Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report

Rochelle S. Newman[a)]
Department of Hearing and Speech Sciences, Program in Neuroscience and Cognitive Science, University of Maryland, Lefrak Hall, College Park, Maryland 20742

(Received 22 May 2001; accepted for publication 20 February 2003)

This paper examines whether correlations between speech perception and speech production exist, and, if so, whether they might provide a way of evaluating different acoustic metrics. The cues listeners use for many phonemic distinctions are not known, often because many different acoustic cues are highly correlated with one another, making it difficult to distinguish among them. Perception-production correlations may provide a new means of doing so. In the present paper, correlations were examined between acoustic measures taken on listeners' perceptual prototypes for a given speech category and on their average production of members of that category. Significant correlations were found for VOT among stop consonants, and for spectral peaks (but not centroids or skewness) for voiceless fricatives. These results suggest that correlations between speech perception and production may provide a methodology for evaluating different proposed acoustic metrics. © *2003 Acoustical Society of America.* [DOI: 10.1121/1.1567280]

PACS numbers: 43.70.Fq, 43.71.Es, 43.71.An [CWT]

I. INTRODUCTION

A great deal of research in speech perception has focused on the cues listeners use to distinguish different phonetic categories. Although the cues to some phonemic distinctions have been well specified (for example, VOT for voicing among stop consonants), the cues to other distinctions (such as place of articulation) are less clear.

One reason for this uncertainty is that the acoustic spectrum for many phonemes is quite complex, and the differences between spectra can therefore be described in a number of ways. Different alternatives are often highly correlated with one another, making it difficult to distinguish among them experimentally. For instance, Syrdal and Gopal (1986) have suggested that differences between formant peaks may be a cue to stop consonant place of articulation, whereas Sussman and colleagues (Sussman *et al.*, 1993, 1991) have suggested that the starting point of the second formant may be a cue by itself. Since both of these cues are based (at least in part) on the location of the second formant, changes in one cue almost necessitate changes in the alternative as well. Thus attempts to specify the acoustic changes to which listeners are sensitive often fail to differentiate between different proposed cues.

The present paper is an attempt to develop a new method of distinguishing between alternative cues, based on links between speech perception and speech production. Some phonemic distinctions can be articulated in multiple ways, with slightly different muscle movements (for example, see Johnson *et al.*, 1993b; Perkell and Matthies, 1992). Different people may articulate the same sound with different combinations of muscle and articulatory action. This could then influence what these individuals expect to hear from other speakers. Those individuals who produce a sound in a particular manner are likely going to judge others' productions according to the same metric. By examining which acoustic properties demonstrate these types of links between perception and production, it is possible to assess the likelihood that a particular cue is being used by a given listener.

There are a number of reasons to predict that these types of links should occur. Infants learn to speak their native language by hearing what other people produce. They must in some way associate the sounds they hear with the proper way of producing them, suggesting some basic sort of linkage between the systems (see, for example, Kuhl and Meltzoff, 1982). Moreover, since people are likely to have heard their own productions more than those of any other single individual, their productions are likely to have an especially important role in their perceptual prototypes. Thus, perceptual expectations should be skewed towards one's own productions, again suggesting at least an indirect link between the two systems.

Evidence for a more direct link comes from studies that have found that particular experiences in either perception or production often result in changes in the other modality as well (Bradlow *et al.*, 1997; Cooper, 1974; Cooper and Lauritsen, 1974; Cooper and Nager, 1975; Jamieson and Rvachew, 1992). For example, Bradlow *et al.* (1997) found that training Japanese speakers on perception of the English /ɹ/-/l/ distinction also resulted in improved production. Cooper (1974) found that after repeated presentation of /pi/, listener's productions of that syllable were more ''/bi/-like'' (that is, had shorter VOTs).

Several theories also make claims regarding the existence of links between perception and production. For example, motor theory (Liberman *et al.*, 1962, 1967; Liberman and Mattingly, 1985) argues that adults perceive speech by making reference to their own articulation. The authors even go so far as to claim that the word ''link'' really is not cor-

a)Electronic mail: rnewman@hesp.umd.edu

 0001-4966/2003/113(5)/2850/11/$19.00

rect, since it implies that speech perception and production ''though tightly bonded, are nevertheless distinct.'' Rather, they feel that ''for language, perception and production are only different sides of the same coin'' (Liberman and Mattingly, 1985, p. 30). Fowler's direct perception theory (Fowler, 1986) suggests that listeners directly perceive the gestures (or productions) of the speaker. Nearey's double weak theory (1992) also claims that the perceptual system has knowledge about relations between speech-production capabilities and the resulting acoustic output, which may require a link between the perception and production systems. Thus, these theories all suggest that there should be some connection between the two systems, although the strength of the predicted linkage varies between theories.

Experimental evidence for the existence of these perception-production correlations is somewhat mixed, however. Bell-Berti *et al.* (1979) found that there are two different manners of producing the tense-lax distinction among American English vowels, and that the strategy selected by different individuals (based on EMG data) was related to how those individuals performed in a perceptual task. Fox (1982) examined perceptual scaling data on vowels, and found that while three dimensions (representing tongue height, tongue ''frontness,'' and the presence of lip-rounding) were an adequate fit to listeners' data, the listeners differed in the weightings (or saliences) they gave to each dimension. Furthermore, there was a relationship between the weightings used by any given listener and acoustic measures of that listener's productions.

There has also been some evidence of correlations between perception and production of consonants, using cues such as VOT (Flege and Eefting, 1986; Hoffman *et al.*, 1984). These correlations appear to be limited to proficient speakers of the language (Flege, 1999), suggesting that they may be related to learning the appropriate pronunciation in the language.

Other studies have failed to find such perception-production correlations, however. For example, Bailey and Haggard (1973, 1980) failed to find a correlation between average produced VOTs for voiced and voiceless consonants and listener's category boundaries on a /g/-/k/ continuum. Ainsworth and Paliwal (1984) asked listeners to both produce English glides and identify synthetic tokens, and measured the F2 and F3 loci for these items. But they found that the variability within subjects was as high as that between subjects, arguing against perception-production links. Many of these failed attempts to find perception-production links have used relatively coarse-grained distinctions between stimuli (for example, Ainsworth and Paliwal, 1984; Bailey and Haggard, 1973, 1980; Paliwal *et al.*, 1983), or have averaged productions across different phoneme categories (such as labial and velar stop consonants; see Bailey and Haggard, 1973, 1980). Other studies have used relatively simple production measures, such as individual formants (Ainsworth and Paliwal, 1984; Frieda *et al.*, 2000; Paliwal *et al.*, 1983).

The variability in these results clearly demonstrates that correlations between speech perception and speech production are inconsistent. Finding such correlations appears to require not only a task that is sensitive to small variations in perception and production, but also an appropriate acoustic correlate as a production measure. In fact, there is no reason to expect correlations between perception and production unless the acoustic property being measured is one that is at least related to the cues actually used by listeners. The variability in previous research may suggest that these correlations could serve as a means of telling us something about the cues being investigated, specifically about the likelihood that listeners actually use those cues.

Another possible reason for this variability in findings is the hyperspace effect (Johnson *et al.*, 1993a). When asked to judge the best examples of a phonetic category, listeners often choose tokens with more extreme articulation than is common in fluent speech. Listeners' perceptual prototypes might better match exaggerated productions than typical ones. Although this would not necessarily eliminate a perception-production correlation, it would be likely to reduce it, making it more difficult to find significant results (especially with relatively insensitive tasks).

The present paper explores the feasibility of using perception-production correlations as a means of evaluating the appropriateness of speech production measures. Experiment 1 demonstrates the existence of perception-production correlations in a case where an appropriate cue is known. It focuses on VOT differences between voiced and voiceless stop consonants, an acoustic measure of voicing that has received substantial support in the literature (Lisker and Abramson, 1964, 1970). If correlations are not clearly apparent in this case, it would suggest either that our methodology is not sufficiently sensitive, or that these correlations vary across speakers. In either case, correlations could not be relied upon as a research tool. Given a significant correlation in experiment 1, experiment 2 then examines a phoneme for which there have been multiple proposed acoustic cues, with the goal of determining whether perception-production correlations can distinguish among related metrics.

II. EXPERIMENT 1

This experiment investigates whether correlations between individuals' perception of speech contrasts and their production of those contrasts can be found when an appropriate acoustic cue is used. Listeners participated in both a production and a perception task focusing on voice onset time (VOT), a result of laryngeal timing differences which are the primary cue to the voiced-voiceless distinction among stop consonants (Lisker and Abramson, 1964, 1970). The relationship between each individual's measures on the two tasks was examined.

The perception task was modeled on work by Miller and Volaitis (1989). In the present version of the task, listeners heard a VOT series ranging from /bɑ/ to /p/ɑ/ to something beyond a good /pɑ/ (labeled as */pɑ/, following Miller and Volaitis). These extreme stimuli sound like a very breathy ''pah,'' and have VOTs that are far longer than would normally occur in speech. Listeners were asked to rate the items for their goodness as members of the category /p/. Miller and Volaitis found that this task results in an orderly rating scale, with only one or a few items receiving the highest rating, and

ratings dropping off to either direction. The item with the highest mean rating was considered the listener's category prototype, and correlations between this perceptual prototype and the individual's production prototype—the average acoustic measure across a number of different productions—were examined.

These two tasks provide measures of each individual's perceptual prototype and average production, using the same acoustic measure (VOT). If correlations exist between perception and production within individuals, those individuals who produce /p/'s with longer VOTs would be expected to also rate items with longer VOTs as being better examples of the category than would individuals who produce tokens with shorter VOTs. Thus, correlations between each individual's perception and production measures would be expected.

A. Method

1. Subjects

The 25 paid participants were native speakers of English with no reported history of a speech or hearing disorder. They participated in two 1-h sessions. Data from two additional participants were dropped for being a non-native speaker ($n=1$) or for missing the second visit ($n=1$). During debriefing one of our listeners reported that he had misunderstood the instructions, and had identified whether the items were /p/'s or not, rather than rating their degree of goodness; his data were removed from analysis, as were that of a speaker whose highest rated item had a VOT more than 4 standard deviations beyond the mean of the other participants (221 ms). Data from three additional participants were removed because a central member of the /pɑ/ category could not be determined from their perceptual data, as discussed in the procedure below. Leaving out these listeners resulted in 20 participants for this experiment.[1]

2. Stimuli

To create models for our production task, a female native talker of English (RSN) recorded one token each of the 48 CV syllables formed from pairings of the English stop consonants (/p/, /b/, /t/, /d/, /k/, /g/) and the eight vowels /i, e, æ, u, o, ɔ, ɑ, ʌ/. These vowels represent the range of vowels in English which can occur in an open syllable. Two additional tokens of the syllable /pɑ/ (for a total of three) were recorded to provide a greater range of examples of the target syllable. All tokens were amplified, low-pass filtered at 9.5 kHz, and digitized via a 16-bit, analog-to-digital converter at a 20-kHz sampling rate.

Rather than create a synthetic speech series for our perceptual task (as did Miller and Volaitis, 1989), items from natural speech were edited in order to make the items as natural-sounding as possible. The same native speaker recorded the tokens /bɑ/, /pɑ/ and */pɑ/. A 21-item continuum ranging from /b/ to /p/ was created from the /bɑ/ base by removing successively longer sections from the /b/ onset and replacing them with the corresponding portions of the /p/ onset. Formant transitions in the original items were approximately 35 ms for /bɑ/ and 75 ms for /pɑ/, but the formants during the transitions were more broad for the /p/. This increase in bandwidth would have helped to mask any mismatches in formant values during these transitions; moreover, any such mismatches would have occurred only in the first two or three items in the series, none of which were within the range of prototypical /p/ tokens. Steady-state formants were at 950, 1350, and 3200 Hz for /b/, and at 1000, 1275, and 2950 Hz for /p/. Although these values are not identical, the formant frequencies are sufficiently close that cross-splicing did not result in sudden changes in formant values. All editing was done at zero crossings in the digital waveform to avoid audible clicks. The first stimulus was created by removing the /b/ release burst and replacing it with the release burst from /pɑ/. The second through twenty-first stimuli were each made by removing one additional vocal pulse from the onset of the /bɑ/ syllable, and replacing this with the equivalent duration of burst release and aspiration from /p/. Durations of vocal pulses were not exactly equal, but averaged 4.2 ms. The next 40 items were each generated by removing an additional 5 ms of aspiration from the */pɑ/ token and adding this to the end of the aspiration in the last item of the /b-p/ series (i.e., the 21st, or most ''p''-like item).

This resulted in a 61-item series, which would have been overly tedious for the participants. Pilot testing showed that most individuals placed their prototypes between 55 and 140 ms VOT (or between stimulus items 13 and 31). In order to maintain sensitivity to small differences between listeners, all stimuli within this range were included in the experiment. Beyond this range, every other stimulus was included in the experiment, and the remaining stimuli were removed. This resulted in a 40-item series, with VOTs ranging from 8.25 to 291 ms. Adjacent stimuli in the series differed in VOT by 4.6 ms at intermediate VOTs and by 9.4 ms at both longer and shorter VOTs.

3. Procedure

Listeners participated individually in both a production task and a perception task across two sessions; the production task occurred at the start of the first session.[2] The production task was an imitation task; pilot work suggested that when asked to read aloud written representations of syllables, talkers tend to speak progressively more quickly as recording continues. To encourage talkers to maintain a fairly even speaking rate, our participants listened to an example of each syllable over a loudspeaker and then repeated that syllable in the way they would normally produce it. This method of recording has previously been used by Forrest *et al.* (1988). Although it is possible that this method could induce listeners to mimic the acoustic characteristics of the model, the variability in participant's productions suggest this was not the case: average /pɑ/ productions ranged from VOTs of 51 to 125 ms.

The model stimuli for this production task were converted to analog from by a 16-bit, digital-to-analog converter at a 20-kHz sampling rate, low-pass filtered at 9.5 kHz, and presented in random order. Trials were repeated if productions were peak-clipped, if the participant failed to respond within 4 s, or if the participant indicated that he or she

wished to redo that trial (either because of uncertainty as to the target syllable, or because some other noise, such as a cough, interfered with recording). Listeners heard (and recorded) each of the 50 syllables in a single block, and participated in two such blocks. This resulted in two recordings of each CV syllable (and six tokens of the target item /pá/) to be used for later acoustic measurement.

Listeners then participated in the perceptual task. The stimuli for this task were converted to analog form in the same manner as above, and were presented binaurally through TDH-39 headphones at a comfortable listening level. Listeners heard the syllables in random order, and were asked to rate each initial phoneme for its goodness as a member of the category /p/. They responded using the numbers 0–9 on a numeric keypad. Listeners were instructed to use the ''0'' label whenever the item did not sound like a ''p'' at all, to use the ''1'' when they were unclear whether the item was a ''p'' or not, and to use the range ''2''–''9'' for items which were definitely members of the category ''p,'' but differed in how good of an example they were. Listeners were given a reference sheet containing this scale in case they wished to refer back to it. Responses from the first block of trials (one repetition of each item) were considered practice and were not included in subsequent data analysis. Listeners then participated in six test blocks (of two repetitions per item) in each of the two sessions, for a total of 24 responses to each stimulus.

4. Acoustic and perceptual measures

A mean perceptual rating was computed for each stimulus for each participant. The single item with the highest rating (regardless of where it occurred) was considered the listener's perceptual prototype, and that item's VOT was recorded. One participant had equally high ratings for two items in the continuum; the VOT values for these items were averaged as that listener's prototype. Listeners were excluded if they did not show at least a half of a ranking difference between their peak item and the final item in the series, suggesting there was no clear peak item.

For the production experiment, the time interval from syllable release to the onset of vocal pulsing was measured for each token produced by each speaker. The six values for the recordings of ''pa'' were averaged as the produced ''pa'' VOT. The values for the 14 other ''p'' recordings were averaged to find a mean VOT for the remaining ''p'' tokens. Likewise, the values for the 16 recordings for each of the other stop consonants were averaged, to determine its mean VOT. Prevoicing was ignored since this is a different cue from bursts/aspiration and it may be inappropriate to average across the two cues.

A second coder remeasured VOT for all 52 voiceless items for two participants for reliability purposes. Only voiceless items were considered because the large difference between voiced and voiceless items would have resulted in a high correlation across coders even had the VOT measures been relatively coarse; by restricting the range, we focus the correlation on the consistency within the category (as well as making it more difficult to find a high correlation in general). Despite this restricted range, correlations were 0.94 and 0.95 for the two talkers (average absolute differences were 3.5 and 5.3 ms, respectively), suggesting that VOT measures were quite reliable.

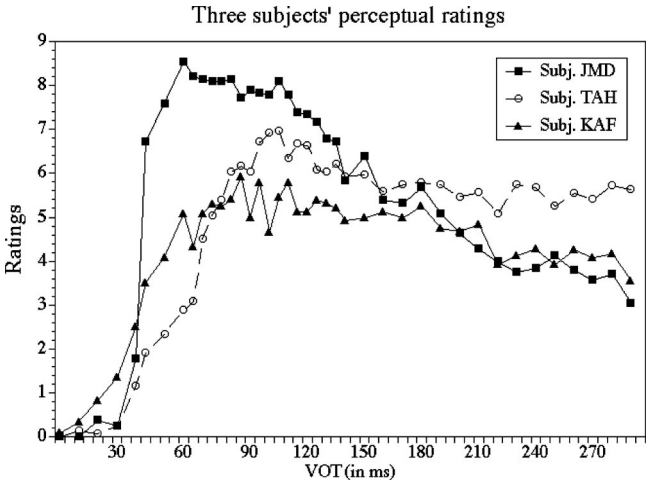

FIG. 1. Three subjects' perception data. Subject's ratings generally increase with increasing VOT until they reach their peak rating, then begin to decrease as the items sound more and more extreme.

B. Results and discussion

Listeners generally showed clear prototypes, with ratings dropping to either side, although they did vary in the number of items receiving high rankings. On average, listeners showed a rating drop-off of 2.5 units between their peak item and the series */pɑ/ endpoint; average rating on the /b/ endpoint was 0.27, suggesting it was not heard as a member of the /p/ category. Listeners' prototypes ranged from VOTs of 60.9 to 150.9 ms, suggesting that this perceptual measure is sensitive to differences between individuals. Figure 1 shows three listeners' perception data; these individuals were selected as demonstrating a range of prototype values (60.9, 88.0, and 105.3 ms) and drop-offs (5.5, 2.4, and 1.3 units).

The calculations resulted in one perceptual measure (VOT of the prototype), and seven production measures (average VOT for /pɑ/, average VOT for other /p/ items, and average VOT for /b/, /t/, /d/, /k/, and /g/ items). A stepwise hierarchical regression was performed using the perceptual measure as the dependent variable, and all seven production measures as independent variables. A hierarchical regression is less likely to capitalize on chance relationships than is a stepwise regression (Cohen and Cohen, 1983), but requires an *a priori* ordering of the IVs in terms of their likelihood of having a correlation with the DV. On the basis of phonological theory it was assumed that items differing from the target in one phonetic feature would be more closely related to the target than those differing from it in two features, and that items matching on the measure of interest (VOT; that is, /t/ and /k/) would be more closely related than the phoneme /b/, which matches on place of articulation. As alveolars tend to be more similar to bilabials than are velars (Dorman *et al.*, 1977; Klatt, 1975; Lisker and Abramson, 1964), alveolars were placed higher in the ordering. This resulted in the ordering /pɑ/, /p/, /t/, /k/, /b/, /d/, /g/.

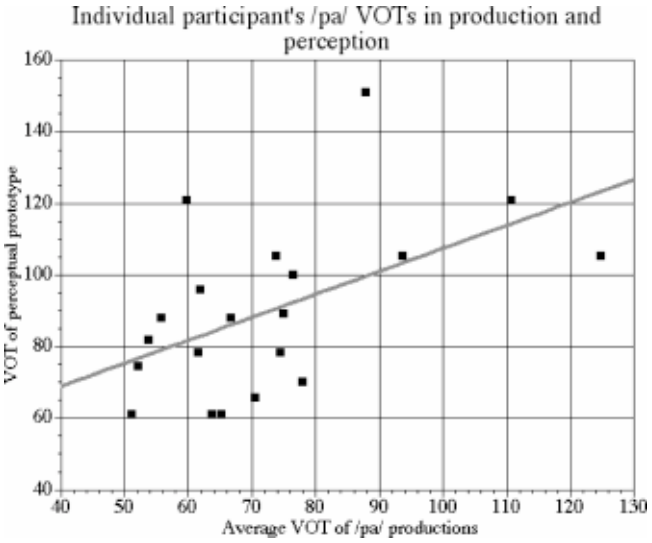

FIG. 2. Scatter-plot showing the correlation between the average VOT in production for /pɑ/ and the VOT of the highest-rated "*p*" in the perception task.

All production measures were independently correlated with the prototype VOT. However, only the VOT values from the /pɑ/ item contributed significantly to the regression formula, and the /b/ items added marginally significant additional information; the other items did not add additional information to the equation. The variation in produced /pɑ/ VOT was responsible for 27% of the variance in listeners' perceptual ratings [$F(1,18)=6.73,\ p<0.02$], whereas the variation in /b/ VOT was responsible for an additional 15% [$F(1,14)=3.94,\ p=0.07$]. Figure 2 shows the regression line with /pɑ/ VOT as the predictor. A complete listing of the regression coefficients, r^2, change in r^2, and statistics are given in Table I. It is apparent from Fig. 2 that the /pɑ/ correlation was augmented by the unusually high perceptual scores for two individuals; however, the correlation remains significant even if data from these individuals are removed, suggesting they are not the primary cause of the correlation ($r=0.49,\ p<0.05$).

These results suggest that the present methodology was successful at finding a link between perception and production. There was a significant relationship between participants' productions and their perceptual prototypes. Individuals whose perceptual prototype for "p" occurred at longer VOTs also tended to produce longer VOTs themselves.

It is interesting to note that while the listeners' productions of the voiceless stops did not provide additional information above and beyond their production of the target item itself, their production of the first voiced stop in the hierarchy approached doing so. This suggests that production of the voiceless items may be highly correlated within each individual, but that production of voiced items may not be as correlated with the voiceless tokens. Indeed, the p, t, and k measures were highly correlated, with correlations ranging from 0.88 to 0.96; the b, d, and g items, while correlated with one another, did not correlate so highly with the voiceless items (correlations among the voiced items ranged from 0.54 to 0.75; correlations between voiced and voiceless items ranged from 0.37 to −0.06). The additional voiceless stops may not have added additional information because they were highly correlated with the production of the target item; the marginal effect of adding the /b/ items into the equation suggest that the voiced stops contained additional information beyond that provided by the production of the target item. Perhaps these provide information about the degree of category separability the individual prefers (participants whose /b/ productions had relatively short VOTs tended to have prototype /p/s with longer VOTs, suggesting a preference for more easily discriminable categories).

Something more akin to listeners' category boundaries were also examined: the minimum VOTs they considered acceptable for a /p/. In production, this was estimated as the minimum /p/ VOT that speakers produced—across listeners this averaged 46.7 ms. In the perception task, the VOT of the earliest member of the continuum that listeners rated as being a member of the /p/ (rather than /b/) category was measured—this averaged 44.7 ms across listeners. These two measures showed a significant correlation across participants ($r=0.48,\ p<0.05$), suggesting that the lower ends of listeners' categories were comparable in perception and production, much as were their category prototypes.

There was also evidence for a hyperspace effect (Frieda *et al.*, 2000; Johnson *et al.*, 1993a), as has previously been found for vowels. Perceptual prototypes had an average VOT of 90 ms, whereas VOTs of the participants' productions averaged only 73 ms. This difference was significant by a paired *t*-test, $t=3.61$, $p<0.002$. Fifteen of the 20 participants showed this pattern of longer VOTs for their perceptual prototype than in their speech production.

The results from this experiment suggest that individual differences in production are related to differences in perception. Both the best exemplar of the category, and the lowest acceptable member of the category, were similar in production and perception within each individual. It appears that production-perception correlations can be found with an ap-

TABLE I. Results from multiple regression from experiment 1.

Step	Individual *r*	Multiple *r*	Multiple r^2	Change in r^2	Change in F	Significance
pɑ	0.522	0.522	0.272	0.272	6.73	0.018
p	0.361	0.550	0.303	0.031	0.75	0.398
t	0.311	0.551	0.303	0.000	0.00	0.960
k	0.362	0.561	0.315	0.012	0.27	0.614
b	−0.442	0.682	0.466	0.150	3.94	0.067
d	−0.297	0.688	0.474	0.008	0.20	0.662
g	−0.323	0.756	0.572	0.098	2.75	0.123

propriate perceptual task and an appropriate acoustic correlate.

One possibility is that this task can be used to evaluate different acoustic cues. Often, there are multiple proposals for how a given phonemic distinction should be described. It might be possible to evaluate different metrics by determining the degree to which perception and production measures using these proposed cues are correlated. Such an approach requires that correlations occur only for those cues that are actually used by listeners. Given that different proposed cues are generally highly correlated with one another, perception-production correlations might be present for all proposed cues. If so, the correlations would not provide additional information to distinguish among them. This approach can only be useful if correlations exist for some acoustic cues, but not for all. Experiment 2 examines this in more detail.

III. EXPERIMENT 2

Unlike the /p/-/b/ distinction examined in experiment 1, there are some phonemic distinctions where multiple metrics or measures appear to be equally plausible. One such phoneme is the fricative /ʃ/ (''sh''). Fricatives are produced by creating a partial obstruction in the mouth. Forcing air through this narrow constriction causes turbulence in the airstream, resulting in a ''noisy'' sound, with energy at a broad range of frequencies (Pickett, 1980). The location of the obstruction differs between an /s/ and an /ʃ/, and a number of studies have examined the possible acoustic correlates of this difference. Research has focused on four attributes as being particularly important for fricatives in general: spectral properties of the fricative noise, noise duration, noise amplitude, and spectral properties of the transition between the fricative and the following vowel (Jongman *et al.*, 2000).

Of these four types of cues, spectral properties of the noise appear to be most important for the /s/-/ʃ/ distinction. Noise duration and overall amplitude appear particularly important for distinguishing the sibilant fricatives (/s/ and /ʃ/) from the nonsibilants (/f/ and /θ/), but do not appear to distinguish between the two sibilants (Behrens and Blumstein, 1988b). Relative amplitude differences do appear to be important, but this may be a result (at least in part) of concomitant changes in spectral properties (Hedrick, 1997; Hedrick and Ohde, 1993). Although some research has examined transition information (Sussman, 1994; Sussman and Shore, 1996; but see Fowler, 1994) as cues to place of articulation within fricatives, most researchers have focused on spectral properties of the noise as being the most important cues for distinguishing /s/ and /ʃ/. This information has been shown to be sufficient for a high degree classification in several studies (Tabain, 1998; Tomiak, 1991).

There have been many proposals as to the best ways to characterize this spectral information. Harris (1958; see also Heinz and Stevens, 1961; Hughes and Halle, 1956; May, 1976) found that the noise center frequency information (roughly the frequency mean) is the primary cue for distinguishing these particular phonemes. Strevens (1960) reported that the frication range for /s/ was shifted higher than that for /ʃ/, which would likewise imply that the mean frequency for /s/ would be higher (although he measured only the range of frequencies at which energy occurred, and did not actually calculate average values). Forrest and colleagues (1988) examined three spectral moments [centroid (or frequency mean), skewness, and kurtosis], and found that skewness of frication was the primary feature distinguishing these phonemes, although centroids might also aid in their discriminability [but see Shadle and Mair (1996) for contrasting data on the role of skewness]. In contrast, other work has examined spectral peaks, which are more akin to a statistical mode than a mean (Behrens and Blumstein, 1988a; Jassem, 1965; Seitz *et al.*, 1987). Thus, while there is wide agreement that the frication noise is the primary acoustic cue for distinguishing /s/ and /ʃ/, there is less agreement on the appropriate way of measuring this cue.

One reason for this disagreement is that these measures are highly correlated in these phonemes. They do not refer to independent information in the spectrum, but instead are different ways of describing the same information. It is therefore very difficult to distinguish between these measures experimentally; any modification of one cue results in changes in the other cues as well.

In recent work, Jongman *et al.* (2000) examined a variety of cues to fricatives. They found that while many of these cues were successful at classifying fricatives, discriminant analysis suggested that spectral peak location was a more important cue than were spectral moments.

The present experiment proposes a different way of evaluating these measures. If the degree of perception-production correlation for a given cue is based on the extent to which that cue is related to the dimensions utilized by the listener, then the degree of correlation can be used as means of evaluating this relation. Correlations should be stronger for the cue most similar to that which listeners are actually using.

Given that the different cues are themselves highly correlated, one concern is that all cues may result in strong perception-production correlations. If so, these types of correlations would not be useful as a means of evaluating metrics. Thus the primary goal in the present experiment is to determine whether perception-production correlations can distinguish among different acoustic measures, even when those measures are themselves related.

A. Method

1. Subjects

Twenty-four volunteers participated in exchange for a cash payment. Ratings from five of these participants did not fall off towards the extremes of the continuum; their data were not analyzed.[3] (For this experiment, in which both endpoints of the series were clearly not /ʃ/ tokens, a minimum of a one category rating drop-off to either side was required for data inclusion.) This left a total of 19 listeners, one of whom was also in experiment 1. The average rating drop-offs were 4.1 units towards the velar side, and 6.3 units toward the /s/ side.

2. Stimuli

For the model stimuli for the production task, a female native talker of English (RSN) recorded four tokens of each

CV syllable consisting of either /s/ or /ʃ/ followed by one of the seven vowels /i, e, æ, u, o, ɑ, ʌ/. The recording manner for these 56 items was identical to that in experiment 1.

For the perception task, the stimuli consisted of series ranging from /sæ/ to /ʃæ/ and from /ʃæ/ to beyond-/ʃæ/ (or */ʃæ/). The vowel /æ/ was chosen because it does not entail lip-rounding or protrusion, which can alter the spectral information in the fricative (Soli, 1981). These stimuli were produced synthetically, as the type of editing used in experiment 1 can only be used to create duration-based series, not frequency-based series. The synthetic stimuli were modeled on a male voice chosen because it is well-mimicked by our speech synthesis program.

Use of a male voice in the perception task and a female voice in the production task should prevent listeners from hearing the items as coming from the same individual, and from judging the voice in the perceptual task on the basis of the speech from the talker in the production task. This avoids one potential criticism of experiment 1, that the talker used in the perceptual study and the model for the production component were the same individual. If listeners in the production task were trying to mimic that talker's speaking style, correlations could have occurred for that reason alone. If perception-production correlations are found in this experiment where the voices clearly differed, it would suggest that these relationships are not an artifact of having used the same talker for both tasks.

The model speaker produced tokens of /sæ/ and /ʃæ/ in the context of the carrier phrase, ''Please say _____ to me.'' The transition and vowel portions of the /s/ and /ʃ/ syllables were temporarily removed, leaving only the 215-ms frication portion of the syllables. These two frications were synthesized using the parallel mode of a cascade/parallel synthesizer (Klatt, 1980). Formant frequencies, amplitudes, and bandwidths were carefully adjusted to make the synthetic tokens both sound as similar to the original items as possible and look as similar as possible in spectral cross-section. The vowel portion from one of the two syllables was likewise synthesized and its formant values, bandwidths, and amplitudes adjusted. This vowel portion was then appended to both the /s/ and /ʃ/ tokens, resulting in two endpoints which had identical synthesis parameters after the first 215 ms (or 43 frames). Values for the initial frication portion were then interpolated between the two endpoints to make a 21-item series. This interpolation was performed on all parameters that differed across the endpoints: the amplitudes and bandwidths of all formants, and the location of formants 2, 4, and 5 (the location of formants 1 and 3 were the same in both endpoints, as was the bandwidth of formant 3). The transition and vowel portions (which occurred after the initial 215 ms) were held constant across items.

Rather than make the series continue beyond /ʃ/ acoustically (by continuing to adjust formant and amplitude values in the same manner as in the first half of the continuum), the series continued beyond /ʃ/ in an articulatory sense, towards a more extreme place of articulation. Continuing to adjust formant and amplitude values in the same manner could have resulted in an endpoint that was not possible from a human vocal tract. A linguist was asked to produce fricatives from a variety of places of articulation: alveolar (as in /s/), palatal (/ʃ/), and velar and uvular fricatives (which do not occur in English but do occur in other languages; uvular fricatives occur in one of the languages in which she was fluent). The first five formant movements between her tokens were analyzed, and the formants, amplitudes, and bandwidths in our synthetic continua were adjusted to move in the same manner. Thus our formant movements beyond /ʃ/ were such that they moved towards a more velar/uvular place of articulation. A 20-item series was created in this manner, resulting in a total of 41 stimuli (the /s/ endpoint, 19 interpolated items between /s/ and /ʃ/, the /ʃ/ endpoint, 19 interpolated items beyond /ʃ/, and the most velar endpoint).

3. Procedure and measures

This study was combined with that of another perceptual rating study using other syllables, not reported here. Participants recorded their speech at the onset of the first session, and then took part in the two perceptual tasks; the order of these tasks was counterbalanced across participants.

Procedures for both the production and perception tasks were identical to those in experiment 1, with the exception that listeners were asked to rate the phonemes as examples of the sound ''sh,'' rather than as examples of the sound ''p.'' There were 16 blocks of trials in the perceptual task and 2 blocks of trials in the production task.

Three types of acoustic measurements were taken on the participants' productions: frication centroid, skewness, and peaks. For centroids and skewness, analysis was modeled after that of Forrest *et al.* (1988); a 20-ms analysis window was used to compute a sequence of Fourier spectrum; the initial analysis window centered on the frication onset and each subsequent spectra was computed over a window centered 10 ms further into the signal, resulting in a series of measurements containing 50% overlap. The speech signal was preemphasized by first differencing (with preemphasis of 0.94), and a 400-point Hamming window was used for analysis. (For the synthetic speech in the perceptual task, the stimuli had a 10-kHz sampling rate, so a 20-ms window resulted in a 200-point Hamming window). The spectra were treated as random probability distributions, and the centroid (or mean) and skewness of the distribution were calculated. The number of analysis windows was set at 10; analysis thus occurred over a total of 110 ms and the means and skewness values were averaged across the 10 frames. [This duration was suggested by Tomiak (1991) to provide a valid estimate of the fricative, based on results from a masking study.] The measured portions were at the onset of the fricatives.

Peak frequency measures were performed using the CSRE software package from AVAAZ. Each production was analyzed using a fast Fourier transform over a 128-point Hamming window with 50% overlap, averaged across the initial 100 ms. This analysis was then treated as a random probability distribution, but instead of finding the moments of the distribution, the mode was found instead (or the frequency at which the greatest amount of energy was present). Some speakers' productions did appear to have more than one peak frequency; however, the single frequency value

TABLE II. Acoustic measures from experiment 2.

Participant	/ʃæ/ centroid	/ʃæ/ skewness	/ʃæ/ peak
ggg	5197	−0.001	2822
ddy	5302	−0.059	3628
clk	5315	−0.051	3461
ic	5174	−0.028	3528
cer	5107	+0.031	2871
hem	5122	+0.001	4111
jg	5202	−0.060	4692
nv	5316	−0.064	3477
cab	5303	−0.088	4199
acy	5450	−0.138	5966
jem	5124	+0.023	3374
bam	5268	−0.053	4590
iaf	5036	+0.089	4316
kjp	5355	−0.117	5429
kfb	4999	+0.041	3828
vjl	5302	−0.092	5410
mlt	5101	+0.040	4423
ksk	5367	−0.082	3916
tlg	5272	−0.064	4033

with the greatest energy was selected. These acoustic measurements are shown in Table II.

For the perceptual task, the single item in the continuum with the highest rating was considered the listener's prototype, as in experiment 1. This prototype was measured for its frication centroid, skewness, and peak, in the same manner as the participants' productions described above. For one listener, three adjacent items received equally high ratings; the values for these three items were averaged on each measure to find the prototype for that listener.

B. Results and discussion

As expected, the three production measures were highly correlated, especially the two spectral moments measures. For centroids and skewness, the correlations for the /ʃ/ measures across participants was -0.94, $p<0.0001$. For centroids and peaks, the correlation was marginal, $r=0.43$, $p<0.07$, and for skewness and peaks it was -0.54, $p<0.02$. Thus, all three measures do seem to be based on related aspects of the same information.

The perception-production correlations were examined using all three measures. In experiment 1, it was found that measurements from the single syllable identical to the perceptual item were the most relevant; the relationship between the /ʃæ/ prototypes and the average /ʃæ/ productions were therefore examined here (that is, the average value across the eight different productions of the /ʃæ/ syllable, rather than the average across all 56 /ʃ/ tokens).

For centroids, the correlation was not significant ($r=-0.30$, $p>0.10$). Moreover, it was actually negative. While negative correlations across different measures would be unsurprising, correlations between production and perception using the same measure should be positive. It would be rather odd for those individuals who produced the most extreme /ʃ/ tokens to prefer the least extreme versions perceptually. Thus finding a negative correlation here suggests that individuals' perceptual prototypes are not determined by the tokens' centroid measures.

For skewness, correlations were again both negative and nonsignificant ($r=-0.19$, $p>0.40$). Skewness likewise does not appear to be a primary factor in individuals' prototypes.

The correlation for peak frequency, however, was not only positive but also was significant ($r=0.50$, $p<0.03$). The correlation was almost identical in size to that found in experiment 1 ($r=0.52$). This suggests that frequency peaks may be a better indication than centroids of what makes a particular fricative token sound better to a listener. Those listeners who showed more extreme values in their frequency peaks for /ʃ/ also preferred listening to more extreme tokens. This was not the case for either fricative centroid or skewness measures.

More importantly for the present purposes, the findings also suggest that correlations between speech perception and speech production can differentiate between different acoustic cues. Only acoustic measurements based on the peaks in the spectrum appeared to be related to listeners' goodness ratings. This is in accord with recent findings by Jongman *et al.* (2000), also suggesting that spectral peaks are better cues to fricative discrimination than are spectral moments. This suggests that even when different acoustic cues are highly correlated, perception-production correlations can be used to discriminate among different measures.

The present results also extend the general finding from experiment 1 that correlations exist between speech perception and speech production. Finding these correlations in two different experiments, for two different phonemic contrasts, suggests that these results are fairly common. Furthermore, one potential problem in experiment 1 was that the talker whose voice served as the model for the production study was the same talker as that judged in the perceptual study. If the correlations in that experiment were actually the result of participants trying to mimic that talker, such results would not be expected in the present experiment, where the talker that served as the base for the perceptual study and the talker that served as the model for the production component were not only different individuals, but also were of different genders.

The next logical step would be to directly test the idea that peaks are more important than centroids by orthogonally varying these dimensions. Unfortunately, it is not possible to vary these two properties in this manner (at least not while maintaining good endpoint stimuli), which is why the finding of an alternative method of differentiating cues is so important. As an example of the difficulty, the synthetic /s/ endpoint in this study had a mean frequency of approximately 4500 Hz, and a peak frequency of approximately 4700 Hz. In order to manipulate peak and centroid independently, it would be necessary to create a series that maintained this centroid at 4500 Hz while the peak frequency moved from 4700 Hz down to a value appropriate for an /ʃ/ (approximately 3400 Hz based on our /ʃ/ endpoint). Lowering the peak is relatively easy in synthetic speech; however, in order to keep the centroid from changing along with the peak, this would require adding diffuse high-frequency energy to compensate for the loss of energy at 4700 Hz (and the increase in energy at 3400 Hz). With an unlimited frequency range, this

would be easily doable. However, the implementation of the Klatt synthesizer used here was limited to 5000 Hz, making it impossible to add sufficient high-frequency energy without creating a high-frequency peak. This limits the comparison of peaks and centroids to indirect measures, such as the correlations discussed here.

IV. GENERAL DISCUSSION

These two experiments demonstrate that individual differences in production are related to differences in perception. Listeners whose productions are more extreme along an acoustic continuum appear to prefer hearing more extreme productions from other speakers as well. This is in addition to a hyperspace effect, in which individuals prefer listening to more extreme tokens than they themselves produce.[4]

However, these production-perception correlations are not ubiquitous. Although they will occur for acoustic cues known to be used by listeners (such as VOT), they do not occur for all possible acoustic measures. Despite the fact that the three acoustic measures used in experiment 2 were highly correlated with one another, significant perception-production correlations were found only for one of them—in particular, for the one most supported by a recent comparative analysis by Jongman *et al.* (2000). Nor did the lack of an effect in the other two measures appear to be caused by a lack of power: the results were not only nonsignificant, but were in the opposite direction as that expected.

These findings suggest that this task can be used to evaluate different acoustic measures. For many phonemic distinctions, there is no apparent ''best'' measure. Many different metrics may be proposed, and it is often difficult to discriminate among such metrics experimentally. Looking for links between perception and production may provide another means for making such comparisons.

Clearly, this conclusion must be taken as tentative at this point. More research is necessary to ensure that these correlations only exist (and consistently exist) when the appropriate measure is used. Furthermore, since the degree of variability among individuals can influence the likelihood of finding a significant correlation, this task is likely to be best used in a converging methods approach, in combination with more traditional ways of contrasting metrics (such as that of Richardson, 1992 and Tomiak, 1991). Still, the results are at least suggestive that this task can provide a better indication of the types of acoustic cues most likely to be used by listeners, and may be particularly useful in situations where other contrastive methods are not possible.

One limitation of the present approach is that it requires the use of a single cue, such as frequency centroid or VOT. For some phonemic distinctions, sets of cues have been proposed that work as a whole. For example, peak differences (Syrdal and Gopal, 1986) have been proposed as cues to stop consonant place of articulation. These were proposed as a set of values, and there is no reason to believe that the individual peak differences would of necessity correlate with one another. If each component is a dimension in multi-dimensional space, the overall location of a value in space would depend on the values from the set of measures, but need not correlate highly with any single measure. Since there is no statistical test that provides an overall measure of the strength of a relationship between two sets of variables (see Cohen and Cohen, 1983), the present methodology of examining correlations between perceptual prototypes and average productions is likely to be limited to cases in which there is a single acoustic property that can be measured.

Although significant correlations between speech production and speech perception were found in both of the present experiments, these correlations were quite modest, accounting for approximately 27% of the variance in listeners' perceptual prototypes. There are several possible reasons why this might be the case. One possibility is that the representations for perception and production are at least partially distinct. If this were the case, it could be taken as an argument against models such as motor theory (Liberman and Mattingly, 1985), which rely on identical representations for both input and output. However, there are other potential explanations for the small size of these effects that limit the strength of this conclusion. The effect size may be due, in part, to the existence of a hyperspace effect (Johnson *et al.*, 1993a); listeners highest-rated item may be more strongly related to a hyperarticulated production than to a typical one. Since participants were not asked to exaggerate their productions, the correlation may be less strong than would otherwise be the case. It is also possible that a stronger correlation might have arisen with more participants, or more acoustic measures per participant.

Anecdotal evidence suggests that people are often surprised by the sound of their own voice when they hear a recording of it. This is due to the fact that one's own voice is heard both via air conduction (as others hear us) and via bone conduction within the head, which emphasizes low frequencies. This makes our own voices sound more resonant to ourselves than to others. Perception-production correlations might therefore be expected to be strongest for temporally-based contrasts and for sounds without vocal fold vibration, as these would be unaffected by a low-frequency emphasis.

As expected from previous research, correlations between speech perception and speech production do appear to exist, although they are not as strong as might be expected on the basis of some theoretical models. These correlations appear to distinguish between different acoustic cues, suggesting that they may be usable as a way of evaluating different proposed metrics. Future research will be needed to examine this proposal in more depth.

ACKNOWLEDGMENTS

Some of this research was from a doctoral dissertation, directed by James R. Sawusch. Special thanks to him, and also to Peter Jusczyk and Paul Luce for helpful discussions, and to Sheryl Clouse for performing the peak measurements described in experiment 2. Also, thanks to Jim Sawusch for the use of his acoustic measurement program, NMEASURE, which was used for the centroid and skewness measures in experiment 2, Nina Azhdam for the reliability measures in experiment 1, and to Rob Fox, Chris Turner, and an anonymous reviewer for comments on a previous draft.

[1]As a comparison, Frieda *et al.* (2000) report that 11 of their 35 participants did not meet their criteria for prototype designation.

[2]Because of a computer recording error, one subject's production task had to be recorded at the start of the second session.

[3]Four of these five participants failed to show a drop-off towards the /s/ end of the series, rather than towards the velar end; that is, they gave high rankings to the most /s/-like member of the series. In contrast, the 19 participants whose data were kept gave this item an average rating of 0.6, indicating that they did not hear this item as a member of the /ʃ/ category at all. One possibility is that the individuals whose data were dropped were confused by English orthography, in which /s/ and /ʃ/ both contain ''s.''

[4]Hyperspace effects could only be explored in experiment 1: in experiment 2, the perceptual items were based on synthetic speech with a limited frequency range. This makes the direct comparison to natural productions difficult.

Ainsworth, W. A., and Paliwal, K. K. (**1984**). ''Correlation between the production and perception of the English glides /w, r, l, j/,'' J. Phonetics **12**, 237–243.

Bailey, P. J., and Haggard, M. P. (**1973**). ''Perception and production: Some correlations on voicing of an initial stop,'' Lang. Speech **16**, 189–195.

Bailey, P. J., and Haggard, M. P. (**1980**). ''Perception-production relations in the voicing contrast for initial stops in 3-year-olds,'' Phonetica **37**, 377–396.

Behrens, S. J., and Blumstein, S. E. (**1988a**). ''Acoustic characteristics of English voiceless fricatives: a descriptive analysis,'' J. Phonetics **16**, 295–298.

Behrens, S. J., and Blumstein, S. E. (**1988b**). ''On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants,'' J. Acoust. Soc. Am. **84**, 861–867.

Bell-Berti, F., Raphael, L. J., Pisoni, D. B., and Sawusch, J. R. (**1979**). ''Some relationships between speech production and perception,'' Phonetica **36**, 373–383.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). ''Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production,'' J. Acoust. Soc. Am. **101**, 2299–2310.

Cohen, J., and Cohen, P. (**1983**). *Applied Multivariate Regression/ Correlation Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ).

Cooper, W. E. (**1974**). ''Perceptual-motor adaptation to a speech feature,'' Percept. Psychophys. **16**(2), 229–234.

Cooper, W. E., and Lauritsen, M. R. (**1974**). ''Feature processing in the perception and production of speech,'' Nature (London) **252**, 121–123.

Cooper, W. E., and Nager, R. M. (**1975**). ''Perceptuo-motor adaptation to speech: an analysis of bisyllabic utterances and a neural model,'' J. Acoust. Soc. Am. **58**, 256–265.

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (**1977**). ''Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context dependent cues,'' Percept. Psychophys. **22**(2), 109–122.

Flege, J. E. (**1999**). ''Age of learning and second-language speech,'' in *Second Language Acquisition and the Critical Period Hypothesis*, edited by D. P. Birdsong (Erlbaum, Mahwah, NJ).

Flege, J. E., and Eefting, W. (**1986**). ''Linguistic and developmental effects on the production and perception of stop consonants,'' Phonetica **43**, 155–171.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). ''Statistical analysis of word-initial voiceless obstruents: Preliminary data,'' J. Acoust. Soc. Am. **84**, 115–123.

Fowler, C. A. (**1986**). ''An event approach to the study of speech perception from a direct-realist approach,'' J. Phonetics **14**, 3–28.

Fowler, C. A. (**1994**). ''Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation,'' Percept. Psychophys. **55**(6), 597–610.

Fox, R. A. (**1982**). ''Individual variation in the perception of vowels: Implications for a perception-production link,'' Phonetica **39**, 1–22.

Frieda, E. M., Walley, A. C., Flege, J. E., and Sloane, M. E. (**2000**). ''Adults' perception and production of the English vowel /i/,'' J. Speech Lang. Hear. Res. **43**, 129–143.

Harris, K. S. (**1958**). ''Cues for the discrimination of American English fricatives in spoken syllables,'' Lang. Speech **1**(1), 1–7.

Hedrick, M. (**1997**). ''Effect of acoustic cues on labeling fricatives and affricates,'' J. Speech Lang. Hear. Res. **40**, 925–938.

Hedrick, M. S., and Ohde, R. N. (**1993**). ''Effect of relative amplitude of frication on perception of place of articulation,'' J. Acoust. Soc. Am. **94**, 2005–2026.

Heinz, J. M., and Stevens, K. N. (**1961**). ''On the properties of voiceless fricative consonants,'' J. Acoust. Soc. Am. **33**, 589–596.

Hoffman, P. R., Daniloff, R. G., Alfonso, P. J., and Schuckers, G. H. (**1984**). ''Multiple-phoneme-misarticulating children's perception and production of voice onset time,'' Percept. Mot. Skills **58**, 603–610.

Hughes, G. W., and Halle, M. (**1956**). ''Spectral properties of fricative consonants,'' J. Acoust. Soc. Am. **28**, 303–310.

Jamieson, D. G., and Rvachew, S. (**1992**). ''Remediating speech production errors with sound identification training,'' J. Speech-Language Patholo. Audiol. **16**, 201–210.

Jassem, W. (**1965**). ''The formants of fricative consonants,'' Lang. Speech **8**, 1–16.

Johnson, K., Flemming, E., and Wright, R. (**1993a**). ''The hyperspace effect: phonetic targets are hyperarticulated,'' Language **69**(3), 505–528.

Johnson, K., Ladefoged, P., and Lindau, M. (**1993b**). ''Individual differences in vowel production,'' J. Acoust. Soc. Am. **94**, 701–714.

Jongman, A., Wayland, R., and Wong, S. (**2000**). ''Acoustic characteristics of English fricatives,'' J. Acoust. Soc. Am. **108**, 1252–1263.

Klatt, D. H. (**1975**). ''Voice onset time, friction, and aspiration in word-initial consonant clusters,'' J. Speech Hear. Res. **18**, 686–706.

Klatt, D. H. (**1980**). ''Software for a cascade/parallel formant synthesizer,'' J. Acoust. Soc. Am. **67**, 971–995.

Kuhl, P. K., and Meltzoff, A. N. (**1982**). ''The bimodal perception of speech in infancy,'' Science **218**, 1138–1141.

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

Liberman, A. M., Cooper, F. S., Harris, K. S., and MacNeilage, P. F. (**1962**). ''A motor theory of speech perception,'' in *Speech Communication Seminar*, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). ''Perception of the speech code,'' Psychol. Rev. **74**(6), 431–461.

Lisker, L., and Abramson, A. S. (**1964**). ''A cross-language study of voicing in initial stops: Acoustical measurements,'' Word **20**(3), 384–422.

Lisker, L., and Abramson, A. S. (**1970**). ''The voicing dimension: Some experiments in comparative phonetics,'' in *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague, 1967 (Academia, Prague), pp. 563–567.

May, J. (**1976**). ''Vocal tract normalization for /s/ and /ʃ/,'' Haskins Laboratories: Status Report on Speech Research, SR-48, pp. 67–73.

Miller, J. L., and Volaitis, L. E. (**1989**). ''Effect of speaking rate on the perceptual structure of a phonetic category,'' Percept. Psychophys. **46**, 505–512.

Nearey, T. M. (**1992**). ''Context effects in a double-weak theory of speech perception,'' Lang. Speech **35**(1,2), 53–171.

Paliwal, K. K., Lindsay, D., and Ainsworth, W. A. (**1983**). ''Correlation between production and perception of English vowels,'' J. Phonetics **11**, 77–83.

Perkell, J. S., and Matthies, M. L. (**1992**). ''Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability,'' J. Acoust. Soc. Am. **91**, 2911–2925.

Pickett, J. M. (**1980**). *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception* (University Park, Baltimore).

Richardson, K. H. (**1992**). ''An analysis of invariance in English stop consonants,'' Diss. Abstr. Int., B **53**(3-B), 1633.

Seitz, P. F. D., Bladon, R. A. W., and Watson, I. M. C. (**1987**). ''Across-speaker and within-speaker variability of British English sibilant spectral characteristics,'' J. Acoust. Soc. Am. Suppl. 1 **82**, S37.

Shadle, C. H., and Mair, S. J. (**1996**). ''Quantifying spectral characteristics of fricatives,'' 4th International Conference on Spoken Language Processing (ICSLP).

Soli, S. D. (**1981**). ''Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation,'' J. Acoust. Soc. Am. **70**, 976–984.

Strevens, P. (**1960**). ''Spectra of fricative noise in human speech,'' Lang. Speech **3**, 32–49.

Sussman, H. M. (**1994**). ''The phonological reality of locus equations across manner class distinctions: Preliminary observations,'' Phonetica **51**, 119–131.

Sussman, H. M., and Shore, J. (**1996**). ''Locus equations as phonetic de-

scriptors of consonantal place of articulation,'' Percept. Psychophys. **58**, 936–946.
Sussman, H. M., Hoemeke, K. A., and Ahmed, F. S. (**1993**). ''A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation,'' J. Acoust. Soc. Am. **94**, 1256–1268.
Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (**1991**). ''An investigation of locus equations as a source of relational invariance for stop place categorization,'' J. Acoust. Soc. Am. **90**, 1309–1325.
Syrdal, A. K., and Gopal, H. S. (**1986**). ''A perceptual model of vowel recognition based on the auditory representation of American English vowels,'' J. Acoust. Soc. Am. **79**, 1086–1100.
Tabain, M. (**1998**). ''Non-sibilant fricatives in English: Spectral information above 10 kHz,'' Phonetica **55**, 107–130.
Tomiak, G. R. (**1991**). ''An acoustic and perceptual analysis of the spectral moments invariant with voiceless fricative obstruents,'' Diss. Abstr. Int., B **51**(8-B), 4082–4083.