1034

# 7

# Normalization of Vowels by Breath Sounds

**D. H. WHALEN**
**SONYA M. SHEFFERT**

*Last night, the phone rang just before midnight, and only half awake, she answered it. There was only breathing on the other end, but instead of hanging up she lay there in the dark, listening to it. She was sure it was Philippe, but did she know the sound of his breathing that well? Could the sound of one person's breathing be that recognizable? Is it like fingerprints—no two breaths are the same?* (Davis, 1994, p. 237)

## 7.1 INTRODUCTION

The acoustic realization of speech sounds depends greatly on the vocal tract that produces them. Adult speakers differ in their vocal anatomy, especially across the sexes, and every listener must take such differences into account. Additionally, every adult begins life as a child ("I started out as a child," as Bill Cosby used to say), and productive language begins as early as 1 year of age, well before the head has reached its full, adult size. So each speaker must be able to normalize his or her parents' speech in order to know what they are supposed to learn. And one's own speech must be normalized for the current size of the vocal tract, not to the size during the early stages of acquisition. Given such variation, humans have had to make use of a communicative system that allows listeners to make compensations for changes in speakers. Listeners clearly do make adjustments, most clearly in the case of vowels (Nearey, 1989), but also with stops (Rand, 1971) and fricatives (May, 1976; Strand & Johnson, 1996). Thus speaker normalization must be considered a feature of speech perception.

*Talker Variability in Speech Processing*

The sources of normalization have been found to be both intrinsic to the signal and extrinsic to it. Intrinsic factors include the ratio of the formants (F) (Chiba & Kajiyama, 1941; Potter & Steinberg, 1950), F0 of the vowel (Miller, 1989; Syrdal & Gopal, 1986), the spacing of F3 and F4 (Fujisaki & Kawashima, 1968), and the shape of the upper range of the spectral envelope (Kitamura & Akagi, 1994). Extrinsic factors include a speaker's overall range of formant values (Gerstman, 1968), presentation of point vowels (Ainsworth, 1975), the sentence preceding the item to be identified (Broadbent & Ladefoged, 1960; Remez, Rubin, Nygaard, & Howell, 1987), and even the assumed sex of the speaker as given visually (Strand & Johnson, 1996). As Nearey (1989) pointed out in his overview of this issue, the existence of intrinsic effects has often led to the insistence that only intrinsic effects should be considered, despite the experimental evidence of extrinsic effects. This tendency is heightened by the fact that the influence of extrinsic factors may be limited to ambiguous stimuli, in light of the lack of an effect of speaker information on error rates in identifying naturally produced vowels (Verbrugge, Strange, Shankweiler, & Edman, 1976). Intrinsic methods of normalization also have the advantage for computational treatments that they can be applied to any stretch of the speech signal without taking the rest into account. But the evidence is fairly clear that extrinsic factors must play a role, and so some combination of extrinsic and intrinsic normalization must occur.

A further question that remains unresolved is whether normalization is carried out on every vowel (and perhaps on accompanying consonants) or if the listener sets up a working model of the speaker in order to interpret what is heard. Recent evidence from short-term memory (STM) studies shows that it is more difficult to remember lists spoken by several speakers than those spoken by a single speaker (Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989). One possible interpretation of this effect is that the need to construct models of the different speakers is more time-consuming than constructing one for a single speaker, and thus the resources available for the memory task are reduced. More explicit testing for such "model construction" is necessary, including testing whether intrinsically ambiguous vowels (i.e., ones that cannot be classified without knowing which vocal tract produced them) are treated unambiguously in a single-speaker context, and whether the model can be retained for some length of time and reactivated either automatically or by explicitly cuing. The alternative to model construction is a more general set of context effects, in which speech is adapted to its current environment but in a way that relates various acoustic properties to each other, not to a proposed speaker. Such models would not predict that any gain would accrue from earlier exposure to a speaker.

Many of the normalization effects rely on features of the acoustics that are not available to consciousness on the part of the listener. The intensity differences due to vowel quality, for example, are fairly automatically taken into account in perception (Lehiste, 1970). Other features, such as speaker sex, are fairly easily reported, even if the exact effect of the normalization is still unavailable for listener report. If the model construction approach is correct, we would expect

that it would be easy to manipulate the level of consciousness attainable by manipulating how much the listener "knows" about the speaker, including visual appearance, name, and "personal" facts (however fictitious!). Again, these experiments remain to be done, but there is an interesting discrepancy between what listeners can report and what they can actually do.

One interesting example of this discrepancy between performance and report is in the judgment of speaker height and weight based only on acoustic speech information. Listeners are quite willing to perform this task, and generally report a moderate level of confidence in their answers (Lass, Phillips, & Bruchey, 1980). Even novelists have noted the expectation of a correlation: "She walked past Sara with a quick 'Hi,' hugged Anthony, and said, 'Hello, darling,' in a voice that had more height than she did" (Davis, 1994, p. 263). However, subjects are terrible at the task. One reason for that poor performance, it turns out, is an almost complete lack of correlation between any measurable speech characteristic and height or weight (Künzel, 1989; Van Dommelen, 1995). The basis for this confidence is fairly clear. There are physical reasons to expect that large bodies will have large resonances and lower pitches, given the constraints on vibrating bodies. So there are ample opportunities for human listeners to associate low sounds with big bodies and vice versa. Within the species *Canis familiaris,* for example, there is a great tendency for large breeds (e.g., German shepherds) to have low fundamentals, while the smaller breeds (e.g., Chihuahuas) have high ones. However, those examples lead listeners astray in two ways. First, the size of the human body does not seem to be terribly well correlated with the size of the vocal tract, and especially so in the case of the vocal folds themselves. Although it is true that heavier folds (as seen, for example, in postpubescent males) result in a lower F0, the weight of the folds does not seem to correlate with body weight by itself. Nature is full of examples of creatures that have exploited the apparent connection between F0 and size to allow some species (certain frogs and toads, for example) to sound much larger than they are. The fact that this trick works is evidence of its abiding appeal to perceptual systems, even our own, in which we can learn from experience that the association is a weak one. We are fooled by the apparent link, and we continue to think that we can judge body size by voice quality.

We normalize successfully, even when we are misled about body size. This is due to the fact that vocal tract size is what is important, and we are better attuned to that feature than to other, irrelevant features. Vocal tract normalization is so important, in fact, that we must assume its operation for every act of speech perception. Speech perception itself has been found to be sensitive to virtually every acoustic property that covaries with speech production (Liberman & Mattingly, 1985; Lisker, 1986; Repp, 1982; Studdert-Kennedy, 1976). If speech perception makes use of all available information, as it seems to, and normalization is a normal part of speech perception, then we would expect that it too would make use of all information available, even if it is somewhat unusual.

With that thought in mind, we set out to see whether normalization could be affected by a nonspeech sound that nonetheless is a typical component of the

speech process, namely, the sound of inspiration. Because speech is generated primarily by controlling the outgoing breath stream, it follows that inspiration is a necessary precondition to speaking. Sometimes that inspiration will be silent, or nearly so. But quite often it is easily audible. We have already found (Whalen, Hoequist, & Sheffert, 1995) that inclusion of such breath sounds can enhance the memorability of synthetic speech, although the exact mechanism of that improvement is unclear. One possibility is that the inspiration helps to normalize the speech, and thus makes the maintenance of the model of the speaker that much easier. However, we currently have no direct evidence that inspiration noise provides any information for normalization. The present study was designed to produce some of that evidence.

We looked for normalization effects of natural inspiration noises preceding synthetic utterances that formed two vowel continua. We made use of the speaker characteristic that has provided the most consistent normalization effects, namely, speaker sex. Would male and female breaths shift the boundary between the synthetic vowels? Additionally, we tested whether there is coarticulatory information in the inspiration noise. The noise is presumably generated primarily at the larynx and at various points in the lungs themselves, but it propagates through the vocal tract and thus will be shaped by it. If there is anticipation of the upcoming vowel in the vocal tract during the inspiration, then it may provide direct coarticulatory evidence of the vowel to be produced. If so, listeners might be able to use that information and so shift the boundary in the direction of the vowel that the breath was originally produced with.

## 7.2 EXPERIMENT

The experiment made use of natural inspiration noises, placed as a precursor to synthetic syllables (similar to Broadbent & Ladefoged, 1960). Two continua were used so that a variety of contexts could be tested.

### 7.2.1 Method

#### 7.2.1.1 Stimuli

The inspiration noises were produced by two native speakers of English, one male and one female (the second author). Each read a randomized list containing five repetitions of the words *bead, bad,* and *bud.* They were instructed to be sure that they took a breath through the mouth before each utterance. From these, a good exemplar was chosen. They were roughly equivalent in duration and amplitude. Spectrograms of all six are presented in Figure 1. Although the spectral characteristics are quite weak, given the source of the noise, there are noticeable differences among the breaths. The male noises have a lower overall frequency than the female ones. The vowel context did not seem to give rise to different

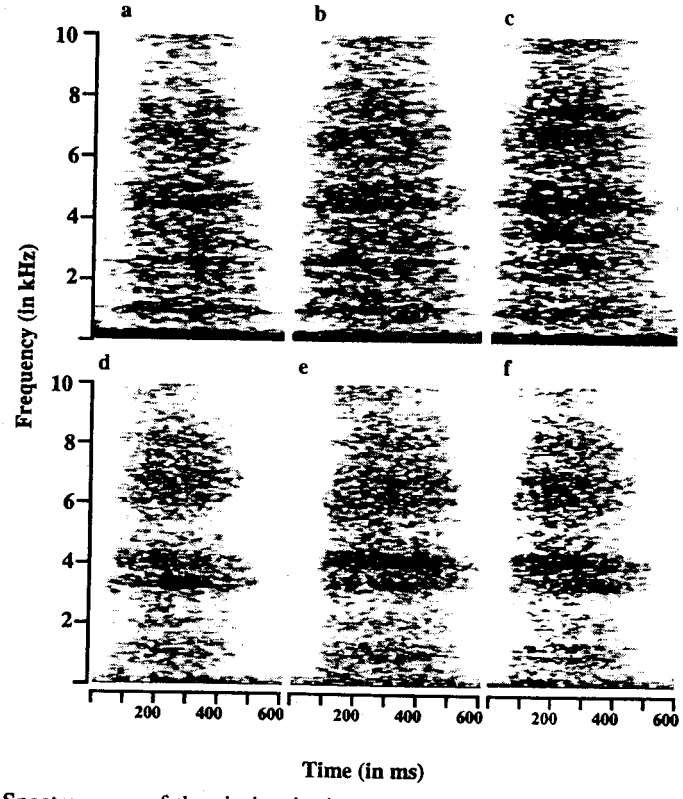


**FIGURE 1** Spectrograms of the six inspiration noises used in the experiment. Panels (a–c) are for the male speaker, (d–f) for the female speaker. Panels (a) and (d) are from the "bud" utterances, (b) and (e) from the "bad" utterances, and (c) and (f) from the "bead" utterances.

resonances, but there were differences in the relative amplitude of the two main resonances, as can be seen in panels d and e of Figure 1.

We also included a speech precursor, namely, the phrase "And the next word is _____." One token of this phrase from each speaker was selected.

The synthetic speech stimuli were created on the serial resonance synthesizer SYN, written by Ignatius Mattingly. There were two continua, one going from /bæd/ to /bəd/ and the other going from /bɪd/ to /bɛd/. Formant values were intermediate between those of the male and the female speakers, as measured from their productions. (We also collected the speakers' productions of *bid* and *bed* for just this purpose.) For the /æ/ə/ continuum, F1 began at 300 Hz and F3 began at 2050 Hz. We found that shifting the onset of F2 along with the steady state value improved the intelligibility of the token. The onsets are listed in Table I. There was a short release burst simultaneous with the first 10–20 ms of the syllable, centered at 2000 Hz. The formant center frequencies, listed in Table I, were attained after 50 ms. There was a steady state of 250 ms, after which F1 changed linearly to 350 Hz and F2 changed to 1700 Hz. We had better results with a variable F3 ending frequency; those are listed in Table I. There was a small

**TABLE I** Formant Values for the æ/ə Continuum[a]

| Vowel | Cont. Number | F1 | F2 Onset | F2 Steady state | F3 Steady state | F3 Offset |
|---|---|---|---|---|---|---|
| ə | 1 | 640 | 850 | 1450 | 2850 | 2900 |
| | 2 | 660 | 900 | 1530 | 2830 | 2825 |
| | 3 | 685 | 975 | 1610 | 2810 | 2750 |
| | 4 | 705 | 1015 | 1740 | 2790 | 2675 |
| æ | 5 | 725 | 1100 | 1875 | 2775 | 2600 |

[a]F, formant.

burst for the final stop simultaneous with the closure. Such "closure bursts" are seldom commented on in the literature, but are readily visible in the speech waveform for many utterances. Inclusion of a very weak closure burst greatly improved the perception of the final stop. F0 was a constant 160 Hz for the first 250 ms, then it fell linearly to 147 Hz. This value is midway between the F0 of the male and that of the female speaker.

For the /ɪ/ϵ/ continuum, the formants began at 250, 1200, and 2275 Hz for F1, F2, and F3 respectively. F3 was the same for all five stimuli, but it did rise from 2500 Hz to 2550 Hz during the "steady state" portion of the syllable (50–250 ms). There was a release burst simultaneous with the first part of the transitions, centered at 2000 Hz and lasting 10–20 ms. The steady state values for F1 and F2 are given in Table II. For the final stop, the transitions lasted 50 ms and ended at 175, 1850, and 2400 Hz for F1, F2, and F3 respectively. There was a small "closure" burst at the end of the syllable. As before, F0 was a constant 160 Hz for the first 250 ms, then it fell linearly to 147 Hz.

Because these formant and F0 values were based on the average of the actual formant and F0 values used by our two speakers, the sex of the synthetic "speaker" was maximally ambiguous, which would presumably allow the speaker information in the inspiration noise to have an effect. It would not be

**TABLE II** Formant Values for the Steady States of F1 and F2, ɪ/ϵ Continuum[a]

| Vowel | Cont. # | F1 | F2 |
|---|---|---|---|
| ɪ | 1 | 470 | 2280 |
| | 2 | 490 | 2260 |
| | 3 | 510 | 2240 |
| | 4 | 530 | 2220 |
| ϵ | 5 | 550 | 2200 |

[a]F, formant.

surprising if relative weak information, such as that presumably contained in the breath sound, would be unable to overcome more robust cues in the speech itself.

#### 7.2.1.2 Procedure

Each of the five members of each continuum were combined with each of the precursors. For the /æ/ə/ continuum, the breaths produced before *bad* and *bud* were used. For the /ɪ/ɛ/ continuum, the breaths produced before *bead* and *bad* were used. The first case is therefore one in which the vowels at the endpoints of the continua were identical to the ones produced with the inspiration noise. The second case is one in which the vowels produced with the breath were more peripheral in the vowel space than the vowels of the test items. In this way, we might perhaps see a coarticulatory effect in the exaggerated case even if we did not find one in the case in which the exact items were used. The full speech precursors (male and female versions of "And the next word is _____") were the same for both continua.

The continua were recorded on digital audio tape (DAT) and presented to one subject at a time over headphones. Ten repetitions of each combination of synthetic syllable and precursor were randomized and presented for identification as *bad* and *bud* or, for the other continuum, *bid* or *bed.* There was a 200-ms pause between the precursor (speech or breath) and the syllable, and a 2.5-s pause in which the answer was written. There was a 5-s pause after every ten items, corresponding to the end of a line on the answer sheet. The order of presentation of the two continua was randomly selected for each listener.

#### 7.2.1.3 Subjects

The subjects were 15 undergraduate students at the University of Connecticut, who received course credit for their participation, along with seven colleagues from Haskins Laboratories and the two authors. All were native speakers of English with no reported hearing problems.

### 7.2.2 Results

Our expectations for the directions of the shifts in identification were as follows. For the coarticulatory information in the breaths, we predicted that the identification would shift toward the vowel that the breath had been produced with. This was the same vowel of the response in the /æ/ə/ continuum, and the vowel that was nearest the endpoint vowel for the /ɪ/ɛ/ continuum (that is, /i/ for /ɪ/, and /æ/ for /ɛ/). Our expectations for the effects of speaker sex on the responses was based on the changes in the significance of F1 for the two speakers. With the smaller vocal tract, the female speaker should have higher F1s than the male. Thus this should lead to more /ə/ responses for a female breath or precursor in the /æ/ə/ continuum and more /ɪ/ responses for a female breath or precursor in the /ɪ/ɛ/ continuum.

**TABLE III** Percent "Bud" Responses to the /æ/ə/ Continuum[a]

| | Speech | /æ/ Breath | /ə/ Breath | Marginal | Without speech |
|---|---|---|---|---|---|
| male | 51.4 | 53.9 | 55.4 | 53.6 | 54.7 |
| female | 53.0 | 51.2 | 54.1 | 52.8 | 52.7 |
| marginal | — | 52.6 | 54.8 | — | — |

[a]The prediction was that the female speaker and the /e/ breath would elicit more "bud" responses.

The mean percentage of *bud* responses (across the whole continuum) to the /æ/ə/ continuum are shown in Table III. The mean percentage of "bid" responses (across the whole continuum) to the /ɪ/ɛ/ continuum are shown in Table IV. The effects were assessed with a series of analysis of variance (ANOVA), using the percentage of responses across the whole continuum. (With a small number of continuum steps and a relatively small number of judgments per step, our experience has been that PROBIT analysis overemphasizes random variation at the endpoints of the continuum.) For each continuum, three analyses were performed. The first, main analysis, included both the speech precursors and the breath precursors, with factors Sex of Speaker (two levels, male and female) and Precursor (three levels, speech and two phonetic contexts of breath). Because the breaths include two effects (speaker normalization and coarticulation), we can expect that this factor will need further elaboration without regard to the speech precursor. Thus the second analysis examined only the breath precursors, with the factors Sex and Precursor (two levels each). The third analysis was a one-way examination of Sex for the speech precursors.

For the /æ/ə/ continuum, Sex was not significant (n.s.) in the main analysis ($F(1,23) < 1$, n.s.). The Precursor factor just missed significance ($F(2,46) = 3.11$, $p = .054$), and the interaction was not significant ($F(2,46) = 2.07$, n.s.). In the analysis of just the breaths, Sex was again not significant ($F(1,23) = 2.38$, n.s.). The Precursor factor was significant ($F(1,23) = 6.30$, $p < .05$), while the interaction was not ($F(1,23) < .1$, n.s.). As can be seen from Table III, the /ə/ breaths elicited a significant 2.2% more *bud* responses, indicating that the coarticulatory information was effective in shifting the perception of these ambiguous synthetic

**TABLE IV** Percent *Bid* Responses to the /ɪ/ɛ/ Continuum[a]

| | Speech | /æ/ Breath | /i/ Breath | Marginal | Without speech |
|---|---|---|---|---|---|
| Male | 52.4 | 44.6 | 43.3 | 46.8 | 44.0 |
| Female | 51.5 | 45.1 | 48.0 | 48.2 | 46.6 |
| Marginal | — | 44.9 | 45.7 | — | — |

[a]The prediction was that the female speaker and the /i/ breath would elicit more "bid" responses.

vowels. In the analysis of just the speech precursors, Sex was not significant ($F(1,23) < .1$, n.s.). Thus the speaker characteristic of sex did not influence the perception of these stimuli.

For the /ɪ/ɛ/ continuum, Sex was not significant in the main analysis ($F(1,23) = 1.49$, n.s.). The Precursor factor was significant ($F(2,46) = 9.96$, $p < .001$), perhaps due to the overall difference between the speech and breath precursors. The interaction just missed significance ($F(2,46) = 3.05$, $p = .057$). In the analysis of the breaths alone, Sex was significant ($F(1,23) = 4.29, p < .05$). The Precursor factor was not significant ($F(1,23) < 1$, n.s.), whereas the interaction was marginal ($F(1,23) = 3.00$, $p < .10$). As can be seen from Table IV, the female breaths elicited 2.6% more *bid* responses than the male breaths, an effect in the direction predicted by the likely effect of normalizing for the speaker's vocal tract. Although the /i/ breaths elicited 0.8% more *bid* responses, as we would expect on coarticulatory grounds, this difference was not significant. In the analysis of just the speech precursors, Sex was not significant ($F(1,23) = 1.95$, n.s.), despite a 2.6% difference in the predicted direction. Thus the speaker characteristic of sex did not influence the perception of these stimuli when the precursor was speech, but it did do so when the precursor was an inspiration noise.

After the completion of the experiment, it seemed of interest to determine how much of the information in the breath sounds was consciously available to listeners. To test this, we ran a small study with eight members of the Speech Research Laboratory of Indiana University. They had not participated in the main experiment. These subjects listened to 10 repetitions of each of the six breath sounds to see whether they could identify the vowel it had preceded. We blocked the breaths into the same pairs that had been used in the perception test, so that subjects judged either /i/ or /æ/, or /ə/ or /æ/. Subjects were 49.5% correct on the first comparison, and 47.5% correct on the other. Neither of these figures differs from the chance level of 50%. Four of the subjects also made judgments of speaker sex. These judgments were 98.3% correct for both sets of breaths.

## 7.3 DISCUSSION

Although the sound produced by inspiration is nonlinguistic, it is nevertheless shaped by the vocal tract. Thus such breath sounds have the potential to provide information about the dimensions and configuration of that vocal tract. In the present experiment, we have found evidence that such information is used in the perception of synthetic speech, leading to the expectation that this information is available in more natural situations. Whether there will be instances in which the contribution of such sounds is apparent is a different question. But these results raise the possibility that they might be used in automatic speech recognition as a converging source of information about the speaker's characteristics. Additional research will be needed in order to determine the conditions that result in the contribution of such sounds.

The degree of conscious awareness of the information seemed to play no role in the effectiveness of that information. Speaker sex was readily perceived from the inspiration sounds, whereas vowel quality was not. Yet speaker sex influenced the judgments in one continuum, whereas coarticulatory vowel information affected the other. This indicates that the effect of the breath sounds is unlikely to be a strategic one, as might arguably be the case in the example of the specification of speaker sex by vision alone (Strand & Johnson, 1996).

Our findings suggest that the coarticulatory information in the breaths may need to match exactly the vowels that are to be perceived. In our data, the coarticulatory influence was present in the continuum in which the breaths were produced with just the vowels that constitute the endpoints of the continuum. In the other continuum, we used more peripheral vowel articulations, in the hopes that this might exaggerate the influence that the coarticulatory information would have. For that case, however, there was no significant influence of that information. Exactly appropriate information is perhaps more easily used in this context. We are currently conducting an experiment to test this further.

The natural speech precursors did not affect the perception of the continua, in contrast to results found with synthetic precursor sentences (Broadbent & Ladefoged, 1960; Remez et al., 1987). We suspect that this lack of an effect is due to the discrepancy between the natural speech and the synthesis. The previous studies used synthesis for both the precursor and the target. This may make it easier for the listeners to integrate the two into a single, coherent percept. The clear difference between the natural speech and the synthesis in the present experiment, on the other hand, may have led to a lack of integration and thus to a lack of an effect. Various manipulations of the stimuli, such as synthesizing the precursor or creating a more natural-sounding synthetic continuum, would be needed to fully resolve this issue.

The fact that speaker information extrinsic to the signal, as shown here and in Strand and Johnson (1996), can influence the perception of speech clearly indicates that normalization procedures should not be exclusively intrinsically based. There is enough evidence now to indicate that the listener may construct a "model" of the speaker as a means of perceiving speech, given the large number of differences in vocal tract size that any listener will confront in a lifetime. Such "model making" leads to the prediction that individual characteristics outside the speech realm (such as a personal name, a habitual gait, or even a well-known belonging) could influence perception of speech in a way appropriate to a known individual. As an alternative to making a model of an individual speaker, extrinsic effects might result from all the information available about a vocal tract as selected from among a range of vocal tract types that are not associated with individuals. In that case, only speech information would be expected to play a role. The present study shows, at least, that the sound of inspiration must be included in the range of speech-relevant information available to the listener.

## ACKNOWLEDGMENTS

This research was supported by the National Institute of Child Health and Human Development (NICHD) grant HD-01994, to Haskins Laboratories. We thank Julia Irwin and Larry Brancazio for help with the experiments. Carol A. Fowler and Leigh Lisker provided helpful comments.

## REFERENCES

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 103–111). New York: Academic Press.

Broadbent, D. E., & Ladefoged, P. (1960). Vowel judgments and adaptation level. *Proceedings of the Royal Society of London, 151,* 384–399.

Chiba, T., & Kajiyama, M. (1941). *The vowel: Its nature and structure.* Tokyo: Tokyo-Kaiseikan Publishing Co.

Davis, P. (1994). *Bondage.* New York: Pocket Books [1995].

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16,* 73–77.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16,* 78–80.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 152–162.

Kitamura, T., & Akagi, M. (1994). Speaker individualities in speech spectral envelopes. *Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan,* Vol. 3, pp. 1183–1186.

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

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.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech, 29,* 3–11.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 676–684.

May, J. G. (1976). Vocal tract normalization for /s/ and /š/. *Haskins Laboratories Status Report on Speech Research, SR48,* 67–74.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Potter, R. K., & Steinberg, J. C. (1950). Towards the specification of speech. *Journal of the Acoustical Society of America, 22,* 807–820.

Rand, T. C. (1971). Vocal tract size normalization in the perception of stop consonants. *Haskins Laboratories Status Report on Speech Research, SR25/26,* 141–146.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 40–61.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92,* 81–110.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In *Proceedings of the 3rd KONVENS Conference.* Berlin, Mouton de Gruyter.

Studdert-Kennedy, M. (1976). Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 243–293). New York: Academic Press.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Van Dommelen, W. A. (1995). Speaker and listener sex for speaker height and weight identification. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. 738–741). Stockholm: Stockholm University.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, J. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America, 60,* 198–212.

Whalen, D. H., Hoequist, C. E., & Sheffert, S. (1995). The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America, 97,* 3147–3153.