The hyperspace effect: Phonetic targets are hyperarticulated

Keith Johnson
Edward Flemming
Richard Wright

Abstract

A common but rarely defended assumption is that phonetic reduction processes apply to hyperarticulated phonetic targets. If this assumption is correct, a two stage model of phonetic implementation is indicated; at the first stage distinctive features are mapped to hyperarticulated phonetic targets, and at the second stage these phonetic targets are reduced. The experiments reported in this paper supported this model by showing that phonetic targets are hyperarticulated. Listeners adjusting the first and second formants of synthetic vowels chose values which were only found in hyperarticulated speech.

Introduction

Most approaches to phonetic realization that directly address the issue of casual speech assume canonical phonetic representations, or targets, which are hyperarticulated, with variation being introduced by reduction processes. For example, Lindblom (1990) conceptualizes speech production as a feedback system in which the input is the goal, and the extent to which it is achieved depends on the "gain" of the feedback loop, so the gain is analogous to something like effort. Browman and Goldstein (1990) propose that casual speech variants of canonical lexical gestural representations are produced by increasing gestural overlap and reducing gestural magnitudes.

In fact it appears that this general approach is widely assumed since almost all discussions of the casual speech - clear speech continuum are cast in terms of "undershoot", "reduction" and related concepts. However, this is not the only possible account of these phenomena. We could postulate the existence of both reduction and hyperarticulation processes, in which case the canonical phonetic representation would be of an intermediate level, perhaps akin to citation forms. We can probably reject the third logical possibility, which is the hypothesis that there are only hyperarticulation processes, since the most reduced forms of words can be so indistinct that if they were the starting point it would be difficult to derive the clear distinctions between them that exist in hyperarticulated speech. This is essentially the same type of argument that leads Jakobson & Halle (1956, p. 6), among others, to state that the most clearly articulated speech is most relevant to phonological analysis since it contains the most information.

In this paper, we report the results of several method of adjustment studies in which listeners controlled the first two formants in synthetic vowels. We will show that listeners' responses in the method of adjustment task differ systematically from their own productions and that this systematic deviation from normal speech is consistent with the view that phonetic targets are hyperarticulated.

Phonetic comparison of vowel systems and the method of adjustment

In the method of adjustment (Nooteboom, 1973; Ganong & Zatorre, 1980; Samuel, 1982; Johnson, 1989) a listener is given control over one or more parameters of a speech synthesizer and is asked to adjust them until the machine pronounces a particular speech sound. Where other methods used in speech perception research focus on the boundaries between phonetic categories, the method of adjustment provides information about linguistic/phonetic targets for speech sounds. Repp & Liberman (1987) discuss the need for data on the internal structure of phonetic categories, as opposed to category boundaries, and assume that the method of adjustment provides information about the "prototypes" of phonetic categories (see also Samuel, 1982 concerning this assumption). However, they note that "until recently, no one had used methods designed to identify prototypes" (p. 90) and that "the application of such methods has so far failed to yield entirely satisfactory results".

The method of adjustment is useful for cross-linguistic comparisons of vowel systems

because personal differences in vocal tract anatomy (vocal tract size, oral cavity to pharynx cavity ratio, palate doming, lip shape, etc.) give rise to acoustic differences between speakers (Ladefoged & Broadbent, 1957). Although research since the late 1940's has shown that the speech signal varies quite considerably from speaker to speaker even within the same language and dialect (Joos, 1948; Peterson & Barney, 1952), cross-linguistic acoustic/phonetic comparisons of vowels confound personal and linguistic differences. Thus, any cross-linguistic acoustic comparison of vowel systems includes some unknown amount of personal variation.

One way to compensate for personal variation in making cross-linguistic comparisons is to scale the measured formant values by a factor which is related to vocal tract size. The scale factor may be derived from the range of observed formant values for a particular speaker (Gerstman, 1968), the mean of the observed formant values (Lobanov, 1971), the mean of the log-transforms of the observed formant values (Nearey, 1977), or a function of the speaker's fundamental frequency (F0) (Miller, 1989; Syrdal & Gopal, 1986). Disner (1980) suggested that it is valid to use the formant mean or range "so long as the data are drawn from a single language or dialect, such that the same set of vowel phonemes is shared by all speakers" (p. 257). But, in making cross-linguistic or cross-dialectal comparisons, reliance on mean formant values or formant range is not usually valid. Methods of formant normalization which use the speaker's F0 are also flawed because they rely on the observed *rough* correlation of F0 with vocal tract length, and are also only valid for comparisons of vowels produced in similar prosodic contexts.

Another class of normalization techniques uses cross-linguistic formant averages as normalizing factors. For instance, Disner (1980) used PARAFAC (a three-mode factor analysis technique, Harshman, 1970) to compare the vowels of English, German, Danish, Swedish, and Dutch. The PARAFAC procedure relates measurements from individuals to the overall mean of the data set and finds speaker constants which scale the individual's vowel space to the overall vowel space. In later work, Disner (1986) compared the vowels of various languages using analysis of variance models which included factors for vowel, speaker and language. In both of these approaches the differences between languages are tested as deviations from cross-linguistic means. Two limitations are inherent in this class of normalization technique. First, only comparable vowel qualities can be tested cross-linguistically, which means that it is not possible to compare whole vowel systems when they contain unequal numbers of vowels. Second, the method assumes that the average individual deviation from the overall mean is not correlated with language. To see the problem with this assumption consider an extreme example. If all of the data for language x is taken from recordings of female speakers while all of the data from language y is taken from recordings of male speakers, an analysis of variance would show quite large differences in formant values as a function of language even though speaker is included as a factor in the model. Thus, although the statistical techniques proposed by Disner may provide a better solution to the normalization problem (for cross-linguistic comparisons) than those provided by other approaches, the problem is not solved by using cross-linguistic formant averages as normalizing factors because the results still depend on the comparability of the groups of speakers who represent each language (see Behne, 1989).

The method of adjustment offers a better way of making cross-linguistic phonetic comparisons of vowel systems. By using a single synthetic voice, the method of adjustment makes it possible to ascertain the listener's expectations for vowel sounds in a particular language *for that voice*. Thus, the linguistic and personal aspects of the speech signal are disentangled.

In the studies of the Southern California English vowel space reported here, we found that listeners chose vowel formants which did not match those produced by *any* speaker in normal speech. The discrepancy between listeners' choices in the method of adjustment and speakers' productions is interesting because it is systematic. The perceptual vowel space was expanded relative to the production space; high vowels were higher, low vowels lower, front vowels more front, and back vowels more back. We will present data which indicates that the perceptual vowel space reflects vowels produced in clear or hyperarticulated speech.

Preliminary study

A preliminary study was designed to give a first indication of the feasibilty of the method of

adjustment for cross-linguistic phonetic research. We sought answers to two questions: (1) can listeners do the task in a reasonable amount of time and with reasonably low within- and between-listener variability? and (2) will the task provide interpretable data for cross-linguistic comparisons?

Subjects. Ten female and four male university students served as volunteer subjects. They had self-reported normal speech and hearing and had recently completed a one-quarter course in phonetic transcription. This pool of subjects represented fairly diverse linguistic backgrounds: four were monolingual Southern Californians, six were English-dominant Southern Californians, one was a native of Maryland, two were native speakers of Serbo-Croatian, and one was a native speaker of Spanish. We will present data collected from the Southern Californians and the Serbo-Croatians.

Materials. 330 steady-state isolated vowel stimuli were synthesized using a software formant synthesizer (Klatt & Klatt, 1990). The stimuli were generated by varying F1 and F2 independently over a large range of values. There were fifteen possible values of F1 and twenty-two possible values of F2. F1 ranged from 250 Hz to 900 Hz with a step size of 0.37 Bark, and F2 ranged from 800 Hz to 2800 Hz also with a step size of 0.37 Bark. These formant step sizes are slightly larger than the just-noticeable-differences reported by Flanagan (1957). Figure 1 shows the stimuli used in Experiment 1 (described below) and illustrates the type of two dimensional array also used in the preliminary experiment, although with a slightly larger range of possible formant values.

Other parameters of the synthesizer were fixed across the set, or were estimated by rule given the values of F1 and F2. The duration was fixed at 250 ms. and the F0 started at 120 Hz and fell over the last half of the vowel. F3 was estimated by the regression formulas published by Nearey (1989). The fourth formant was constrained to be at least 300 Hz higher than F3 and no lower than 3500 Hz. The bandwidths of the formants were estimated by regression formulas relating the bandwidth values suggested by Klatt (1980) to F1, F2 and F3. The formulas for the

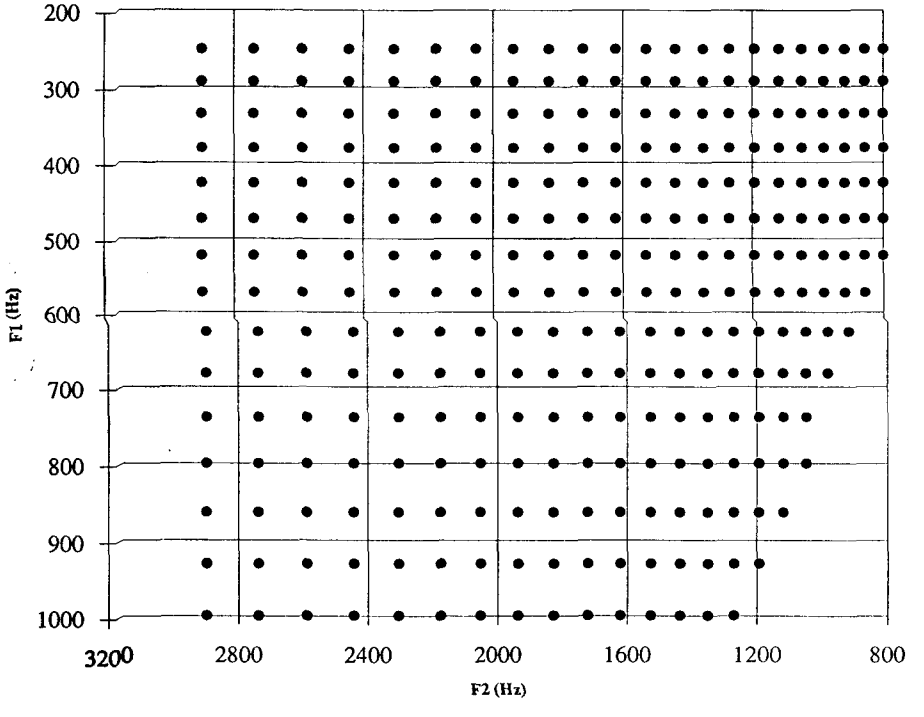


Figure 1 Formant values of stimuli used in the method of adjustment task in Experiment 1. Each filled circle represents a stimulus. The stimuli are equidistant on the Bark scale and therefore are separated by larger Hz intervals as the frequencies of F1 or F2 increased.

bandwidths of F1 through F3 are given in (1) through (3) respectively.

(1) B1 (in Hz) = 29.27 + 0.061*F1 - 0.027*F2 + 0.02*F3, r^2=0.605
(2) B2 (in Hz) = -120.22 - 0.116*F1 + 0.107*F3, r^2=0.497
(3) B3 (in Hz) = -432.1 + 0.053*F1 + 0.142*F2 + 0.151*F3, r^2=0.595

As the r^2 values suggest, formulas (1)-(3) provide only a rough fit to the bandwidth values suggested by Klatt, and extreme formant values resulted in unnatural bandwidths. This contributed to the unnaturalness of stimuli which were already unnatural due to their formant values, while simultaneously contributing to the naturalness of tokens which had humanly possible combinations of formants. The bandwidth of F4 was fixed at 200 Hz. To further increase the naturalness of the stimuli, the "natural" voice-source in the synthesizer was used and the values of amplitude of aspiration, open quotient and glottal tilt varied over time to simulate the changes in glottal vibration seen in naturally produced syllables (see Klatt & Klatt, 1990).

In addition to these synthetic stimuli used in the perception part of the experiment, we compiled a list of common English words illustrating the vowels of English for use in the production part of the experiment. The list was: heed, hid, aid, head, had, HUD, odd, awed, owed, hood, who'd. These words have either a glottal stop or /h/ initially and a final /d/. They were also used as the visual stimuli in the perception experiment.

Procedure. The experiment involved two tasks. First, the subjects were asked to read ten repetitions of the list of English words (in the carrier phrase say ___ again). The order was randomized separately each time through the list. The subjects were seated in a sound booth and recordings were made using high quality equipment (Sennheiser microphone, Symetrix SX202 preamplifier, and Tascam 122 cassette recorder). Formant values from these recorded utterances were measured using CSpeech (Paul Milenkovic) from an LPC spectrum which was calculated at a point early in each vowel as determined in a digital waveform display.

The second task was the method of adjustment task using the same list of words as visual stimuli. This part of the experiment was run online by an IBM PC-AT. Stimuli were stored on disk and were converted to analog waveforms by a Data Translation DT2801A board in the PC. The sampling rate was 10kHz and the signal was low-pass filtered at 4.2 kHz before being amplified (BGW Systems, Model 85) and presented diotically over headphones (Sony MDR-V4). (For more details concerning the setup see Johnson & Teheranizadeh, 1992.) The listener saw a word at the top of a CRT screen, and a two dimensional grid (see Figure 2 for a sample display). Each square in the grid corresponded to one of the vowel sounds in the F1, F2 matrix, and the listener used a mouse to select a particular square and clicked a mouse button to hear the synthetic vowel associated with that square. The task was then to find the location in the grid which produced a synthetic vowel which sounded like the vowel in the word. After choosing the F1 and F2 values for the vowel of a particular word, the listeners were asked to rate their choice on a scale from one to ten. The rating data were used to eliminate mistakes, primarily accidental terminations of trials. This task was repeated 10 times for each of 11 words in the list.

Note one complication in relating the grid shown in Figure 2 to a set of vowel stimuli such as that shown in Figure 1. In the region of the vowel quadrangle where F1 and F2 are close to each other the acoustic vowel space has a corner cut off (F1 was always at least 250Hz below F2), while in the visual display this is not true. This complication was handled by filling in the corner of the visual display with copies of nearby tokens. If the listener were to choose the square which would correspond to an F1 of 1000Hz and an F2 of 1200Hz, for example, a token with the same F1 and the next higher F2 value would be presented. So, the vertical dimension of the display always corresponded to different F1 values but, in one corner of the space, changes in the horizontal dimension of the grid did not result in changes of F2.

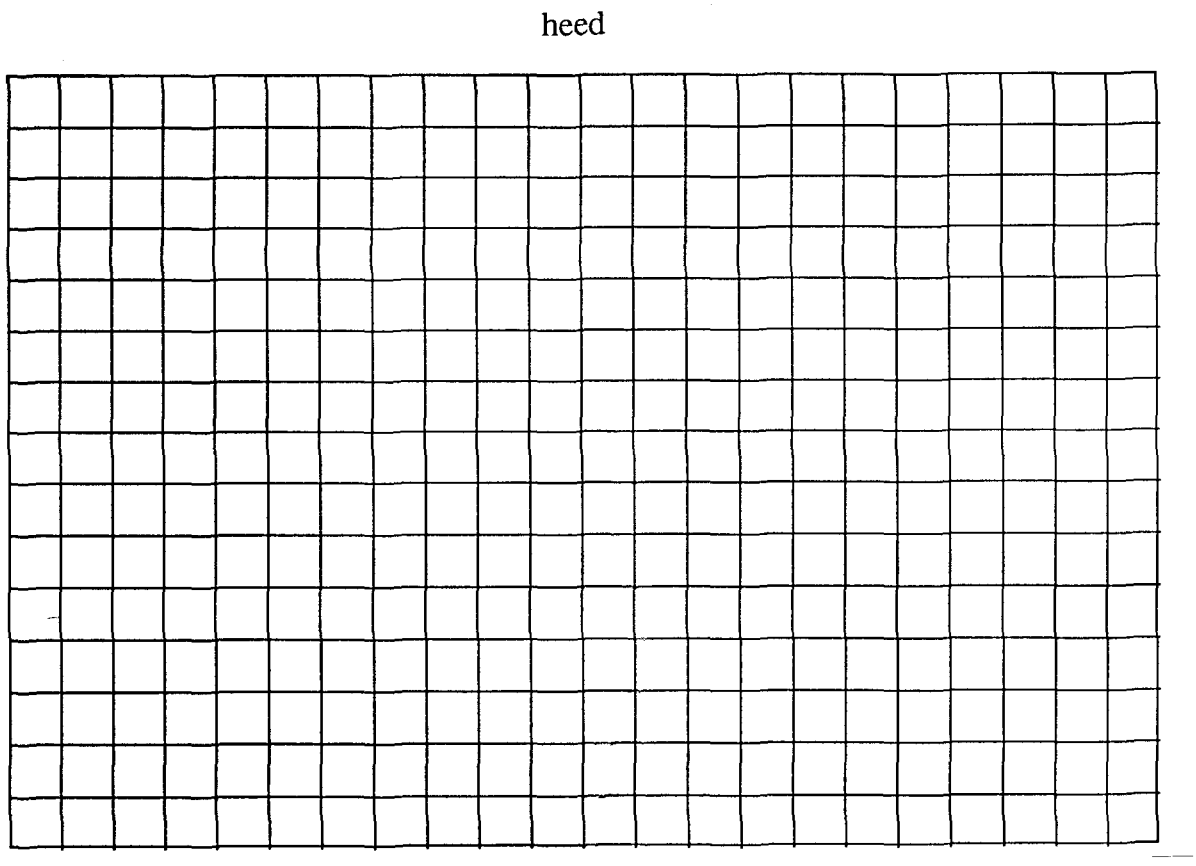


Figure 2 An example of the visual display presented to listeners in the method of adjustment task. The word at the top of the screen changed from trial to trial, and each square (except in the region of vowels with low F2 and high F1) in the display corresponded to a different F1, F2 combination (see text for further details).

Because we wanted to collect several adjustment trials for each of several vowels, and we did not want the listeners to simply rely on visual cues in making their judgements, we changed the orientation of the acoustic vowel space on the screen randomly from trial to trial. So, on 50% of the trials the high F1 stimuli were at the top of the screen and on the other 50% of the trials they were at the bottom of the screen. Similarly, the relationship of F2 to the horizontal dimension of the grid also changed from trial to trial.

Results and Discussion. Table 1 shows the standard deviations in the method of adjustment task for the native Californians. The first two columns show the between-listener standard deviations (calculated across all responses) of F1 and F2 while the next two columns show the average within-listener standard deviations of F1 and F2 (calculated for each listener separately and then averaged). The average ratio of within- to between-listener standard deviation for F1 is 0.83. The ratio of within- to between-listener standard deviation for F2 is 0.73. These ratios suggest that most of the variability in the method of adjustment task occurred within the responses of individual listeners rather than appearing as between-listener variability in the formant values chosen. Note also that standard deviations tend to be higher for higher formant values (SDF2 is higher than SDF1). This reflects the fact that the equal Bark increments in the stimulus set (Figure 1) resulted in larger acoustic differences between the stimuli as the frequency increased. One surprizing observation to be noted in Table 1 is that /i/ showed more between-listener variation in F2 than did the other vowels and /u/ showed more between-listener variation in F1. Listeners were internally consistent in their choices for these vowels but showed relatively more discrepancy

Table 1. Comparison of variability between and within listeners in the preliminary experiment. The first two columns show the overall standard deviations (in Hz) of F1 and F2 in the method of adjustment trials. The middle two columns show the average within-subject standard deviations (in Hz) for the same data. The last two columns show the ratio of within to between listener variability.

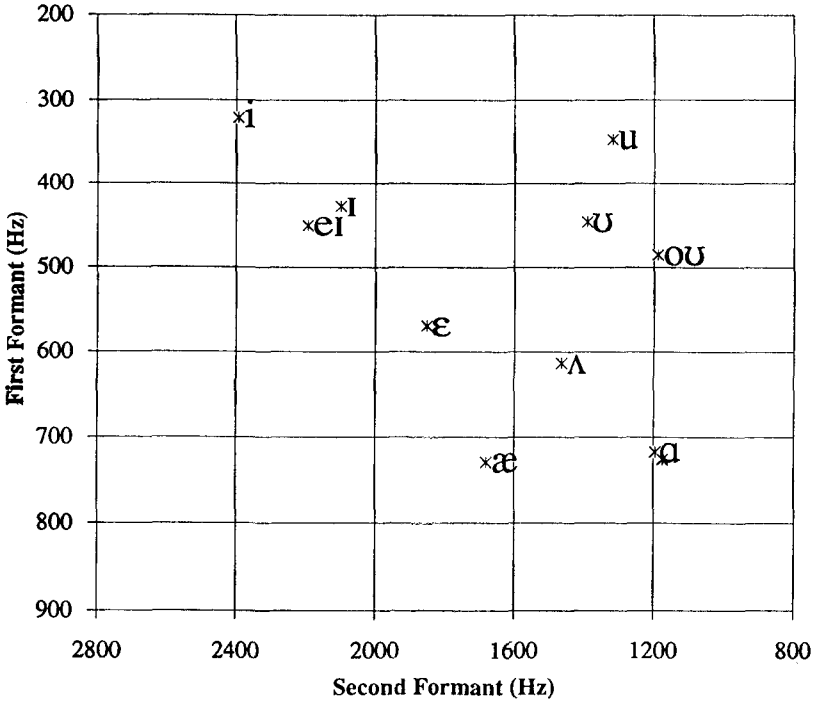
	Between Subj.		Within Subj		Ratios	
word	SDF1	SDF2	$\overline{SDF1}$	$\overline{SDF2}$	F1	F2
heed	19.5	164.6	16.2	99.9	.83	.61
hid	29.7	158.4	26.3	124.9	.89	.79
aid	55.9	229.6	37.1	151.2	.66	.66
head	49.7	173.0	41.7	114.6	.84	.66
had	61.4	175.5	56.4	126.3	.92	.72
odd	56.5	56.6	50.0	52.2	.89	.92
awed	55.7	42.4	48.8	35.8	.88	.84
HUD	48.9	98.5	41.1	70.6	.84	.72
owed	35.5	57.6	31.5	46.0	.89	.80
hood	36.2	96.9	35.7	85.8	.99	.89
who'd	67.9	112.4	41.9	86.2	.61	.77
average	47.0	124.1	38.8	90.3	.83	.73

between listeners than with most of the other vowels. Also note that the choices for F1 and F2 of aid were more variable between listeners than they were within listeners. This is probably due to the fact that the synthetic stimuli were steady-state vowels while the target vowel is diphthongal. The vowel in owed is less diphthongal in this dialect than it is in other dialects of English. Although there are some interesting patterns in the variability found in this preliminary study, the most important finding is that the variability is low; averaged standard deviations of about 50Hz for F1 and 100Hz for F2.

Measurements of the acoustic vowel space produced by 8 Southern Californian males in the citation readings of Experiment 1 (shown in Figure 3a) suggested that the vowels in odd and awed are merged. (These production data from Experiment 1 are shown here because there were too few male subjects in the preliminary experiment and because it is necessary to compare these method of adjustment results with male formant values because the synthetic stimuli had a male voice.) Note also that /eɪ/ and /ɪ/ had very similar formant values.

The merger of the vowels in odd and awed was also found in the perception results (Figure 3b) from the native Southern California subjects in the preliminary experiment. In addition to the merger of the low back vowels, the data indicate that the listeners kept the vowels of aid and hid more spectrally separated in the method of adjustment than they did in production. This tendency was noted in an earlier method of adjustment study of vowels (Johnson, 1989). As suggested in that earlier report, it may be that when potential cues such as intrinsic vowel duration, pitch, and formant movement are not available, listeners will exaggerate an existing small spectral difference in the method of adjustment in order to maintain a linguistic distinction. There is also the possibility that the production data in Figure 3a don't accurately represent the spectral properties of /eɪ/ because of our choice of measurement location. This possible explanation of the discrepancy between the production and perception result seems rather unlikely because we made the acoustic measurements early in the vowel. Therefore, we would expect if anything to see an even lower F1 and higher F2 for /eɪ/ if we were to measure later in the vowel. Thus, if we were to make the formant measurements later in the vowel we would expect to see even more discrepancy rather than less.

(a)



(b)

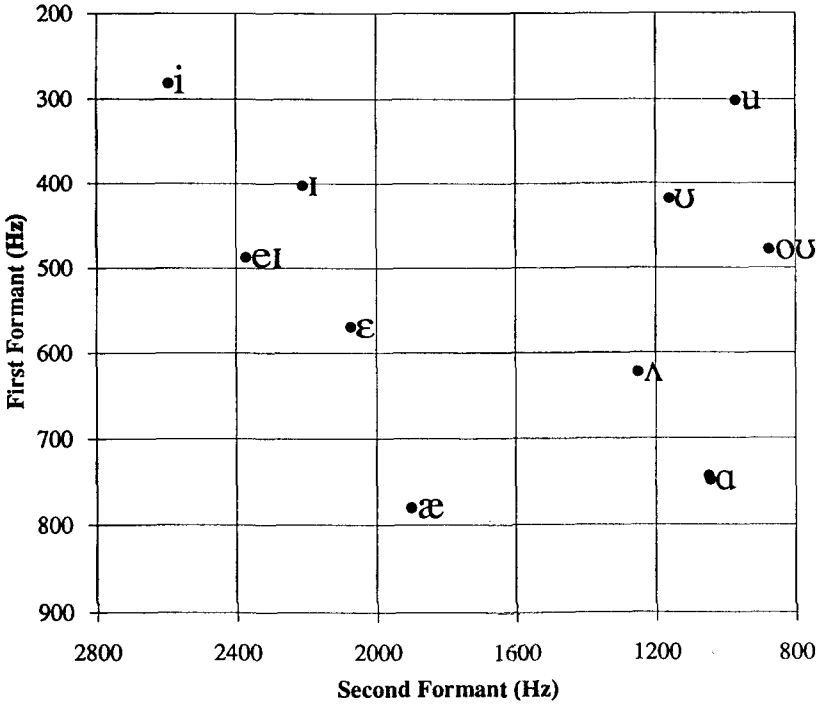
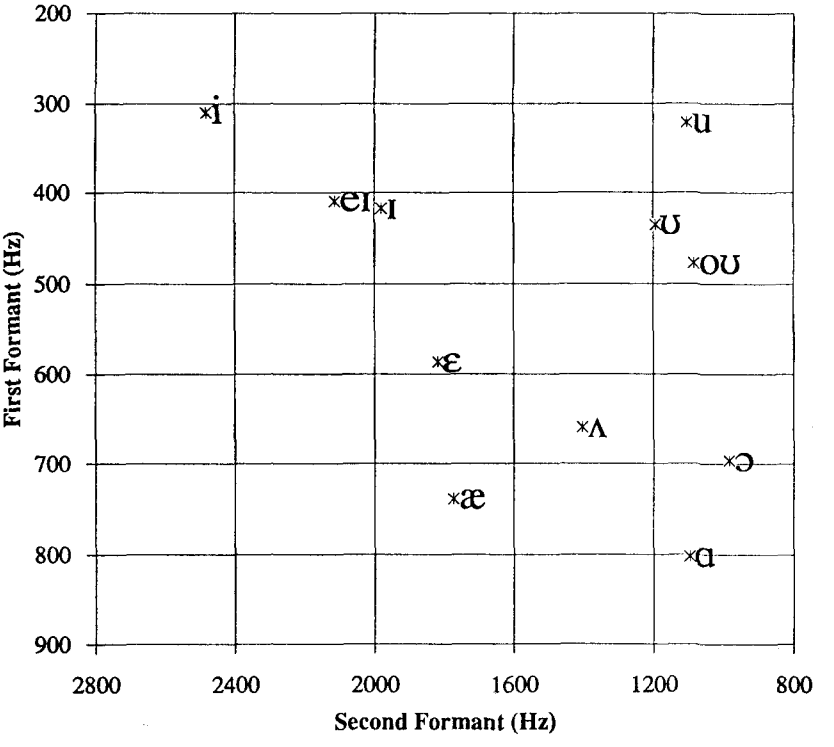


Figure 3 (a) Average measured formant values of vowels produced by the eight male native Southern California English speakers from Experiment 1. These vowels were produced in the "citation" reading condition. (b) Average method of adjustment results for the native Southern California English speakers in the preliminary experiment.

(a)



(b)

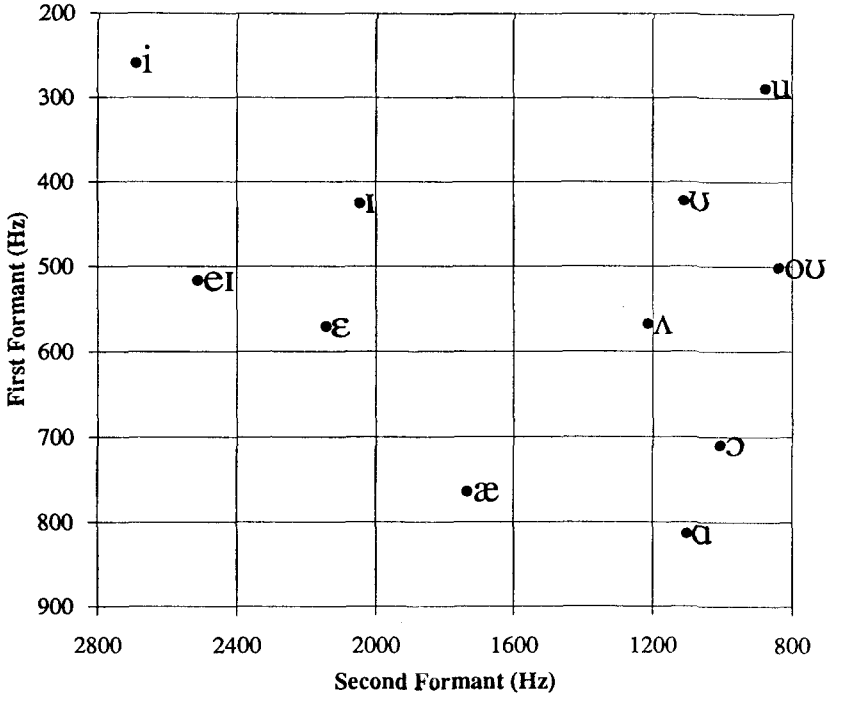
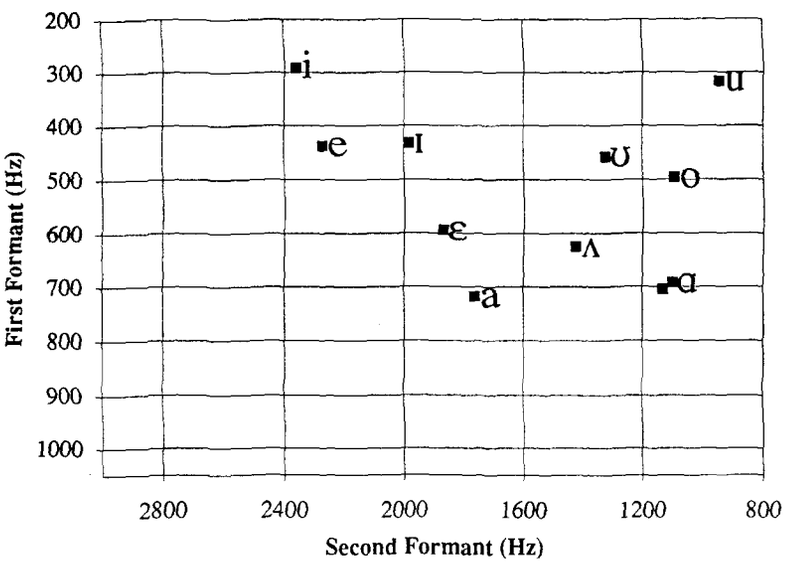


Figure 4 (a) Average measured formant values of vowels produced by one male speaker in the preliminary experiment. This speaker maintained a distinction between the vowels in awed and odd. (b) Average method of adjustment results obtained from the same speaker.

One of the listeners in the preliminary experiment was older than the other listeners and was born in New York City. Figure 4a shows that he maintained the distinction between odd and awed in production. In other respects his acoustic vowel space is similar to the average vowel space shown in Figure 3a. Figure 4b shows that this subject also selected different formant values for the vowels in odd and awed in the perception experiment (the difference between /eɪ/ and /ɪ/ was also quite expanded in the perception space). Keep in mind that although the comparison between production vowel spaces (Figures 3a & 4a) gives us about the same picture of the difference between this speaker and the others, the perception vowel spaces (Figures 3b & 4b) allow for a better comparison because the listeners are telling us their expectations for the vowel space of a single synthetic speaker, thus speaker and dialect information are not confounded.

(a)



(b)

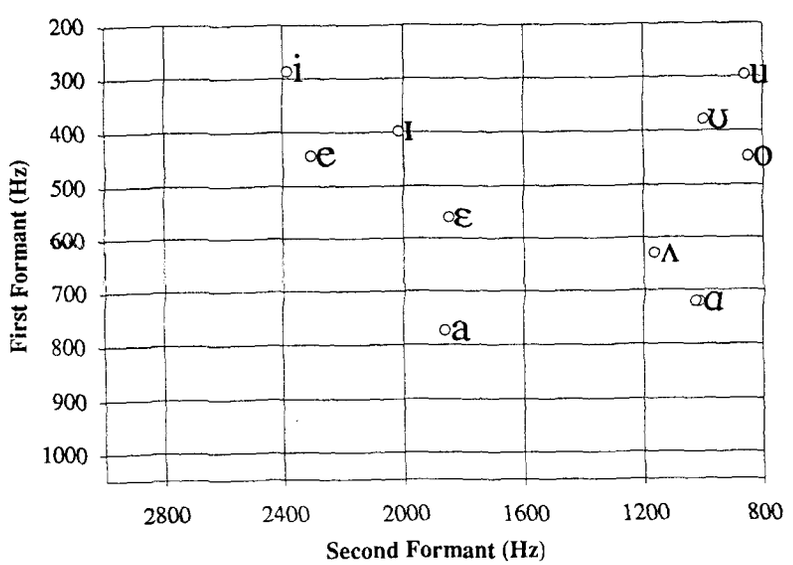


Figure 5 (a) Average measured formant values of English vowels produced by the male Serbo-Croatian speaker. (b) Average method of adjustment results for English vowels obtained from the male Serbo-Croatian speaker.

Two of the listeners in the preliminary study were native speakers of Serbo-Croatian who moved to Los Angeles at different ages. The male speaker was 7 years old when he moved to LA, and both his production and perception data for English were generally comparable to the native English speakers' data (Figure 5). One interesting difference is that his back vowels in both production and perception have lower F2 values than the native speakers. This corresponds to the lower F2 found for Serbo-Croatian /u/ and /o/.

The female Serbo-Croatian speaker (who had moved to Los Angeles at the age of 19) showed a much more striking pattern of deviation from the native Southern Californians (Figure 6). Her production vowel space showed some interesting deviations from the native speakers' space. In particular, the vowel in had was more central and lower and the vowel in owed was also lower (compare Figure 6a with Figure 3a). Also, her back vowels had relatively lower values of

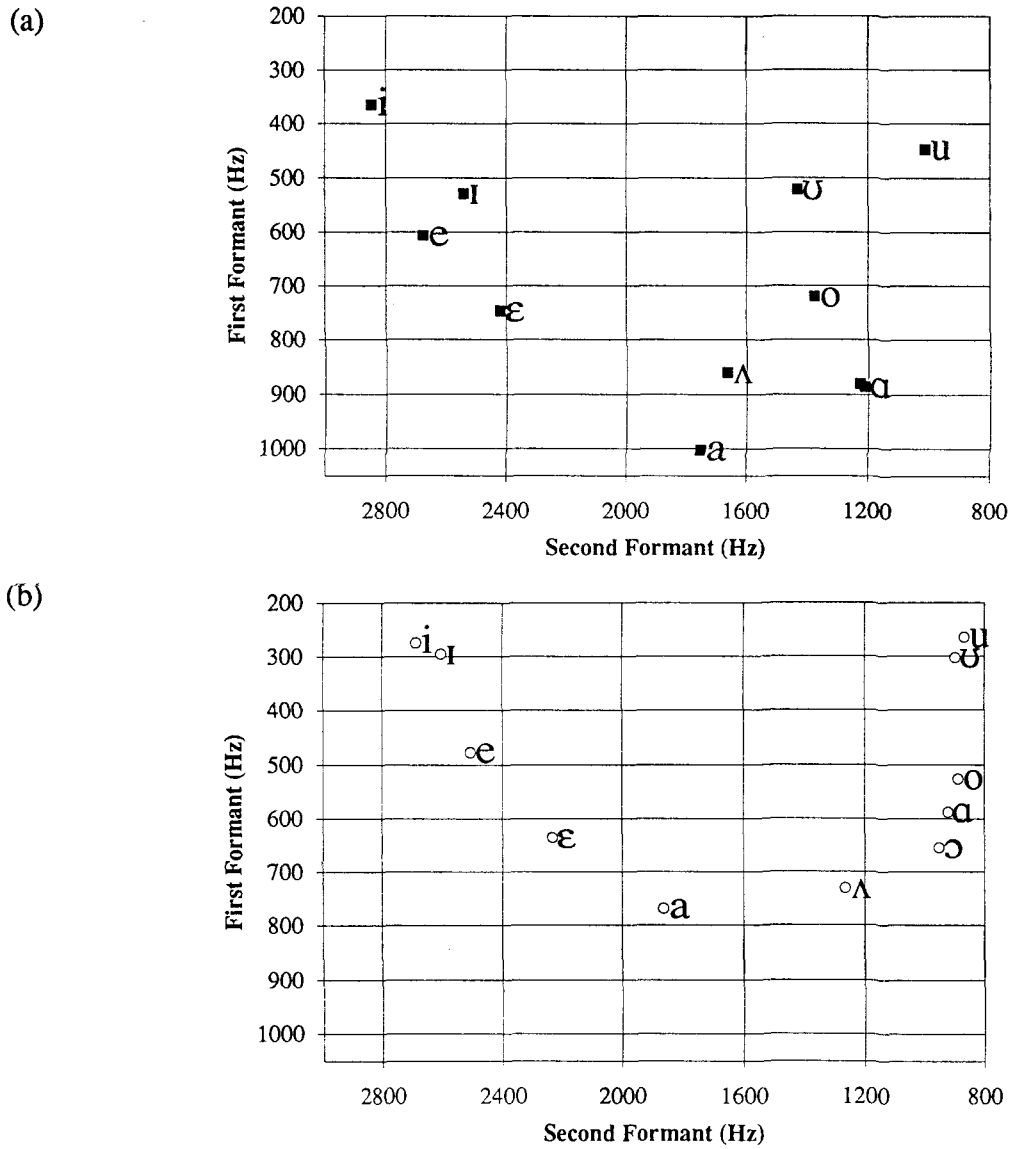


Figure 6 (a) Average measured formant values of English vowels produced by the female Serbo-Croatian speaker. (b) Average method of adjustment results for English (solid circles) and Serbo-Croatian (open circles) vowels obtained from the female Serbo-Croatian speaker.

F2, similar to the pattern seen in the male Serbo-Croatian speaker's production and perception results. However, the most surprizing result from this speaker has to do with her perception vowel space (Figure 6b). She chose peripheral formant values for all of the English vowels and merged the high tense/lax pairs. Figure 6b also shows this speaker's average responses when asked to select vowel sounds for words illustrating the five vowels of Serbo-Croatian. Comparison of her English and Serbo-Croatian perception data indicates that the high vowels of English were associated with the high vowels of Serbo-Croatian but that the rest of the vowels of English (with the possible exception of /o/) were assigned unique values of F1 and F2. Interestingly, however, the entire English vowel space was restricted to the periphery of the acoustic vowel space. This pattern suggests that the listener's perceptual expectations for second language may be influenced globally as well as locally by the first language, and this despite the second language speaker's relative success in producing the second language.

Experiment 1: Instruction set

The preliminary data suggest that the method of adjustment may be a useful tool in the study of dialect differences, cross-linguistic differences and second-language acquisition. In the remainder of this paper we will focus on a methodological puzzle and its significance both for the use of the method of adjustment in studying vowel spaces and for theories of phonetic representation.

The puzzle is illustrated by a comparison of the average perception vowel space from the preliminary experiment (Figure 3b) with the average citation production space from the 8 male speakers in Experiment 1 (Figure 3a). This comparison shows that the vowel space chosen in the method of adjustment was expanded relative to the production vowel space. In other words, listeners' expectations for vowels produced by a male synthetic voice were quite different from the vowels as actually produced by male speakers in citation speech. This is a conundrum if we assume that listeners' perceptual expectations are based on experience.

There were a couple of aspects of the preliminary experiment which made us doubt the validity of this discrepancy between production and perception vowel spaces. The listeners were not phonetically naive; they had completed an undergraduate course in phonetics and thus knew the cardinal vowel system. So, they may have been inclined to select extreme cardinal vowel qualities in the method of adjustment task while more naive speakers might choose formant values more similar to those found in production. Additionally, we suspected that the instructions given to the listeners may have biased them toward extreme vowel qualities. We asked the listeners to find the "best" vowel sound for each word. After the fact we realized that this instruction could have been interpreted to mean, "find the most distinct example of the vowel".

Experiment 1 was designed to investigate these issues by using (1) naive listeners and (2) a careful manipulation of instruction set. One group of listeners was instructed to find the best example of the vowel in each word (the *best* condition). While another group of listeners was instructed to find the vowel sound which most closely matched their own pronunciation of the vowel in each word (the *as you say it* condition).

Subjects. Ten females and eight male university students were recruited through the university newspaper and paid a small sum for their participation. They were monolingual English speakers who reported normal speech and hearing ability and had attended high school in Southern California. The subjects were divided into two groups as described below, with five females and four males in each group.

Materials. As in the preliminary experiment, 330 steady-state isolated vowel stimuli with fifteen possible values of F1 and twenty-two possible values of F2 were synthesized using a software formant synthesizer (Klatt & Klatt, 1990). The formant ranges in Experiment 1 (shown in Figure 1) were larger than those used in the preliminary experiment because the listeners in the preliminary experiment chose formant values which were more extreme than we had anticipated. F1 ranged from 250 Hz to 1000 Hz in increments of 0.42 Bark, while F2 ranged from 800 Hz to 2900 Hz in 0.39 Bark increments.

Procedure. Experimental sessions in Experiment 1 were very much like sessions in the preliminary experiment. However, after the subjects had completed the perception part of the

experiment they were asked to read the word list a second time. In this second reading of the words we elicited hyperarticulated or clear-speech versions of the words by saying "what?" or "huh?" after each sentence, prompting the speaker to read each sentence again more clearly. This procedure was explained to the speakers prior to starting the tape recorder. We will call the first reading the *citation* reading and the second the *hyperarticulated* reading. One other procedural difference in Experiment 1 concerned the instructions given to the listeners in the method of adjustment task. We asked one group of listeners (5 female, 4 male) to find the best examples of the vowels and another group of listeners (5 female, 4 male) to find the vowel sound which most closely matched their own pronunciation of the vowel in each word. We will call the first instruction set the *best* condition and the second set the *as you say it* condition.

Finally, the recordings were analysed using CSL (Kay Elemetrics) rather than CSpeech. In analysing these productions we chose measurement points from spectrographic displays (where we had only used waveform displays in the preliminary experiment) and were therefore able to identify an early steady-state portion of the vowel as the point representing the acoustic vowel "target".

Results. The perception results from Experiment 1 are shown in Figure 7a. In separate repeated-measures analyses of variance there were no reliable effects of instruction set on the choices made by the listeners on F1 or F2 for any of the vowels. The most robust difference as a function of instruction set was for the F1 of awed which tended to be greater in the *as you say it* group than in the *best* group (806 Hz versus 758 Hz respectively) but this difference was only marginally reliable ($F[1,16]=3.19$, $p=0.093$), no other differences between groups proved to be statistically reliable.

Although the vowel spaces chosen were not affected by instruction set, the ratings given to the synthetic stimuli (shown in Table 2) were. Listeners in the *best* condition were consistently more critical of the stimuli than were the listeners in the *as you say it* condition. The statistical comparison of these conditions was complicated by ceiling effects in the rating data, but the trend is clear. The result is that the instruction set manipulation had an effect on ratings, but it did not have an effect on the formant values chosen in the method of adjustment.

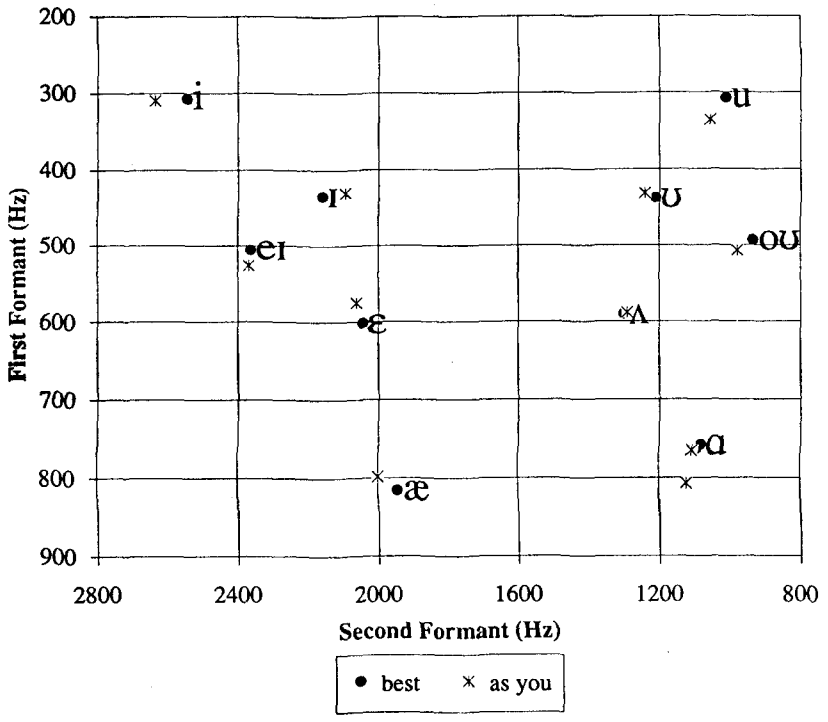
In addition, a comparison (shown in Figure 7b) of the perceptual vowel spaces of naive listeners from Experiment 1 and phonetically trained listeners from the preliminary experiment

Table 2. Rating values given to synthetic vowels in the perception part of Experiment 1. Rating values (on a scale from 1 to 10) were averaged without including vowels rated 1 because the listeners were asked to use 1 to indicate that they had accidently terminated a trial early. Data are presented by vowel and by listening condition.

word	*best*	*as you say it*
heed	8.9	9.7
hid	**8.6**	9.0
aid	**8.7**	9.5
head	**8.9**	9.4
had	**8.6**	9.4
odd	**9.0**	9.7
awed	**8.8**	9.9*
HUD	**8.8**	9.5
owed	**8.1**	9.8**
hood	**8.2**	9.4**
who'd	**8.6**	9.2
average	8.7	9.5

*p < 0.1 **p < 0.05

(a)



(b)

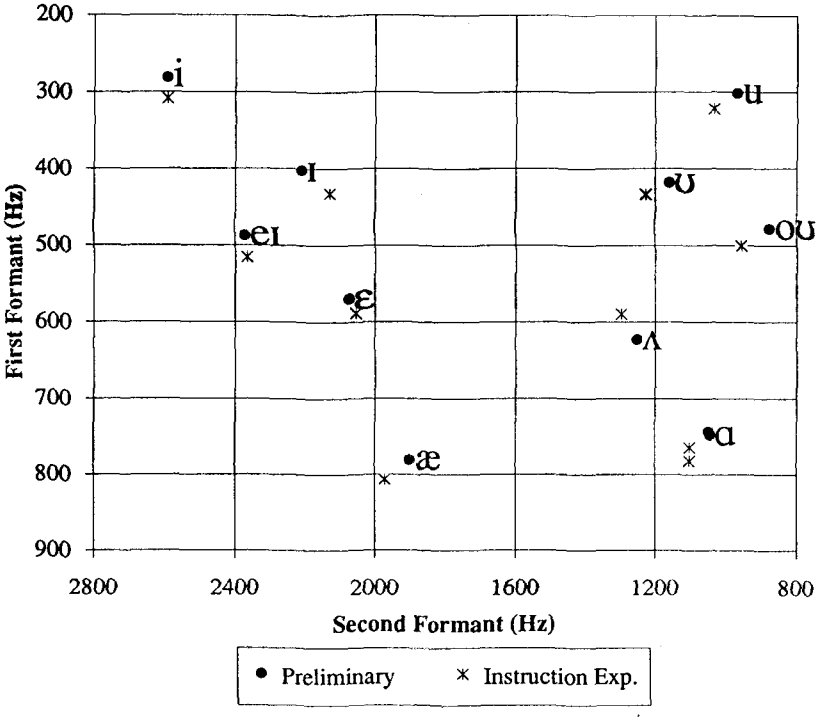


Figure 7 (a) Method of adjustment results of the instruction set experiment. Filled circles are the average values chosen by the listeners in the *best* group, and stars are the average values chosen by the listeners in the *as you say it* group. (b) Comparison of the acoustic vowel spaces chosen by listeners in the preliminary experiment and in Experiment 1. Filled circles are the average responses of the listeners in the preliminary experiment, and stars are the responses (averaged across instruction condition) of the listeners in Experiment 1.

averaged over instruction conditions suggests that phonetic training (at least from the first author) had no effect on the results of the method of adjustment task. It is not valid to attempt a statistical comparison of the data shown in Figure 7b because there were several small changes in the method (particularly the range of possible F1/F2 combinations was expanded in Experiment 1). Still the differences appear to be of the same magnitude as the nonsignificant differences found as a function of instruction set (Figure 7a) and are certainly nothing like the differences between measured values from citation forms and the method of adjustment results (Figure 3a versus Figure 3b).

So, the perceptual vowel space resulting from the method of adjustment task is robust. Speakers of the same dialect give the same answers regardless of instruction set or their previous training in phonetics. This robust pattern of performance raises an interesting question. Namely, what in the experience of the listener underlies the method of adjustment vowel space which is so different from the acoustic vowel space found in normal productions of the same words?

One hypothesis is that the perceptual vowel space which is found in the method of adjustment task reflects hyperarticulated versions of the vowels, rather than the vowel qualities found less carefully produced speech. We will call this the hyperspace hypothesis. With this hypothesis in mind, we asked the speakers to read the word list in a hyperarticulated style. As has been found before (Picheny, Durlach, & Braida, 1986; Moon & Lindblom, 1989), the hyperarticulated versions of the vowels had generally more extreme vowel formants than did the less carefully produced vowels (see Figure 8a). This is just the sort of vowel space expansion that we saw in comparing the perception results with citation readings of the words.

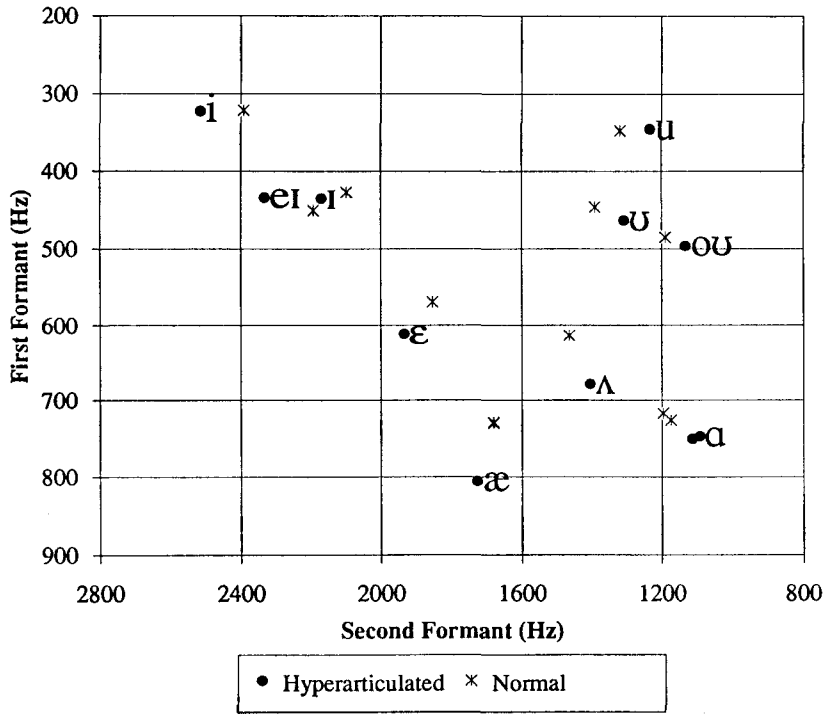
A comparison of the average vowel formants in hyperarticulated productions and the method of adjustment results from Experiment 1 (shown in Figure 8b) suggests that the hyperspace hypothesis is on the right track. Further, when we looked at the hyperarticulated vowel spaces for individual speakers we found that all of the formant values chosen in the perception task were represented in the productions of at least one speaker. Thus, it seems that the listeners' responses in the method of adjustment experiment are not only robust, but are also based upon their experience of very clearly articulated versions of the vowels.

Experiment 2: Intrinsic F0 and duration

One factor which may have had an effect on the method of adjustment results in both the preliminary experiment and in Experiment 1 is that the stimuli were impoverished relative to natural speech. While in English vowels differ in intrinsic pitch, duration and formant trajectories (Peterson & Barney, 1952, Peterson & Lehiste, 1960, Lehiste & Peterson, 1961), the stimuli which we used in the method of adjustment task did not vary along these dimensions. The stimuli were impoverished in this way because we were interested in designing a tool for the cross-linguistic comparison of vowel spaces, and therefore we avoided maipulating pitch, duration and formant trajectories because the redundancies oberserved in English vowels are not cross-linguistic universals of vowel systems (Lehiste, 1970; Keating, 1985). This decision complicates any interpretation of the method of adjustment results because listeners may have attempted to compensate for a loss in the overall distinctiveness of the different vowel qualities (resulting from the absence of redundant cues) by increasing distinctiveness in the spectral domain. Therefore, the hyperspace effect may have been an artifact of the experimental design. We tested this possibility in a second experiment.

In Experiment 2, American English intrinsic vowel F0 and duration were modelled in the synthetic stimuli used in a method of adjustment study. F0 and duration were made to vary as a function of F1 and F2 in a way which is similar to their observed variation in English. Thus, some portion of the redundant information which was missing from the stimuli used in the preliminary experiment and in Experiment 1 was present in these stimuli. If the hyperspace effect occurred in these earlier experiments because of the lack of redundant information in the stimuli, we should find a reduction (but probably not a total elimination) of the effect in Experiment 2.

(a)



(b)

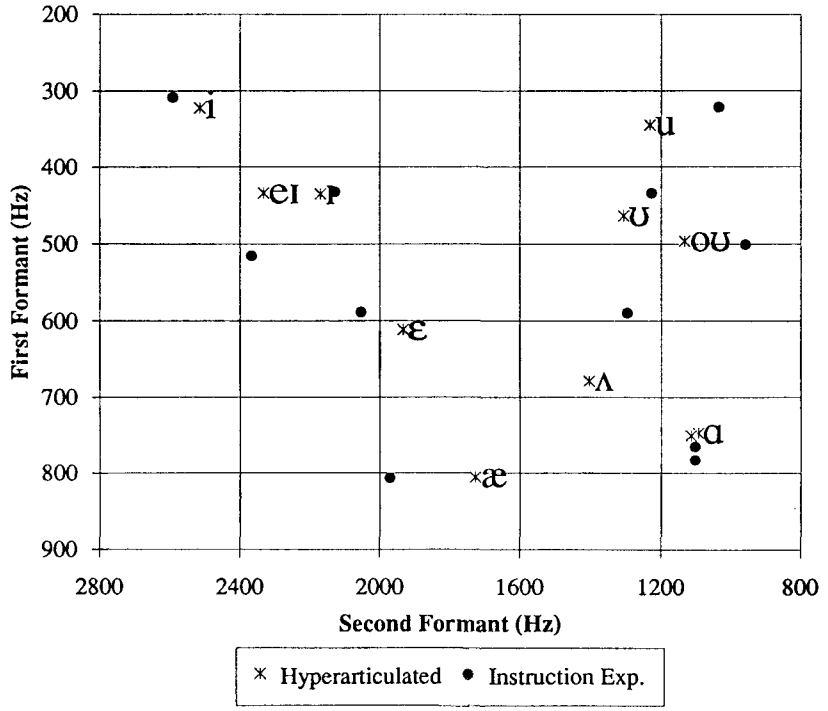


Figure 8 (a) Average measured formant values of the vowels produced by the eight male speakers in Experiment 1. Filled circles are the average values in the hyperarticulated condition and stars are the average values of the vowels in the citation-form condition. (b) Comparison of hyperarticulated productions (8 male speakers from Experiment 1) with method of adjustment results. Stars are the average measured formant values from hyperarticulated versions of the vowels and filled circles are the method of adjustment results (averaged across listener and instruction condition) from Experiment 1.

Subjects. Two male and one female native speakers of Southern Californian English (by the criteria used in Experiment 1) volunteered for the experiment. The listeners reported normal speech and hearing abilities and had completed two introductory phonetics courses. Because we found no differences in performance in the task as a function of phonetic training between the preliminary experiment and Experiment 1, these listeners were taken to be representative of Southern California English.

Materials. As in the earlier experiments, 330 isolated steady-state vowels were synthesized. The formants and bandwidths were the same as those in the stimuli synthesized for Experiment 1, however F0 and duration varied from stimulus to stimulus rather than being fixed as they were in the earlier experiments.

The method used to derive F0 and duration values for the stimuli was analogous to that used in the earlier sets to derive bandwidth values by rule (formulas 1-3 above). Average F0 and formant values for male speakers from Peterson & Barney's (1952) study of American English vowels were entered into a regression analysis in which F0 was predicted by F1 and F2. The resulting regression formula, shown in (4), indicates that F0 is negatively correlated with both F1 and F2. As a result of using this formula to calculate F0 values for the synthetic stimuli, F0 ranged from 110Hz to 142Hz. As in the earlier experiments, F0 was steady over the first half of the vowel and then fell gradually to about 85% of its original value over the last half.

Similarly, average duration measurements from Peterson & Lehiste (1960) and formant measurements from Peterson & Barney (1952) for tense vowels in English were analysed and a regression formula (5) was calculated for duration as a function of F1 and F2. The resulting durations ranged from 210ms to 305ms for the range of F1, F2 combinations in the vowel array. The duration equation is problematic for English because lax vowels have much shorter durations than their tense counterparts, even though their formant values are comparable. Thus, the duration formula only captures vowel variation which is correlated with F1 and F2 variation and the duration differences between tense and lax vowels are not captured by the manipulation. All stimuli had 50ms on- and off-ramps for the amplitude of voicing.

(4) F0 (in Hz) = 153.44 - 0.035*F1 - 0.00275*F2, $r^2 = 0.939$

(5) Dur (in ms) = 191.754 + 0.121*F1 - 0.00347*F2, $r^2 = 0.792$

Procedure. Unlike the earlier experiments, no production data were collected in Experiment 2. The method of adjustment task was conducted using the same equipment and software as in the earlier experiments. The listeners were instructed to find vowels which sounded like the ones they produced in the words (the *as you say it* condition of Experiment 1). Each of the eleven English words used in the earlier experiments was presented in random order 7 times.

Results and Discussion. Figure 9 shows the results of Experiment 2 compared with the average results of Experiment 1. This figure indicates that there were only very minor differences between the vowel formants chosen when F0 and duration varied as a function of F1 and F2 and the vowel formants chosen when these two redundant parameters were held constant across the vowel array. These results suggest that the expanded vowel space which was found in the preliminary experiment and in Experiment 1 was not an artifact of the synthetic stimuli. If the absence of redundant information such as intrinsic F0 differences or duration differences between vowels had caused an expansion of the vowel space we would have expected some contraction of the vowel space in this experiment. This result did not occur; the hyperspace effect persisted.

Conclusion

Why do listeners choose a hyperspace in the method of adjustment task? One answer is that the experimental situation biases listeners in this direction. Although we have tried to test some ways in which this might have happened (instruction set, phonetic training, and lack of redundant cues), there may still be some aspect of the stimuli or of task itself which biases listeners toward a hyperspace. For example, if we were to present synthesized versions of whole words

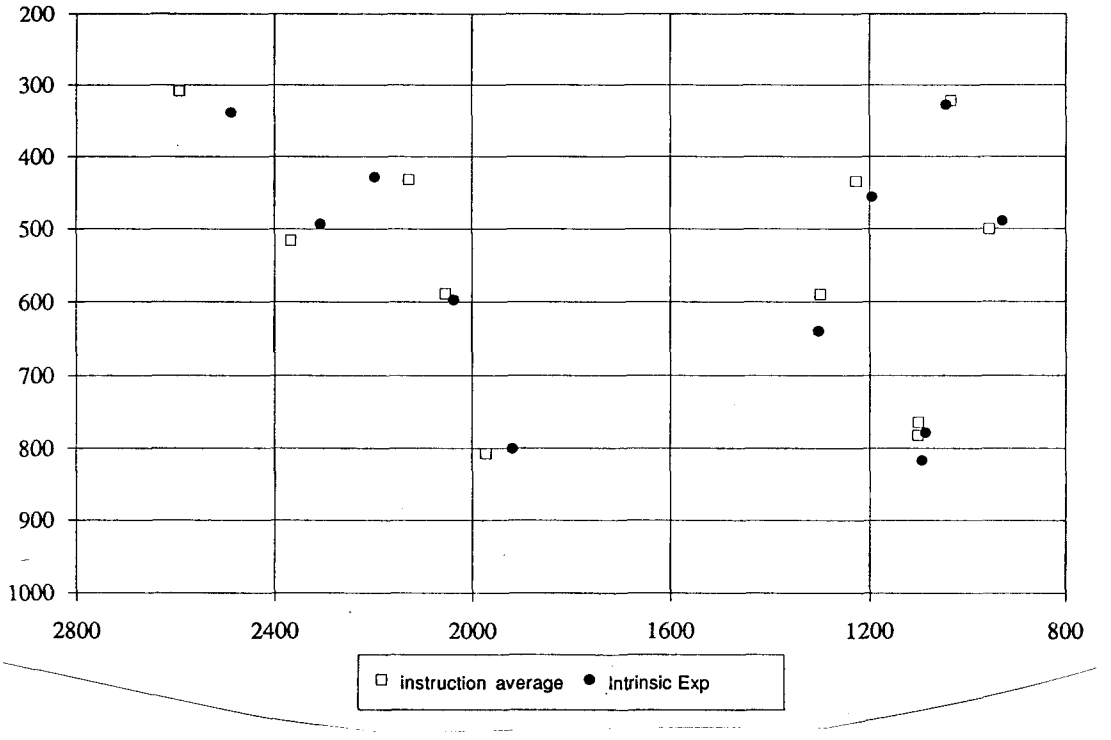


Figure 9 Average method of adjustment results from Experiment 2 (filled circles) and Experiment 1 (open squares).

rather than isolated vowels, listeners might be inclined to choose less extreme vowel qualities (Lindblom & Studdert-Kennedy, 1967). This hypothesis, although reasonable, is probably not right because measured formant values from /hVd/ contexts are very similar to those found in isolated vowel productions (Peterson & Barney, 1952).

One other explanation of the effect is that the formality of the test situation biased the listeners toward a hyperspace in the method of adjustment task. While it is difficult to elicit casual speech in experimental situations, it should be noted that it is also quite difficult to elicit hyperarticulated speech. We found that speakers, when simply instructed to speak clearly, would initially produce quite hyperarticulated speech, but after only a few utterances would revert back to the same style of speaking that they used in the citation reading. This is why we adopted a special procedure to elicit hyperarticulated speech in Experiment 1. So, it is not clear why hyperarticulation would be the *listener's* response to the formality of an experimental situation and not the *speaker's*.

The results reported here suggest that hyperarticulated versions of speech sounds are more basic than less clearly articulated versions of those same sounds (or as one of our speakers put it, the hyperarticulated versions are the *real* sounds). Jakobson & Halle expressed this view when they said, "The slurred fashion of pronunciation is but an abbreviated derivative from the explicit clear-speech form which carries the highest amount of information. ... When analyzing the pattern of phonemes and distinctive features composing them, one must resort to the fullest, optimal code at the command of the given speakers" (1956, p. 6). The hyperspace result can be interpreted as empirical support for this generally assumed, although not explicitly defended, point of view since the notion that speakers utilize hyperarticulated phonetic representations or targets provides a natural rationale for the peripheral vowels selected by the listeners in the method of adjustment.

An alternative account is that phonetic implementation rules are context sensitive. For instance, the feature [+high] might be realized as particular F1 targets which differ depending on prosodic context or the degree of effor the speaker is willing to expend. In this model, the

parametric output is determined as a function of both the distinctive feature and various parameters of the performance context. A conceptual difficulty with this type of implementation model is that the different contextually determined realizations of a feature all have equal status as phonetic realizations of that feature (this is also a problem for Keating's, 1988 window model of coarticulation). Thus, a reduced schwa-like version of /i/ is just as good an example of a high vowel as is a hyperarticulated, maximally distinct /i/. As shown in the method of adjustment task, this runs counter to the intuitions of naive listeners.

The theory of phonetic realization must account for the wide range of realizations of the same utterance that a single speaker produces in differing situations. Some of the variation may be introduced by optional categorical phonological rules, but much of the variation is continuous and very low-level in nature, and must surely be the result of phonetic implementation. The details of a phonetic implementation model consistent with our experimental results have not been fully worked out, but an outline is clear. This type of model includes (1) a mapping from categorical representations to parametric representations corresponding to hyperarticulated speech, and (2) a second mapping from maximally distinct parametric representations to reduced forms. The first mapping is what we normally think of as phonetic implementation. It maps distinctive features to phonetic parameters like vocal tract shapes or formant values. By including a second mapping in the model, as suggested by the experimental data, some of the complications that arise in devising schemes of phonetic implementation may be alleviated. In particular, the wide range of realizations of the same utterance that a single speaker can produce does not have to be accounted for by a single mapping from phonological features to phonetic parameters. One noteworthy description of a mapping from hyperarticulated parametric representations to reduced parametric representations is Browman & Goldstein's (1986, 1989) articulatory phonology in which reduction phenomena have been described as gestural overlap, hiding, and blending. There are conceptual difficulties in interpreting a gestural model as the second mapping in a two stage model of phonetic implementation (clearly this is not what Browman and Goldstein have in mind) but there are some compatibilities between a gestural model and the requirements of the mapping because the gestures in articulatory phonology specify articulatorily extreme targets which become reduced in normal speech.

Acknowledgments

Special thanks to Doug Whalen who suggested Experiment 2, Bob Port who offered encouragement early on, and Peter Ladefoged whose interest and support inspired us.

References

Behne, D.M. (1989) A comparison of the first and second formants of vowels common to English and French. *Research on Speech Perception Progress Report no. 15*, 269-282. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.

Browman, C.P. & Goldstein, L. (1986) Towards an articulatory phonology. *Phonology*, **3**, 219-252.

Browman, C.P. & Goldstein, L. (1989) Articulatory gestures as phonological units. *Phonology*, **6**, 201-231.

Browman, C.P. & Goldstein, L. (1990) Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, **18**, 299-320.

Disner, S.F. (1980) Evaluation of vowel normalization procedures. *J. Acoust. Soc. Am*,67, 253-261.

Disner, S.F. (1986) On describing vowel quality. In (Eds.) *Experimental phonology*, ed. by J.J. Ohala & J.J. Jaeger, 69-79. Orlando: Academic Press.

Flanagan, J. (1957) Estimates of the maximum precision necessary in quantizing certain 'dimensions' of vowel sounds. *J. Acoust. Soc. Am*,**29**, 533-534.

Ganong, W.F. & Zatorre, R.J. (1980) Measuring phoneme boundaries four ways. *J. Acoust. Soc. Am*, **68**, 431-439.

Gerstman, L.H. (1968) Classification of self-normalized vowels. *IEEE Trans. Audio*

*Electroacoust.***AU-16**, 78-80.
Harshman, R. (1970) Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA Working Papers in Phonetics*, **16**.
Jakobson, R. & Halle, M. (1956) *Fundamentals of Language*. 'S-Gravenhage: Mouton & Co.
Joos, M. (1948) *Acoustic Phonetics*. Linguistic Society of America Language Monograph No. 23 (Baltimore: Waverly Press).
Johnson, K. (1989) On the perceptual representation of vowel categories. *Research on Speech Perception Progress Report no. 15,* 343-58. Bloomington, IN: Speech Research Laboratory, Psychology Department, Indiana University.
Johnson, K. & Teheranizadeh, H. (1992) Facilities for speech perception research at the UCLA phonetics lab. *UCLA Working Papers in Phonetics,* **81**.
Keating, P.A. (1985) Universal phonetics and the organization of grammars. *Phonetic linguistics: essays in honor of Peter Ladefoged*, ed. by V. Fromkin, 115-32. Orlando: Academic Press.
Keating, P.A. (1988) The window model of coarticulation: articulatory evidence. *UCLA Working Papers in Phonetics,* **69**, 3-29.
Klatt, D. (1980) Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* **67**, 971-995.
Klatt, D. & Klatt, L. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am*, **87**, 820-857.
Ladefoged, P. & Broadbent, D. (1957) Information conveyed by vowels. *J. Acoust. Soc. Am,* **29**, 98-104.
Lehiste, I. (1970) *Suprasegmentals*. Cambridge, MIT Press.
Lehiste, I. & Peterson, G.E. (1961) Transitions, glides and diphthongs. *J. Acoust. Soc. Am,* **33**, 268-277.
Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H&H theory. *Speech production and speech modelling*, ed. by W.J. Hardcastle & A. Marchal, 403-439. Dordrecht: Kluwer Academic.
Lindblom, B. & Studdert-Kennedy, M. (1967) On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am*, **42**, 830-843.
Lobanov, B.M. (1971) Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am,* **49**, 606-608.
Miller, J.D. (1989) Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am*, **85**, 2114-2134.
Moon, S.J. & Lindblom, B. (1989) Formant undershoot in clear and citation-form speech: A second progress report. *STL-QPSR*, **1**, 121-123.
Nearey, T.M. (1977) Phonetic feature systems for vowels. PhD Dissertation, University of Connecticut, Storrs, CT.
Nearey, T. (1989) Static, dynamic and relational properties in vowel perception. *J. Acoust. Soc. Am*, **85**, 2088-2113.
Nooteboom, S.G. (1973) The perceptual reality of some prosodic durations. *J. Phon,* **1**, 25-46.
Peterson, G.E. & Barney, H.L. (1952) Control methods used in a study of vowels. *J. Acoust. Soc. Am*, **24**, 175-184.
Peterson, G.E. & Lehiste, I. (1960) Duration of syllable nuclei in English. *J. Acoust. Soc. Am*, **32**, 693-703.
Picheny, M.A., Durlach, N.I. & Braida, L.D. (1986) Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *J. Speech & Hearing Res*, **29**, 434-446.
Repp, B.H. & Liberman, A.M. (1987) Phonetic category boundaries are flexible. *Categorical Perception*, ed by S.N. Harnad. New York: Cambridge University Press.
Samuel, A. (1982) Phonetic prototypes. *Perc. & Psychophys*, **31**, 307-314.
Syrdal, A. & Gopal, H. (1986) A perceptual model of vowel recognition based on the auditory respresentations of American English vowels. *J. Acoust. Soc. Am*, **79**, 1086-1100.