Published in final edited form as:
Psychol Sci. 1994 January 1; 5(1): 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x.

SPEECH PERCEPTION AS A TALKER-CONTINGENT PROCESS

Lynne C. Nygaard, **Mitchell S. Sommers**, and **David B. Pisoni**
Indiana University

Abstract

To determine how familiarity with a talker's voice affects perception of spoken words, we trained two groups of subjects to recognize a set of voices over a 9-day period. One group then identified novel words produced by the same set of talkers at four signal-to-noise ratios. Control subjects identified the same words produced by a different set of talkers. The results showed that the ability to identify a talker's voice improved intelligibility of novel words produced by that talker. The results suggest that speech perception may involve talker-contingent processes whereby perceptual learning of aspects of the vocal source facilitates the subsequent phonetic analysis of the acoustic signal.

During the perception of speech, listeners must extract stable phonetic percepts from acoustic signals that are highly variable. Variations in talker characteristics, in particular, have been shown to produce profound effects on the acoustic realization of speech sounds (Nearey, 1978; Peterson & Barney, 1952). Traditionally, models of speech perception have characterized variation in the acoustic speech signal as a perceptual problem that perceivers must solve (Shankweiler, Strange, & Verbrugge, 1976). Listeners are thought to contend with variation in speech signals due to talker differences through a compensatory process in which speech sounds are normalized with reference to specific voice characteristics. According to a strict interpretation of this view, information about a talker is stripped away during the perception of speech to arrive at the abstract, canonical linguistic units that are presumed to be the basic building blocks of perception (Halle, 1985; Joos, 1948; Summerfield & Haggard, 1973).

Unfortunately, this standard view of talker normalization begs the question of how the processing of a talker's voice is related to the perception of the phonetic content of speech. Although a talker's voice carries important information about the social and physical aspects of that talker into the communicative setting (Laver & Trudgill, 1979), the encoding of voice-specific information for the identification and discrimination of talkers has generally been considered to be a problem quite separate from apprehending the linguistic content of an utterance. On the one hand, researchers have investigated the ability of listeners to explicitly recognize and discriminate familiar and unfamiliar voices (e.g., Legge, Grossmann, & Pieper, 1984; Van Lancker, Kreiman, & Emmorey, 1985). In this case, the speech signal is viewed simply as a carrier of talker information. On the other hand, research in speech perception has been devoted to studying the linguistic content of speech—either entirely independently of any variability in talker or source characteristics or from the point of view that variation due to changes in talkers is noise that must be normalized or discarded quickly in order to recover the linguistic content of an utterance. Consequently, the emphasis in speech perception research has been on identifying and defining short-term, presumably

Address correspondence to Lynne C. Nygaard, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405; lnygaard@indiana.edu.

automatic adaptations to differences in source characteristics (Garvin & Ladefoged, 1963; Johnson, 1990; Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989).

The theoretical and empirical dissociation of the encoding of talker characteristics and the processing of the phonetic content of an utterance assumes that the analysis of these two kinds of information is independent (Laver & Trudgill, 1979). Only recently has this assumption been questioned on the basis of a growing body of research demonstrating effects of talker variability on both perceptual (Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Summerfield & Haggard, 1973) and memory (Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989) processes. For example, using a continuous recognition memory procedure, Palmeri, Goldinger, and Pisoni (1993) recently found that specific voice information was retained in memory along with item information, and these attributes were found to aid later recognition memory. These findings suggest that talker information may not be discarded in the process of speech perception, but rather variation in a talker's voice may become part of a rich and highly detailed representation of the speaker's utterance.

Although previous experiments have demonstrated that short-term adjustments may occur in the analysis of speech produced by different talkers (Ladefoged & Broadbent, 1957) and that talker information may be retained in long-term memory, the question remains whether the talker information that is retained in memory has any relationship to the ongoing analysis of linguistic content during the perception of speech. The purpose of the present experiment was to address this question by determining if differences in a listener's familiarity with a vocal source have any effect on the encoding of the phonetic content of a talker's utterance. To accomplish this, we asked two groups of listeners explicitly to learn to recognize the voices of 10 talkers over a 9-day period. At the end of the training period, we evaluated the role of talker recognition on the perception of spoken words to determine if the ability to identify a talker's voice was independent of phonetic analyses. It should be noted that independence between talker recognition and phonetic analysis is implicitly assumed by all current theoretical accounts of speech perception (Fowler, 1986; Liberman & Mattingly, 1985; McClelland & Elman, 1986; Stevens & Blumstein, 1978). If learning to identify a talker's voice is found to affect subsequent word recognition performance, the mechanisms responsible for the encoding of talker information would seem to be linked directly to those that underlie phonetic perception. Establishing such a link would require a fundamental change in present conceptualizations of the nature of mechanisms contributing to speech perception.

METHOD

Subjects

Subjects were 38 undergraduate and graduate students at Indiana University. Nineteen subjects served in each condition—experimental and control. All subjects were native speakers of American English and reported no history of a speech or hearing disorder at the time of testing. The subjects were paid for their services.

Stimulus Materials

Three sets of stimuli were used in this experiment. All were selected from a data base of 360 monosyllabic words produced by 10 male and 10 female talkers. Word identification tests in quiet showed greater than 90% intelligibility for all words. In addition, all words were rated to be highly familiar (Nusbaum, Pisoni, & Davis, 1984). The stimuli were originally recorded on audiotape and digitized at a sampling rate of 10 kHz on a PDP 11/34 computer

using a 12-bit analog-to-digital converter. The root mean squared (RMS) amplitude levels for all words were digitally equated.

Procedure

Training—Two groups of 19 listeners each completed 9 days of training to familiarize themselves with the voices of 10 talkers. Listeners were asked to learn to recognize each talker's voice and to associate that voice with one of 10 common names (see Lightfoot, 1989). Digitized stimuli were presented using a 12-bit digital-to-analog converter and were low-pass filtered at 4.8 kHz. Stimuli were presented to listeners over matched and calibrated TDH-39 headphones at approximately 80 dB SPL (sound pressure level).

On each of the 9 training days, both groups of listeners completed three different phases. The first phase consisted of a familiarization task. Five words from each of the 10 talkers were presented in succession to the listeners. Subjects then heard a 10-word list composed of 1 word from each talker in succession. Each time a token was presented to the listeners, the name of the appropriate talker was displayed on a computer screen. Listeners were asked to listen carefully to the words presented and to attend specifically to the talker's voice so they could learn the name.

The second phase of training consisted of a recognition task in which subjects were asked to identify the talker who had produced each token. The 100 words used did not overlap with those used in the first phase. Ten words from each of the 10 talkers were presented in random order to listeners who were asked to recognize each voice by pressing the appropriate button on a keyboard. The keys were labeled with 10 names. Keys 1 through 5 were labeled with male names; Keys 6 through 10 were labeled with female names. On each trial, after all subjects had entered their responses, the correct name appeared on the computer screen.

After subjects completed two repetitions of the first two phases of training, we administered a test phase on each day. As in the second training phase, 10 words from each of the 10 talkers were presented in random order. Subjects were asked to indicate who each speaker was by pressing on a keyboard the button corresponding to the appropriate name. However, feedback was not given.

Although the words used in the test phase were drawn from the same 100 words used in the second training phase, on each day of training subjects never heard the same item produced by the same talker in both the test and the training phase. In addition, training stimuli were reselected from the data base on each day so that subjects never heard the same word produced by the same talker in training. This training procedure was designed to expose listeners to a diverse set of tokens from each of the talkers.

Generalization—On the 10th day of the experiment, both groups of subjects completed a generalization test. One hundred new words produced by each of the 10 familiar talkers were used. As in the test phase used during training, 10 words from each of the 10 talkers were presented in random order. Subjects were asked to name the talker on each trial. No feedback was given. Thus, the generalization test was identical to the training test phase except that listeners had never heard any of the words before.

Word intelligibility—In addition to the generalization test, we administered a speech intelligibility test in which subjects were asked to identify words presented in noise. In this transfer task, 100 novel words were presented at either 80, 75, 70, or 65 dB (SPL) in continuous white noise low-pass filtered at 4.8 kHz and presented at 70 dB (SPL), yielding four signal-to-noise ratios: + 10, +5, 0, and −5. Equal numbers of words were presented at

each of the four signal-to-noise ratios. In this test, subjects were simply asked to identify the word itself (rather than explicitly recognize the talker's voice) by typing the word on a keyboard. Subjects in the experimental condition were presented with words produced by the 10 talkers they had learned in the training phase. Subjects in the control condition were presented with words produced by 10 new talkers they had not heard in the training phases.

RESULTS AND DISCUSSION

Training

Most subjects showed continuous improvement across the 9 days in their ability to recognize talkers from isolated words. However, individual differences were found in performance. Consequently, we selected a criterion of 70% correct for talker recognition on the last day of training for inclusion in the experiment. Our rationale for choosing this criterion was simply that to determine whether learning a talker's voice affects perceptual processing, we needed to ensure we had identified a group of subjects who did, in fact, learn to recognize the talkers' voices from isolated words. On the basis of this criterion, 9 subjects from each training group were included in the final analysis.[1] Both groups of listeners identified talkers consistently above chance even on the 1st day of training, and performance rose to nearly 80% correct by the last day of training. A repeated measures analysis of variance (ANOVA) with learning and days of training as factors showed a significant main effect of day of training, $F(9, 144) = 73.55$, $p < .0001$, but no difference between the two groups over days of training, $F(1, 16) = 0.14$, $p > .7$.

Generalization

The generalization test showed almost identical recognition of voices from novel words on the 10th day as on the final day of training. The implication of this result is that listeners acquired detailed knowledge about the talkers' voices that was not necessarily dependent on the specific words that carried that information. In other words, the perceptual learning that took place in the course of the nine training sessions was not dependent on the training stimuli but rather readily generalized to novel utterances produced by the same set of talkers.

Word Intelligibility

Figure 1 shows the percentage of correct word identification as a function of signal-to-noise ratio for both groups of trained subjects. As expected, identification performance decreased from the +10 to the −5 signal-to-noise ratio for both groups. However, subjects tested with words produced by familiar voices were significantly better in recognizing novel words at each signal-to-noise ratio than were subjects tested with unfamiliar voices.[2] A repeated measures ANOVA with training and signal-to-noise ratio as factors revealed highly significant main effects of both signal-to-noise ratio, $F(3, 48) = 173.27$, $p < .0001$, and experimental condition (experimental vs. control group), $F(1, 16) = 13.62$, $p < .002$.[3]

[1]It should be noted that the task of learning to identify voices from isolated words is extremely difficult (see Williams, 1964).Therefore, it was necessary to set a somewhat arbitrary training period and then select subjects who had learned to our criterion by the end of that period. Given additional training with isolated words or with sentences or larger passage of speech, a greater percentage of our subjects would have reached a criterion level of performance.

[2]It should be noted that the subjects who did not meet the criterion of 70% correct voice identification on Day 9 of training were also tested in the word recognition task. Among these "poor" learners, subjects who received words produced by talkers previously heard in training showed no advantage over subjects who received words produced by talkers not heard previously. This finding suggests that simple exposure to the voices heard in training was not sufficient for listeners to obtain the perceptual learning necessary for improved word recognition ability.

[3]Four items from the control condition were eliminated from the overall analyses. After the experiment had been run, these items were found to be mispronounced.

To ensure that the overall intelligibility of the two sets of voices did not differ, two additional groups of 18 untrained subjects who were not familiar with either set of talkers were given the same word intelligibility test. One untrained control group received the stimulus tokens produced by the talkers who were used in the training phase; the other untrained control group received the stimulus tokens from the talkers who were presented to the trained control group in the intelligibility test. Identification performance for the trained and untrained control groups did not differ. A separate repeated measures ANOVA including the two untrained and the one trained control conditions revealed a significant main effect of signal-to-noise ratio, $F(3, 102) = 221.38$, $p < .001$, but no significant main effect of control condition, $F(2, 34) = 0.16$, $p > .9$. This finding confirms that the difference in performance between the experimental group and the trained control subjects was not due to inherent differences in the intelligibility of the voices or the words used.

GENERAL DISCUSSION

The present study found that voice recognition and processing of the phonetic content of a linguistic utterance were not independent. Listeners who learned to recognize a set of talkers apparently encoded and retained in long-term memory talker-specific information that facilitated the subsequent perceptual analysis and identification of novel words produced by the same talkers. These findings provide the first demonstration that experience identifying a talker's voice facilitates perceptual processing of the phonetic content of that speaker's novel utterances. Not only does the perceptual learning that results from the talker recognition task generalize to the recognition of familiar voices producing novel words, but that learning also transfers to a completely different task involving the perceptual analysis of the phonetic content of novel words produced by the same talkers in a speech intelligibility test. Listeners who were presented with one set of voices but were tested with another set of voices failed to show any benefit from the experience gained by learning to recognize those voices explicitly. Only experience with the specific voices used in the intelligibility test facilitated the phonetic processing of novel words. The implication of this result is that phonetic perception and spoken word recognition appear to be affected by knowledge of specific information about a talker's voice. Experience with specific acoustic attributes of a talker's voice appears to facilitate the analysis of spoken words.

Our results are consistent with the view that the learning that occurs when listeners are trained to recognize and identify talkers' voices involves the modification of the procedures or perceptual operations necessary for the extraction of voice information from the speech signal (Kolers, 1976; Kolers & Roediger, 1984). That is, over time during training, listeners may learn to attend to and modify the specific perceptual operations used to analyze and encode each talker's voice during perception, and it is these talker-specific changes that are retained in memory. This procedural knowledge would then allow listeners to more efficiently analyze novel words produced by familiar talkers. We believe this situation may be very similar to the case of reading and remembering inverted text. Kolers and Ostry (1974) found that the operations necessary to read inverted text were retained in long-term memory and facilitated subsequent tasks involving reading inverted text. The type of detailed procedural knowledge that Kolers and Ostry described may be responsible for subjects' superior performance in identifying words spoken by familiar talkers in the present experiment.

The present findings demonstrate that the process that contends with variation in talker characteristics can be modified by experience and training with a specific talker's voice. The interaction of learning to identify a talker's voice and processing the phonetic content of a talker's utterance suggests that the speech perception mechanism is susceptible to general processes of perceptual learning and attention. Thus, the processing of a talker's voice may

demand time and resources if the voice is unfamiliar to a listener (Martin et al., 1989; Mullennix et al., 1989; Summerfield & Haggard, 1973), but may become much more efficient if the voice can be identified as familiar (Lightfoot, 1989). The fact that the speech-processing system is susceptible to such modification argues against a strictly modular view of phonetic processing (Fodor, 1983; Liberman & Mattingly, 1985). We find that encoding of a talker's voice interacts extensively with the analysis of spoken words. If word recognition were a separable process or module distinct from voice recognition, then training listeners to identify voices should have no effect on speech intelligibility. However, this experiment shows that learning to identify voices does facilitate perceptual analysis of words produced by those voices, indicating that encoding of voice characteristics and the perception of speech are highly integral processes that work together to perceptually organize the interleaved talker and linguistic information present in the acoustic signal.

Finally, this study provides the first direct demonstration of the role of long-term memory and perceptual learning of source characteristics in speech perception and spoken word recognition. The perceptual learning acquired through a task involving explicit identification and labeling of talkers' voices was found to transfer to an entirely different task involving the perception of the linguistic content of an utterance. It appears that the analysis involved when different voices are encountered in the perceptual process is not limited to short-term, online normalization, as supposed by most current theories of speech perception, but rather is a highly modifiable process that is subject to the perceptual learning of talker-specific information. Indeed, the present findings suggest that the phonetic coding of speech is carried out in a talker-contingent manner. Phonetic perception and spoken word recognition appear to be integrally related to knowledge of characteristics of a talker's vocal tract and, consequently, attributes of a talker's voice.

Acknowledgments

This research was supported by Training Grant DC-00012-13 and Research Grant DC-0111-16 from the National Institutes of Health to Indiana University. We thank Carol A. Fowler, Robert E. Remez, Peter D. Eimas, Scott E. Lively, and Stephen D. Goldinger for their helpful comments and suggestions on an earlier version of this manuscript.

References

Fodor, J. The modularity of mind. Cambridge, MA: MIT Press; 1983.

Fowler CA. An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics. 1986; 14:3–28.

Garvin PL, Ladefoged PL. Speaker identification and message identification in speech recognition. Phonetica. 1963; 9:193–199.

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

Halle, M. Speculations about the representation of words in memory. In: Fromkin, VA., editor. Phonetic linguistics. New York: Academic Press; 1985. p. 101-104.

Johnson K. The role of perceived speaker identity in F0 normalization of vowels. Journal of the Acoustical Society of America. 1990; 88:642654.

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

Kolers PA. Pattern analyzing memory. Science. 1976; 191:1280–1281. [PubMed: 1257750]

Kolers PA, Ostry DJ. Time course of loss of information regarding pattern analyzing operations. Journal of Verbal Learning and Verbal Behavior. 1974; 13:599–612.

Kolers PA, Roediger HL III. Procedures of mind. Journal of Verbal Learning and Verbal Behavior. 1984; 23:425–449.

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

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. 1-32.

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

Liberman AM, Mattingly IG. The motor theory of speech perception revised. Cognition. 1985; 21:1–36. [PubMed: 4075760]

Lightfoot, N. Research on Speech Perception Progress Report No. 15. Bloomington: Indiana University; 1989. Effects of talker familiarity on serial recall of spoken word lists.

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; 15:676–681.

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18:1–86. [PubMed: 3753912]

Miller JD. Auditorerceptual interpretation of the vowel. Journal of the Acoustical Society of America. 1989; 85:2114–2134. [PubMed: 2659639]

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

Mullennix JW, Pisoni DB, Martin CS. Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America. 1989; 85:365–378. [PubMed: 2921419]

Nearey, T. Phonetic features for vowels. Bloomington: Indiana University Linguistics Club; 1978.

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

Nusbaum, HC.; Pisoni, DB.; Davis, DK. Research on Speech Perception Progress Report No. 10. Bloomington: Indiana University; 1984. Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words.

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; 19:309–328.

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

Shankweiler, DP.; Strange, W.; Verbrugge, RR. Speech and the problem of perceptual constancy. In: Shaw, R.; Bransford, J., editors. Perceiving, acting, knowing: Toward an ecological psychology. Hillsdale, NJ: Erlbaum; 1976. p. 315-345.

Stevens KN, Blumstein SE. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America. 1978; 64:1358–1368. [PubMed: 744836]

Summerfield, Q.; Haggard, MP. Report on Research in Progress in Speech Perception No. 2. Belfast, Northern Ireland: Queen's University of Belfast; 1973. Vocal tract normalisation as demonstrated by reaction times.

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

Williams, CE. The effects of selected factors on the aural identification of speakers (Section III of Report EDS-TDR-65–153). Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command; 1964.

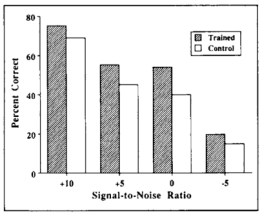

Fig. 1.

Mean intelligibility of words presented in noise for trained and control subjects. Trained, or experimental, subjects were trained with one set of talkers and tested with words produced by these familiar talkers. Control subjects were trained with one set of talkers and tested with words produced by a novel set of talkers. Percentage of correct word recognition is plotted at each signal-to-noise ratio.