et al., 1989) reported that recognition improved when a voice sample was divided into segments and presented over several days, as compared to presentation of the entire sample at one session. Yarmey and Matthys (1992) found variable effects of such distributed stimulus presentations. Hit rates improved when stimuli were presented in two pieces, but not three, and false-alarm rates were not affected by the distributed presentation. Playing a *distractor* (nontarget) voice in a mock one-voice lineup between learning and final test tripled the false-alarm rate (Thompson, 1985b).

5.1.4.3 Passive versus Active Learning

Hammersley and Read (1985) compared recognition of tape-recorded voices with recognition after participation in a conversation. After a 48-hr delay, recognition of the taped voices was at chance levels, but listeners were able to identify voices heard in active conversation with better than chance accuracy.

5.1.4.4 Does Seeing a Face Improve Voice Recognition?

Seeing a face when a target voice is initially presented may or may not help listeners identify that voice later. Legge et al. (1984) found that the percentage of correct identifications was higher when a face was presented at both learning and testing; seeing a face at learning, but not recall, provided no advantage. Yarmey (1993) also reported that seeing a face when learning a voice produced higher hit rates and lower false-alarm rates than when the voice was learned and recognized alone. In contrast, McAllister, Dale, Bregman, McCabe, and Cotton (1993) reported that hit rates were higher and false-alarm rates unchanged when a face was *not* present during learning.

5.1.4.5 Retention Interval

Studies comparing recognition accuracy for unfamiliar voices with short and long delays between learning and testing have produced highly variable results. Kreiman and Papcun (1991) found no differences between discrimination scores and recognition accuracy after 1 week, although patterns of confusions and perceptual strategies differed across delays. (This study is discussed further below). Saslove and Yarmey (1980) reported no difference in accuracy between immediate testing and a delay of 24 hr. Clifford (1980; Clifford et al., 1981) found better performance after a 10-min delay than after 1, 7, or 14 days; the longer delays did not differ, with performance near chance after 24 hr. In contrast, Legge et al. (1984) found no difference between delays of 15 min and 10 days in percent correct recognition; and Yarmey (1991) reported no difference between immediate recall and recognition after 2–3 days. Yarmey and Matthys (1992) compared immediate testing and delays of 24 hr and 1 week. They also found no differences in hit or false-alarm rates among conditions, but reported many confusing interactions between delay, stimulus duration, and frequency of exposure to the target voice.

Studies comparing long retention intervals have also produced conflicting results. Papcun et al. (1989) reported a linear decline in recognition accuracy as retention interval increased from 1 to 4 weeks, with scores above chance at all delays. Read and Craik (1995) found no differences among delays ranging from 6–37 days. Finally, Hollien, Bennett, and Gelfer (1983) reported no difference in performance for delays of 1 day, 1 week, and 2 weeks, with performance at chance for all delays.

5.1.5 Studies Focusing on Effects of Different Transmission Systems

Studies assessing the effects of acoustic filtering on voice recognizability or discriminability are reviewed in Bricker and Pruzansky (1976). Filtering generally reduces recognition and discrimination scores, but not to chance levels, suggesting that listener performance does not depend on any particular part of the speech-frequency spectrum. Linear predictive coding (LPC) also reduces voice recognizability, but not to chance levels: hit rates for familiar voices were lower for coded than for unprocessed recordings, but remained well above chance in both conditions (88 vs. 69%; Schmidt-Nielsen & Stern, 1985).

Several studies have examined voice recognizability over the telephone. McGonegal, Rabiner, and McDermott (1978) compared recognition of taped and live voices to recognition of voices heard over the telephone, and found no difference between systems as long as learning and testing took place over the same system. Rathborn, Bull, and Clifford (1981) reported that hit rates were significantly higher when voices were presented live than over the telephone, but confirmed that the best performance occurs when learning and testing use the same system.

5.1.6 Studies Focusing on Listeners

5.1.6.1 Voice Perception Abilities of Different Listener Populations

Although not a uniquely human ability (Cheney & Seyfarth, 1980; Goldman, Phillips, & Fentress, 1995; White, White, & Thorpe, 1970), the ability to recognize voices appears to be innate, and voice learning may begin in utero. Fetuses at 36-weeks gestation did not discriminate between the voice of their mother and other maternal voices when stimuli were played via a loudspeaker on the abdomen, but did discriminate between their mother's voice played over the loudspeaker and their mother speaking normally (Hepper, Scott, & Shahidullah, 1993). Two-day old infants preferred voices filtered to mimic the uterine environment to unfiltered voices (Fifer & Moon, 1989), and could discriminate among male voices (but not recognize their father's voice; DeCasper & Prescott, 1984). Three-day old infants can recognize the voice of their mother from a set of maternal voices (DeCasper & Fifer, 1980). Four-month-old infants can discriminate their parents' voices from those of same-sex strangers (Brown, 1979). By age 7–8 months, infants can learn to recognize unfamiliar voices (Jusczyk, Hohne, Jusczyk, & Redanz, 1993); and voice recognition abilities reach adult levels by age 10 (Mann, Diamond, & Carey, 1979).

Normal adult listeners differ widely in their abilities to recognize or discriminate among speakers (e.g., Hollien et al., 1982; Nygaard & Pisoni, 1996). Kreiman (1987) reported that discrimination performance declined significantly with a listener's age. Male and female listeners apparently do not differ in voice recognition accuracy (Thompson, 1985a).

Phoneticians are not more accurate than naive listeners in voice discrimination tasks (Shirt, 1983). Expert and naive listeners do differ significantly in the strategies they use when evaluating pathological voice quality (Kreiman, Gerratt, & Precoda, 1990). Naive listeners used relatively simple perceptual strategies and generally attended to a consistent set of voice "features," whereas experts differed considerably from one another and from the naive listeners in the strategies applied when judging quality (Kreiman, Gerratt, Precoda, & Berke, 1992).

The voice recognition abilities of blind listeners may exceed those of sighted controls (Ball, Rathborn, & Clifford, 1983) or may not (Winograd, Kerr, & Spence, 1984). Neurological disease may interfere with voice recognition or discrimination abilities. Studies by Van Lancker and colleagues (e.g., Van Lancker, Kreiman, & Cummings, 1989) showed that deficits in voice discrimination are associated with temporal lobe damage to either hemisphere, whereas deficits in recognizing familiar voices are associated with damage to the right parietal lobe. (These studies are reviewed further in the next sections.) A case study of a blind Capgras patient revealed a deficit in recognizing familiar voices (Reid, Young, & Hellawell, 1993), and Korsakoff amnesia also interferes with familiar-voice recognition (although memories from the remote past are relatively preserved; Meudell, Northen, Snowden, & Neary, 1980). Adults with dyslexia showed significant impairment compared to control subjects when judging if a presented voice was a

new or old stimulus, but not on a similar facial recognition task (Brachacki, Fawcett, & Nicolson, 1994).

5.1.6.2 Confidence in Responses and Listener Accuracy

Many studies have reported that listeners' confidence in their responses is not a significant predictor of response accuracy (Bull & Clifford, 1984; Hollien et al., 1983; Yarmey, 1986, cited in Deffenbacher et al., 1989; Yarmey, 1991; Yarmey & Matthys, 1992; Yarmey, Yarmey, & Yarmey, 1994). Many others have reported significant but small associations between confidence and accuracy (Rathborn et al., 1981; Saslove & Yarmey, 1980; Thompson, 1985a, 1985b). Read and Craik (1995) argued that confidence predicts performance when the task is easy (for example, when the stimuli played at learning and test match in text and emotional tone), but not otherwise. Finally, Rose and Duncan (1995) reported large positive correlations (ρ = .90–.92) between confidence and accuracy when speakers are well known to listeners.

5.1.6.3 Studies Examining Listeners' Perceptual Strategies

A number of studies (Carterette & Barnebey, 1975; Fagel, van Herpt, & Boves, 1983; Gelfer, 1993; Kreiman & Papcun, 1991; Matsumoto, Hiki, Sone, & Nimura, 1973; Murry & Singh, 1980; Singh & Murry, 1978; Walden, Montgomery, Gibeily, Prosek, & Schwartz, 1978) have used multidimensional scaling (MDS) or factor analysis to examine the features that underlie perceived differences in the quality of normal voices. Results are summarized in Table I. Fundamental frequency or pitch characteristics, vowel formant frequencies, and stimulus duration have emerged as perceptually significant features of voice on a fairly regular basis, as has a "voice quality" dimension (whose appearance in studies attempting to derive the dimensions underlying judgments of voice quality is somewhat disappointing). However, many discrepancies exist between studies. Some may be attributable to the variety of subject and listener populations studied and to the variety of stimuli employed. Singh and Murry (1978) suggested that different perceptual features may be used in processing male and female voices, and argued that the cues used to judge voices depend on both the stimulus and the gender of the speaker (Murry & Singh, 1980). However, the dimensions reported by Gelfer (1993) for female voices have also emerged from studies of male voices, and she argued that gender differences probably do not contribute significantly to differences among studies.

Few studies have attempted to relate perceptual dimensions to characteristics of the stimuli or the listeners. Listeners' experience may contribute to differences in perceptual strategies. In a study of pathological voices, Kreiman et al. (1990) found that different dimensions underlie quality judgments by expert and naive listeners, and Gelfer (1993) reported similar findings for expert and naive judgments of normal female quality. Individual expert listeners also differ from one another in their perceptual strategies (Kreiman, Gerratt, & Berke, 1994; Kreiman et al., 1992). Listener agreement on ratings of pathological voice can be predicted

TABLE I Multidimensional Scaling and Factor Analysis Studies of Normal Voice Quality

Study	Dimensions/features[a]
Matsumoto et al. (1973)	F0; glottal source spectrum; jitter; formant frequencies
Carterette and Barnebey (1975)	F0; intensity; intonation pattern
Walden et al., (1978)	F0; stimulus duration; speaker's age; "superior" vs. "inferior" voice quality
Singh and Murry (1978)	Speaker gender; pitch (male speakers only); stimulus duration (female speakers only)
Murry and Singh (1980)	F0; formants; nasality; F2 (male voices, vowel)
	F0; hoarseness; formants (male voices, phrase)
	F0; breathiness; formants; breathiness (female voices, vowel)
	Effort/nasality; F0; stimulus duration (female voices, phrase)
Fagel et al. (1983)	Melodiousness; articulation quality; voice quality; pitch; tempo (factor analysis)
Kreiman and Papcun (1991)	Masculinity; creakiness; variability; mood (discrimination task)
	Masculinity; breathiness; liveliness (long-term memory task)
Gelfer (1993)	Pitch; loudness; age/stimulus duration; perceived variability; quality (expert listeners)
	Pitch/resonant quality; variability/age/rate (naive listeners)

[a]F0, Fundamental frequency; F2, formant 2.

from perceptual strategies (as estimated with MDS). Listeners with similar perceptual spaces agreed better on ratings of breathiness and roughness than did listeners whose perceptual spaces differed (Kreiman et al., 1994). To some extent, listeners' attention to particular stimulus dimensions can be predicted by the extent of stimulus variability on those dimensions. Variability above a certain threshold is a necessary, but not sufficient, condition for a dimension to emerge as perceptually salient (Kreiman et al., 1992). Listener preferences for different dimensions with equal variability are presumably related to differences in their perceptual habits and experience with the class of stimuli, but the origin and nature of these differences are not well understood. Studies relating characteristics of stimuli, listener perceptual strategies, and performance on voice perception tasks have not been undertaken for normal voices, and it is not presently possible to explain why listeners differ in accuracy, why they attend to a particular characteristic on one occasion and not on another, or how or why listeners differ from one another in perceptual strategy when judging normal voice quality.

5.1.7 Theoretical Perspectives

As this review indicates, a significant body of research on voice perception has produced many discrepant results. Such discrepancies have led authors to call

for (and frequently to conduct) further research, but seldom to speculate about why a particular finding emerged, or why other studies have produced different results, and rather little theory exists to explain why results differ from study to study.

A few authors have attempted to investigate the perceptual and memory processes underlying recognition, discrimination, or evaluative tasks in voice perception. One group of studies (Kreiman, 1987; Kreiman & Papcun, 1991; Papcun et al., 1989) examined long- (LTM) and short-term memory (STM) for unfamiliar voices. Listeners in Papcun et al. (1989) heard a speaker making a telephone survey call, and attempted to identify that speaker from a set of 10 recordings after a delay of 1, 2, or 4 weeks. (Listeners also rated their confidence in their response and the similarity of the probe voice to the remembered target.) Signal-detection analyses showed that recognition accuracy declined significantly with delay. Target voices did not differ in how likely they were to be correctly identified. However, foils that were "average-sounding" were more likely to be falsely identified as a target than were more distinctive-sounding foils. These findings were interpreted in terms of a prototype model of memory for voices. The authors speculated that listeners encode vocal quality in terms of a central category member (a "prototype") and a set of deviations from that prototype. Over time, the deviations are forgotten, and recognition responses converge on the most average-sounding voice, whatever target was originally heard.

A second study (Kreiman & Papcun, 1991) compared the discriminability of the voices studied by Papcun et al. (1989) to their recognizability after 1 week, to determine how stored representations of voice quality change over time. The recognition task was identical to that used by Papcun et al.; in the discrimination task, listeners heard pairs of sentences excerpted from the telephone surveys. They were asked to determine whether the speakers were the same or different, to rate their confidence in their response, and to rate the similarity of the two voices. MDS analyses of similarity ratings from the two tasks indicated that rather different perceptual strategies were used. The perceptual space for the LTM task included fewer dimensions, but accounted for more variance, than did the space for the discrimination task. A single perceptual dimension (labeled "masculinity") was shared by the two scaling solutions, but no other significant correspondences between spaces emerged. However, the changes in stored representations over time did not correspond to a loss of ability to distinguish among voices: signal-detection analyses indicated that discrimination accuracy did not differ significantly from recognition accuracy. Patterns of confusions did vary across tasks, in a manner consistent with the prototype model proposed by Papcun et al. Identification responses converged on the least-distinctive voices in the LTM task, again suggesting that over time the "distinctive characteristics" of a voice are forgotten and a "typical" pattern is retained in memory.

A group of studies by Van Lancker and colleagues examined perception of and memory for familiar voices. Several studies of normal listeners (Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985) used naming and multiple-choice tasks to examine recognition of famous voices under

various conditions of signal distortion. These studies suggested that the cues to personal identity vary from voice to voice (see also Van Dommelen, 1990). Furthermore, the perceptual context in which a cue occurred was important in evaluating its usefulness. For example, a distinctive accent was not critical to recognizing a voice if that voice was also very distinctive on some other dimension(s). These data suggested that no single set of voice features exists that characterizes all voices. Instead, the authors proposed, only the salient features are stored for any familiar voice, and each voice constitutes a unique pattern of features. In the framework of Papcun et al. (1989), familiar voices are no longer remembered in terms of a prototype and deviations from that prototype. Rather, only the deviations (the distinctive aspects) are retained in LTM. (Note, however, that Walden et al. (1978) found no obvious effects of familiarity on perceptual strategy in an MDS study.)

Direct evidence for differences in the processing of familiar and unfamiliar voices comes from studies of subjects with focal brain damage (e.g., Van Lancker, Cummings, Kreiman, & Dobkin, 1988; Van Lancker & Kreiman, 1987; Van Lancker et al., 1989). In these studies, performance on a voice discrimination task [using a subset of the unfamiliar voice stimuli studied by Kreiman and Papcun (1991)] and on a familiar voice recognition task was correlated with the site of lesion. Subjects were observed who could recognize familiar voices but not discriminate among unfamiliar voices, and vice versa, indicating that voice recognition and discrimination are dissociated neurological functions. As a group, left-brain-damaged subjects did not differ from normal controls on the voice recognition task, but right-brain-damaged patients were significantly impaired in recognizing familiar voices. Both brain-damaged groups differed overall from controls on the voice discrimination task. Impaired familiar voice recognition was associated with damage to the right parietal lobe, whereas deficits in voice discrimination were associated with temporal lobe damage to either hemisphere.

Finally, recent studies of pathological voice quality (Kreiman & Gerratt, 1996; Kreiman et al., 1990, 1992, 1994) suggest that listeners may contribute nearly as much variance to ratings of quality as do differences between voices. In these studies, listeners judged the dissimilarity of pairs of pathological voices, and the resultant ratings were analyzed with MDS to determine the perceptually important voice "features" that underlie similarity judgments. Results of these studies indicated that quality judgments depend critically on the particular listener and act of listening. Although individual listeners were quite reliable in their judgments, perceived similarity was not constant across listeners, and the scaling analyses revealed only a single perceptual dimension (severity of pathology) that was common to all the listeners. These studies used pathological voices and evaluation tasks, which limits their generalizability to recognition and discrimination of normal voices, but they suggest that voice perception cannot be modeled solely as a function of differences between voices, and that listeners' contributions to voice perception cannot be neglected. If listeners did not differ in their perceptual strategies, vocal quality could be modeled solely as an attribute of voices. That is, when listeners are interchangeable, observed differences in the ratings different voices receive must be due to differences among voices. However, differences among listeners in perceptual strategies are large enough to suggest that it is imprudent to continue the traditional practice of treating differences among listeners as "noise" to be ignored or controlled.

Many limitations to these studies (and to the models they propose) are evident. Although large sets of listeners have been employed, the voice sets studied were rather small (only 10 unfamiliar voices in the Papcun et al. studies, and 26 pairs of unfamiliar voices and 25 familiar voices in the Van Lancker et al. studies). Results have not been replicated in different laboratories or with different samples of speakers and listeners. Finally, the models proposed are mere outlines, and competing explanations are possible for the data presented. Unfortunately, efforts to refine these proposals have not been undertaken.

Although previous studies point to prototype models of voice perception, there is not enough evidence at present to determine whether such models will ultimately serve to describe the processes by which listeners evaluate and recognize voices. However, existing data do suggest that other classes of model are probably inappropriate. For example, linguistic-style distinctive feature models of vocal quality (e.g., Gelfer, 1988; Hammarberg & Gauffin, 1995; Laver, 1980) are unlikely to predict listener performance in voice perception tasks. However elegant or detailed the manner in which such models specify the ways in which voices may differ, they do not model listeners' contributions to the perceptual process (nor are they intended to model listeners' behavior). Thus they cannot explain the relative perceptual importance of different quality "settings," or predict which cues listeners will attend to; and they cannot account for variability in the cues used for different voices or in perceptual strategy across listeners.

5.2 DISCUSSION

As this review indicates, very little is actually known about voice perception (despite common belief; cf. Hollien's statement that "we now know a great deal about people's ability to make aural perceptual speaker identifications" [1990, p. 205]). Experimenters have varied the speaker, stimulus, task, transmission system, and listener population, but conflicting results have emerged in nearly every area of study.[2] No basis exists for linking variability in listener performance to manipulations in other domains, in part because performance is generally measured in terms of response accuracy. Other data that might shed light on how perceptual strategies vary with experimental manipulations (for example, changes

[2] A similar state of affairs exists in the study of pathologic vocal quality, where the literature contains many papers with titles like, "Have the major issues in voice disorders been answered by research in speech science? A 50-year retrospective" (Moore, 1977) and "Voice therapy: A need for research" (Reed, 1980).

in patterns of confusions or similarity ratings) generally are not examined. Thus it is unlikely such studies will ever explain why results differ from paper to paper.

Response accuracy is not without interest, and a reasonable amount of description is necessary before theory can be generated. However, the majority of voice perception studies are not designed to investigate the perceptual processes by which listeners recognize, discriminate among, or evaluate voice signals, and thus do not contribute to developing an understanding of perceptual processes. Instead, studies often seem to "fill in the cells" in a paradigm. (A particularly egregious example of this is a series of papers using direct estimations and paired comparison tasks to examine the effects of disguise, filtering, temporal alterations, phonetic complexity, and voicing on listeners' evaluations of height, weight, gender, and race.) Other papers repeatedly replicate previous findings, without obvious motivation for the replication (for example, the many investigations of the relationship between listeners' confidence and accuracy, or of the effects of response interval on recognition accuracy). In this system, discrepant results are defined with reference to previous findings, not in terms of the predictions of a model. Consequently, they do not lead to modifications of a model and new predictions, but instead generate more studies with more discrepant results. Thus, research questions are never resolved, but are investigated over and over again, as the review above demonstrates. Combined with a general lack of new research foci, this pattern suggests that the descriptive level of study is saturated. Only the development of a theoretical basis for research into voice quality can stem the proliferation of repetitive, inconclusive studies.

Lack of a viable model has also begun to limit progress in other disciplines. For example, broader questions of the acoustic correlates and physiological underpinnings of vocal quality cannot be answered in the absence of an understanding of voice perception. Furthermore, voice research provides no guidance for investigators studying spoken-word recognition, who as a result must select the particular manipulations of voice they use in a somewhat arbitrary fashion. For example, recent work in speech perception has examined how differences in a speaker's gender, speaking rate, and dialect affect spoken-word recognition, but it is unclear why these particular factors were examined, and not others (see Pisoni, 1993, for review). Better models of voice perception would provide a principled basis for developing hypotheses about the acoustic and physiological correlates of perceived quality, and about how voice quality interacts with lexical information in speech recognition.

Finally, systematic investigation of the role of listeners in voice perception has not taken place. Quality is *perceptual* in nature: It is a function of both listeners and stimuli, not just of stimuli. However, the great majority of voice perception studies focus on how a manipulation affects response accuracy, the goal being to determine how well voices are recognized, not to learn what listeners are doing. This approach is essentially "stimulus-driven": stimuli or test conditions are manipulated, and responses tabulated in terms of the effects of these manipulations, but changes in listeners' perceptual processes across conditions

and differences among listeners within conditions are generally ignored. In other words, these studies examine main effects of stimulus or task conditions, but neglect the main effect of listeners and interactions between stimuli, tasks, and listeners.

The importance of modeling both listener and interaction effects in voice perception has long been known. For example, a factor analysis study by Voiers (1964) produced both listener-specific terms and interactions between specific listeners and voices, which he termed *extrastimulus factors* in speaker recognition. Recent studies of pathological voice quality also suggest that differences among listeners may explain as much variance in ratings of quality as differences among voices do (Kreiman & Gerratt, 1996), as described above. Note that the traditional view that voice recognition and discrimination depend on the extent to which interspeaker variability in quality exceeds intraspeaker variability also assumes that voice perception can be modeled as a function of differences among signals, and consequently may require modification.

One possible cause of the lack of theoretical development, relative neglect of listener factors, and reliance on measures of response accuracy that characterize voice perception research, is the persistent focus of most voice research on specific, practical questions. Bricker and Pruzansky's statement remains true today: "While some studies have been undertaken solely to increase our basic knowledge of the process, by far the greater number have been motivated by some practical considerations" (1976, p. 296). Researchers in different disciplines ask different questions: Forensic scientists are interested in earwitness accuracy under different conditions, neurologists are interested in the effects of disorders on perception, otolaryngologists are interested in the effects of disease or treatment on voice quality, and so on. The variety of practical questions motivating voice perception studies has resulted in a literature that is fragmented across disciplines. A reference database maintained in the University of California, Los Angeles (UCLA) Voice Laboratory lists studies in more than 200 journals, spanning a huge range of disciplines (Table II). Lack of a body of theory promotes this fragmentation, and fragmentation also inhibits the development of theory, perpetuating the problem. Thus the pursuit of answers to specific questions has limited the development of voice perception research as a coherent discipline, and has paradoxically contributed to the development of a research climate in which these very questions cannot be answered.

The fragmentation of voice research across disciplines may also have acted to limit the number of researchers investigating voice perception. The UCLA reference database, which covers 63 years of research, lists 1740 authors who have published on some aspect of voice perception or production, 84% of whom have produced only one or two papers on voice-related topics. Less than 6% of the authors indexed have authored more than five papers on voice. Many factors serve to perpetuate this situation. Students at the beginning of their studies may find it difficult to acquire the necessary interdisciplinary expertise within the constraints of doctoral program requirements, to assemble a suitable dissertation

TABLE II Academic Disciplines that Include Studies of Voice

Acoustics
Animal behavior
Electrical engineering/computer science/signal processing
Forensic/police science
Linguistics
Medicine
Developmental biology
Gerontology
Neurology
Obstetrics/Gynecology
Otolaryngology
Pediatrics
Respiration
Surgery
Music
Phonetics
Psychology
Cognitive
Clinical
Social
Neuropsychology
Psychophysics
Psycholinguistics
Sociology
Speech pathology
Speech science

committee, and to convince that committee that voice studies are an appropriate dissertation topic. Students who pursue the study of voice perception may find it difficult to affiliate with a particular department when their graduate work is completed. Given its interdisciplinary nature, the study of vocal quality has no clear academic home. Maintaining the necessary expertise on voice is difficult: Even the most diligent researcher may find it difficult to consistently monitor the contents of 200 or more journals. Finally, it can be difficult to assemble and maintain an interdisciplinary laboratory group, even at a large university.

A number of changes in research practices are required to remedy this situation. First, scientific method must be consistently applied in studies of voice perception. Authors, reviewers, and editors must insist that the questions addressed are motivated by theoretical concerns, and that the manipulations introduced will resolve these questions, not simply add one more study to the weight of evidence on one side or the other of a question. Studies must address the issues of *why* a particular result occurred, and not merely tally the number of voices correctly recognized under particular circumstances. More researchers must work

systematically on voice perception, and funding for basic, rather than applied, studies of voice must be secured. (Development of a theoretical base for such studies should facilitate this process.) Researchers must resist the temptation to reduce the study of voice perception to the study of stimuli, unless they provide a clear explanation of how the particular manipulation will elucidate the larger issues at hand. Finally, much more effort must be devoted to understanding listeners' contribution to voice perception. The problem of how speakers perceive and identify voices is a difficult one, but it will not be solved if the current practice of neglecting theory in favor of simplification of the questions continues.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by grant DC01797 from the National Institute on Deafness and Other Communication Disorders. Thanks to Abeer Alwan, Bruce Gerratt, Peter Ladefoged, and the UCLA voice community for many helpful comments on earlier versions.

REFERENCES

Abberton, E., & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech, 21,* 305–318.

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). The effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology, 37,* 715–727.

Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *Journal of Social Psychology, 99,* 207–220.

Berke, G. S., & Gerratt, B. R. (1993). Laryngeal biomechanics: An overview of mucosal wave mechanics. *Journal of Voice, 7,* 123–128.

Brachacki, G. W., Fawcett, A. J., & Nicolson, R. I. (1994). Adults with dyslexia have a deficit in voice recognition. *Perceptual and Motor Skills, 78,* 304–306.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40,* 1441–1449.

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 295–326). New York: Academic Press.

Brown, C. J. (1979). Reactions of infants to their parents' voices. *Infant Behavior and Development, 2,* 295–300.

Bull, R., & Clifford, B. R. (1984). Earwitness voice recognition accuracy. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony* (pp. 92–123). New York: Cambridge University Press.

Bull, R., Rathborn, H., & Clifford, B. R. (1983). The voice-recognition accuracy of blind listeners. *Perception, 12,* 223–226.

Carterette, E. C., & Barnebey, A. (1975). Recognition memory for voices. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 246–265). New York: Springer.

Cheney, D. L., & Seyfarth, R. (1980). Vocal recognition in free-ranging vervet monkeys. *Animal Behaviour, 28,* 362–367.

Childers, D. G., & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America, 90,* 2394–2410.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4,* 373–394.

Clifford, B. R., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior, 5,* 201–208.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voice. *Science, 208,* 1174–1176.

DeCasper, A. J., & Prescott, P. A. (1984). Human newborns' perception of male voices: Preference, discrimination, and reinforcing value. *Developmental Psychobiology, 17,* 481–491.

Deffenbacher, K. A., Cross, J. F., Handkins, R. E., Chance, J. E., Goldstein, A. G., Hammersley, R., & Read, J. D. (1989). Relevance of voice identification research to criteria for evaluating reliability of an identification. *Journal of Psychology, 123,* 109–119.

Denes, P. B., & Pinson, E. N. (1993). *The speech chain: The physics and biology of spoken language* (2nd ed.). New York: Freeman.

Disner, S. F. (1982). Stress evaluation and voice lie detection: A review. *UCLA Working Papers in Phonetics, 54,* 78–92.

Fagel, W. P. F., van Herpt, L. W. A., & Boves, L. (1983). Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. *Speech Communication, 2,* 315–326.

Fifer, W. P., & Moon, C. (1989). Psychobiology of newborn auditory preferences. *Seminars in Perinatology, 13,* 430–433.

Gallois, C., & Callan, V. J. (1981). Personality impressions elicited by accented English speech. *Journal of Cross-Cultural Psychology, 12,* 347–359.

Garvin, P. L., & Ladefoged, P. (1963). Speaker identification and message identification in speech recognition. *Phonetica, 9,* 193–199.

Gelfer, M. P. (1988). Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice, 2,* 320–326.

Gelfer, M. P. (1993). A multidimensional scaling study of voice quality in females. *Phonetica, 50,* 15–27.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19,* 448–458.

Goldman, J. A., Phillips, D. P., & Fentress, J. C. (1995). An acoustic basis for maternal recognition in timber wolves (*Canis lupus*)? *Journal of the Acoustical Society of America, 97,* 1970–1973.

Goldstein, A. G., & Chance, J. E. (1985). *Voice recognition: The effects of faces, temporal distribution of "practice," and social distance.* Paper presented at the meeting of the Midwestern Psychological Association, Chicago.

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society, 17,* 217–220.

Granström, B. (1992). The use of speech synthesis in exploring different speaking styles. *Speech Communication, 11,* 347–355.

Gunter, C. D., & Manning, W. H. (1982). Listener estimations of speaker height and weight in unfiltered and filtered conditions. *Journal of Phonetics, 10,* 251–257.

Hall, M., & Tosi, O. I. (1975). Spectrographic and aural examination of professionally mimicked voices [Abstract]. *Journal of the Acoustical Society of America, 58,* S107.

Hammarberg, B., & Gauffin, J. (1995). Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. In O. Fujimura & M. Hirano (Eds.), *Vocal fold physiology: Voice quality control* (pp. 283–303). San Diego, CA: Singular.

Hammersley, R., & Read, J. D. (1985). The effect of participation in conversation on recognition and identification of the speakers' voices. *Law and Human Behavior, 9,* 71–81.

Hartman, D. E., & Danauer, J. L. (1976). Perceptual features of speech for males in four perceived age decades. *Journal of the Acoustical Society of America, 59,* 713–715.

Hecker, M. H. L. (1971). Speaker recognition: An interpretive survey of the literature. *ASHA Monographs, 16.*

Helfrich, H. (1979). Age markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 63–108). Cambridge, England: Cambridge University Press.

Hepper, P. G., Scott, D., & Shahidullah, S. (1993). Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology, 11,* 147–153.

Hirson, A., & Duckworth, M. (1993). Glottal fry and voice disguise: A case study in forensic phonetics. *Journal of Biomedical Engineering, 15,* 193–200.

Hollien, H. (1990). *The acoustics of crime: The new science of forensic phonetics.* New York: Plenum.

Hollien, H., Bennett, G. T., & Gelfer, M. P. (1983). Criminal identification comparison: Aural vs. visual identifications resulting from a simulated crime. *Journal of Forensic Sciences, 28,* 208–221.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress, and disguise speaking conditions. *Journal of Phonetics, 10,* 139–148.

Hunt, R. G., & Lin, T. K. (1967). Accuracy of judgments of personal attributes from speech. *Journal of Personality and Social Psychology, 6,* 450–453.

Jacques, R., & Rastatter, M. (1990). Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. *Folia Phoniatrica, 42,* 118–124.

Jusczyk, P., Hohne, E., Jusczyk, A. M., & Redanz, N. J. (1993) Do infants remember voices? [Abstract]. *Journal of the Acoustical Society of America, 93,* 2373.

Karlsson, I. (1992). *Analysis and synthesis of different voices with emphasis on female speech.* Stockholm: Royal Institute of Technology (KTH), Department of Speech Communication and Musical Acoustics.

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

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin, 60,* 408–420.

Kreiman, J. (1987). *Human memory for unfamiliar voices.* Unpublished doctoral dissertation, University of Chicago, Chicago.

Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America, 100,* 1787–1795.

Kreiman, J., Gerratt, B. R., & Berke, G. S. (1994). The multidimensional nature of pathologic vocal quality. *Journal of the Acoustical Society of America, 96,* 1291–1302.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21–40.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33,* 103–115.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512–520.

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication, 10,* 265–275.

Kunzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica, 46,* 117–125.

Kuwabara, H., & Takagi, T. (1991). Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Communication, 10,* 491–495.

Ladefoged, P. (1956). The classification of vowels. *Lingua, 5,* 113–128.

Ladefoged, P., & Ladefoged, J. (1980). The ability of listeners to identify voices. *UCLA Working Papers in Phonetics, 49,* 43–51.

Lass, N. J., Almerino, C. A., Jordan, L. F., & Walsh, J. M. (1980). The effect of filtered speech on speaker race and sex identifications. *Journal of Phonetics, 8,* 101–112.

Lass, N. J., Barry, P. J., Reed, R. A., Walsh, J. M., & Amuso, T. A. (1979). The effect of temporal speech alterations on speaker height and weight identification. *Language and Speech, 22,* 163–171.

Lass, N. J., Beverly, A. S., Nicosia, D. K., & Simpson, L. A. (1978). An investigation of speaker height and weight identification by means of direct estimations. *Journal of Phonetics, 6,* 69–76.

Lass, N. J., & Davis, M. (1976). An investigation of speaker height and weight identification. *Journal of the Acoustical Society of America, 60,* 700–703.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered and filtered isolated vowels. *Journal of the Acoustical Society of America, 59,* 675–678.

Lass, N. J., Mertz, P. J., & Kimmel, K. L. (1978). The effect of temporal speech alterations on speaker race and sex identifications. *Language and Speech, 21,* 279–290.

Lass, N. J., Phillips, J. K., & Bruchey, C. A. (1980). The effect of filtered speech on speaker height and weight identification. *Journal of Phonetics, 8,* 91–100.

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

Laver, J. (1981). The analysis of vocal quality: From the classical period to the 20th century. In R. Asher & E. Henderson (Eds.), *Toward a history of phonetics* (pp. 79–99). Edinburgh: Edinburgh University Press.

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 298–303.

Linville, S. E., & Fisher, H. B. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *Journal of the Acoustical Society of America, 78,* 40–48.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology, 27,* 153–165.

Matsumoto, H., Hiki, S., Sone, T., & Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics, AU-21,* 428–436.

McAllister, H. A., Dale, R. H. I., Bregman, N. J., McCabe, A., & Cotton, C. R. (1993). When eyewitnesses are also earwitnesses: Effects on visual and voice identifications. *Basic and Applied Social Psychology, 14,* 161–170.

McGonegal, C., Rabiner, L., & McDermott, B. (1978). Speaker verification by human listeners over several speech transmission systems. *Bell System Technical Journal, 57,* 2887–2900.

Meudell, P. R., Northen, B., Snowden, J. S., & Neary, D. (1980). Long term memory for famous voices in amnesic and normal subjects. *Neuropsychologia, 18,* 133–139.

Moore, P. (1977). Have the major issues in voice disorders been answered by research in speech science? A 50-year retrospective. *Journal of Speech and Hearing Disorders, 42,* 152–160.

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., & Farnsworth, L. M. (1995). The perceptual representation of voice gender. *Journal of the Acoustical Society of America, 98,* 3080–3095.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *Journal of the Acoustical Society of America, 68,* 1294–1300.

Nygaard, L. C., and Pisoni, D. B. (1996). *Learning voices.* Paper presented at the 131st meeting of the Acoustical Society of America, Indianapolis, IN.

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

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85,* 913–925.

Peters, R. W. (1954). *Studies in extra messages: Listener identification of speakers' voices under conditions of certain restrictions imposed upon the voice signal* (Joint Project Report No. 30, Project No. NM 001-064-01). Pensacola, FL: Naval School of Aviation Medicine, Naval Air Station.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication, 13,* 109–125.

Pisoni, D. B., & Martin, C. S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research, 13,* 577–587.

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185–197). New York: Guilford Press.

Pollack, I., Pickett, J., & Sumby, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America, 26,* 403–406.

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of Speech and Hearing Research, 9,* 273–277.

Rathborn, H., Bull, R., & Clifford, B. R. (1981). Voice recognition over the telephone. *Journal of Police Science and Administration, 9,* 280–284.

Read, J. D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1,* 6–18.

Reed, C. (1980). Voice therapy: A need for research. *Journal of Speech and Hearing Disorders, 45,* 157–189.

Reich, A. R. (1981). Detecting the presence of vocal disguise in the male voice. *Journal of the Acoustical Society of America, 69,* 1458–1461.

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguise upon speaker identification by listening. *Journal of the Acoustical Society of America, 66,* 1023–1028.

Reid, I., Young, A. W., & Hellawell, D. J. (1993). Voice recognition impairment in a blind Capgras patient. *Behavioral Neurology, 6,* 225–228.

Rose, P., & Duncan, S. (1995). Naive auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics, 2,* 1–17.

Rosenberg, A. E. (1973). Listener performance in speaker verification tasks. *IEEE Transactions on Audio and Electroacoustics, AU-21,* 221–225.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65,* 111–116.

Scherer, K. R. (1979). Personality markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 147–210). Cambridge, England: Cambridge University Press.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143–165.

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

Shipp, T., Qi, Y., Huntley, R., & Hollien, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of Voice, 6,* 211–216.

Shirt, M. (1983). An auditory speaker-recognition experiment comparing the performance of trained phoneticians and phonetically naive listeners. *Papers in Linguistics and Phonetics, 1,* 115–117.

Siegman, A. W. (1987). The telltale voice: Nonverbal messages of verbal communication. In A. W. Siegman & S. Feldstein (Eds.), *Nonverbal behavior and communication* (pp. 351–434). Hillsdale, NJ: Erlbaum.

Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *Journal of the Acoustical Society of America, 64,* 81–87.

Strongman, K., & Woosley, J. (1967). Stereotyped reactions to regional accents. *British Journal of Social and Clinical Psychology, 6,* 164–167.

Tartter, V. C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perception & Psychophysics, 49,* 365–372.

Taylor, H. C. (1934). Social agreement on personality traits as judged from speech. *Journal of Social Psychology, 5,* 244–248.

Thompson, C. P. (1985a). Voice identification: Speaker identifiability and a correction of the record regarding sex effects. *Human Learning, 4,* 19–27.

Thompson, C. P. (1985b). Voice identification: Attempted recovery from a biased procedure. *Human Learning, 4,* 213–224.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1,* 121–131.

Titze, I. R. (1994). *Principles of voice production.* Englewood Cliffs, NJ: Prentice-Hall.

Van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language and Speech, 33,* 259–272.

Van Dommelen, W. A. (1993). Speaker height and weight identification: A re-evaluation of some old data. *Journal of Phonetics, 21,* 337–341.

Van Lancker, D., Cummings, J., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24,* 195–209.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia, 25,* 829–834.

Van Lancker, D., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11,* 665–674.

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

Van Lancker, D., Kreiman, J., & Wickens, T. D. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics, 13,* 39–52.

Voiers, W. D. (1964). Perceptual bases of speaker identity. *Journal of the Acoustical Society of America, 36,* 1065–1073.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech and Hearing Research, 21,* 265–275.

Walton, J. H., & Orlikoff, R. F. (1994). Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech and Hearing Research, 37,* 738–745.

White, S., White, R., & Thorpe, W. (1970). Acoustic basis for individual recognition by voice in the gannet. *Nature (London), 225,* 1156–1158.

Williams, C. E. (1964). *The effects of selected factors on the aural identification of speakers* (Section III of Report ESD-TDR-65-153). Hanscom Field, MA: Air Force Systems Command, Electronics Systems Division, Hanscom AFB.

Winograd, E., Kerr, N., & Spence, M. (1984). Voice recognition: Effect of orienting task, and a test of blind versus sighted listeners. *American Journal of Psychology, 97,* 57–70.

Yarmey, A. D. (1986). Verbal, visual, and voice identification of rape suspect under different levels of illumination. *Journal of Applied Psychology, 71,* 363–370.

Yarmey, A. D. (1991). Voice identification over the telephone. *Journal of Applied Social Psychology, 21,* 1868–1876.

Yarmey, A. D. (1993). Stereotypes and recognition memory for faces and voices of good guys and bad guys. *Applied Cognitive Psychology, 7,* 419–431.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1,* 792–816.

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology, 6,* 367–377.

Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology, 8,* 453–464.

6

Talker Normalization

Phonetic Constancy as a Cognitive Process

HOWARD NUSBAUM
JAMES MAGNUSON

6.1 LACK OF INVARIANCE AND THE PROBLEM OF PHONETIC CONSTANCY

Human listeners recognize and understand spoken language quite effectively regardless of the vocal characteristics of the talker, or how quickly the speech is produced, or what the talker has said previously. Even at the most basic level of recognizing spoken consonants and vowels, most humans have little difficulty maintaining phonetic constancy—stable recognition of the phonetic structure of utterances (Shankweiler, Strange, & Verbrugge, 1977) in spite of variation in the relationship between the acoustic patterns of speech and phonetic categories that results from these sources of variability (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Indeed, the perceptual ability of human listeners has still not been matched in engineering efforts to develop computer speech-recognition systems.

Furthermore, even after more than 30 years of scientific endeavor, there are no theories of speech perception that can adequately explain how humans recognize spoken consonants and vowels (see Nusbaum & Henly, in press). Although the theoretical problem posed by the lack of invariance in the relationship between linguistic categories and their acoustic manifestations in the speech signal has been attacked from a number of different perspectives, such as the use of articu-

Talker Variability in Speech Processing