

Static and dynamic approaches to understanding vowel perception.

James M. Hillenbrand

Western Michigan University, Kalamazoo MI 49008

Chapter to appear in G.S. Morrison and P.F. Assmann (Eds), *Vowel Inherent Spectral Change*, Heidelberg: Springer-Verlag.

Introduction

This chapter provides a broad overview of work examining the role of vowel inherent spectral change (VISC) in the recognition of vowel identity. Although seldom explicitly stated, the view that implicitly guided vowel perception research for many years was the idea that nearly all of the information that was needed to specify vowel quality was to be found in a cross section of the vowel spectrum sampled at a reasonably steady portion of the vowel. There is now a considerable body of evidence, most of it based on the study of North American English vowels, showing that VISC plays a secondary but quite important role in the recognition of vowel identity. Evidence comes from a variety of experimental techniques, including: (1) measurement data showing that many nominally monophthongal English vowels show significant spectral change throughout the course of the vowel; (2) statistical pattern recognition studies showing that vowel categories can be separated with greater accuracy, and better agreement is seen with labeling data from human listeners, when the underlying measurements incorporate spectral change; (3) perceptual experiments with “silent center” vowels showing that static vowel targets can be removed or obscured by noise with little or no effect on vowel intelligibility; and (4) perceptual experiments with both naturally spoken and synthetic speech signals showing that vowels with stationary spectral patterns are not especially well identified.

Figure 1 (P&B scatter plot)

The starting point in this discussion will be familiar to most readers. Figure 1 shows F1 and F2 measurements for vowels in /hVd/ syllables recorded by Peterson and Barney (1952; hereafter PB) from 33 men, 28 women, and 15 children. The formants

were sampled at steady-state, i.e., "... a part of the vowel following the influence of the [h] and preceding the influence of the [d] during which a practically steady state is reached ..." (p. 177). The /hVd/ syllables were presented in scrambled order to a large group of listeners with no training in phonetics. Listeners were asked to circle one of ten key words corresponding to the monophthongal vowels /i,ɪ,ɛ,æ,ɑ,ɔ,ʊ,u,ʌ,ɔ̃/. The listening test results were quite simple: signals were identified as the vowel intended by the talker just over 94% of the time. The difficulty, of course, is reconciling the excellent intelligibility of these signals with the extensive crowding and overlap that is seen in the F1-F2 measurements of Figure 1. It is obvious, then, that listeners must be attending to features other than F1 and F2 at steady-state. Many possibilities have been explored over

Figure 2 (P&B amplitude sections)

the years, but the one that is central to the topic of this book is related to the fact that the measurements were made at a single time slice. Figure 2, from PB, shows spectrograms and spectral slices from ten /hVd/ syllables. The arrows show the times at which the acoustic measurements were made. The decisions that were made about locating the steadiest point in the vowel seem reasonable enough, but it can also be seen that several of these vowels show quite a bit of spectral movement (see especially /ɪ/, /ɛ/, /æ/, /ʊ/, and /ʌ/). It turns out that PB, along with many of their contemporaries, were well aware of the limitations of representing vowels with a single time slice. The passage below is very nearly the last thing that PB say in their paper.

It is the present belief that the complex acoustical patterns represented by the words are not adequately represented by a single section, but require a more complex portrayal. The initial and final influences often shown in the bar

movements of the spectrogram are of importance here. The evaluation of these changing bar patterns ... is, of course, a problem of major importance in the study of the fundamental information bearing elements of speech. (p. 184)

Below is a very similar comment from Tiffany (1953):

It has been commonly assumed or implied that the essential physical specification of a vowel phoneme could be accomplished in terms of its acoustic spectrum as measured over a single fundamental period, or over a short interval including at most a few cycles of the fundamental frequency. That is to say, each vowel has been assumed to have a unique energy vs. frequency distribution, with the significant physical variables all accounted for by an essentially cross-sectional analysis of the vowel's harmonic composition. (p. 290)

Tiffany goes on to argue that this single-cross-section view is almost certainly too simplistic. Very similar remarks can be found in Potter and Steinberg (1950) and Stevens and House (1963). Looking back at these comments, which appeared in widely read and influential papers going back to 1950, it is curious that the potential role of spectral change did not receive any steady or focused attention for roughly another 30 years.

Figure 3 (Nearey-Assmann spectral change)

Measurement data

Figure 3, from Nearey and Assmann (1986), shows F1 and F2 values measured from the beginnings and ends of ten Western Canadian vowels spoken in isolation by five men and five women. (The initial formant measurements were taken from the first frame with measurable formants and overall amplitude within 15 dB of peak amplitude, and the final formant measurements were taken from the last frame with measurable formants

and overall amplitude within 15 dB of peak amplitude.) It can be seen that there are a few vowels that do not show much spectral movement, especially /i/, /u/ and /ɒ/. The phonetic diphthongs /e/ and /o/ show exactly the kinds of offglide patterns one would expect. The main thing to notice, though, is that the amount of spectral change for /e/ and /o/ is no greater than it is for the nominally monophthongal vowels /ɪ/, /ɛ/, and /æ/.

Figure 4 (H95 spectral change)

Figure 4 shows similar data for speakers from the Upper Midwest, predominately Southern Michigan (Hillenbrand et al., 1995, hereafter H95). The figure shows formants sampled at 20% and 80% of vowel duration. The data are from /hVd/ syllables spoken by 45 men, but spectral change patterns for women and children are quite similar. There are some differences from the Western Canadian vowels, but there are quite a few features in common: /i/ and /u/ are more-or-less stationary, and there are several nominally monophthongal vowels that show roughly as much spectral movement as /e/ and /o/. The /æ/-/ɛ/ pair is of special interest: /æ/ is raised and fronted in this dialect, placing the two vowels almost on top of one another (see Figure 5). Listeners in this study, however, rarely confused /æ/ and /ɛ/, with intelligibility at ~94-95% for the two vowels. High intelligibility is maintained in spite of the large overlap in part because of duration differences between the two vowels (see below), but differences in spectral change patterns may also play a role. Figure 5 also shows extensive overlap between /e/ and /ɪ/, yet listeners almost never confused the two vowels with one another. But note the highly distinctive spectral change patterns for /e/ and /ɪ/ in Figure 4.

Figure 5 (H95 scatter plot)

Pattern recognition

Evidence from pattern recognition studies shows that vowels can be classified with greater accuracy, and better agreement is seen between recognition-model output and human listener data, when the recognition model takes spectral movement into account. Zahorian and Jagharghi (1993) recorded CVC syllables with nine vowel types, nine initial consonants, and eight final consonants (not fully crossed) from 10 men, 10 women, and 10 children. The signals were analyzed using both formants and a cepstrum-like spectral-shape representation based on the discrete cosine transform (see Chapter 3 for more details on this method). For both formant and spectral-shape representations, classification accuracy was consistently higher when feature trajectories were used as the basis for categorization as compared to static feature vectors. The authors also reported better agreement between model outputs and confusion matrices derived from human listeners when classification was based on dynamic rather than static features.

Table 1 (H95 discriminant analysis)

Hillenbrand et al. (1995) used a quadratic discriminant classifier to categorize 12 vowel types (the ten PB vowels plus /e/ and /o/) in /hVd/ syllables spoken by 45 men, 48 women, and 46 children using various combinations of F0 and the three lowest formant frequencies. The recognition model was trained on spectral features sampled once at steady-state, twice (at 20% and 80% of vowel duration), and three times (at 20%, 50% and 80% of vowel duration). A sampling of the results is shown in Table 1. A substantial improvement was seen in classification accuracy from a single sample to two samples. However, adding a third sample at the center of the vowel produced little additional benefit, which might suggest that vowel identity is associated primarily with information

in the onsets and offsets of vowels (see Chapter 3 for further discussion of these findings).

Hillenbrand, Clark, and Nearey (2001) made recordings from six men and six women producing eight vowel types (/i,ɪ,ɛ,æ,ɑ,ʊ,u,ʌ/) in isolation and in CVC syllables consisting of all combinations of seven initial consonants (/h,b,d,g,p,t,k/) and six final consonants (/b,d,g,p,t,k/). As with an earlier study by Stevens and House (1963), there were many highly reliable effects of phonetic environment on vowel formants. While most of the context effects were small to modest in absolute terms, a few were quite large, especially an upward shift in F2 of ~500 Hz in men and nearly 700 Hz in women for /u/ in the environment of initial alveolars, and an upward shift in F2 of ~200 Hz in men and ~250 Hz in women for /ʊ/ in the environment of initial alveolars. Despite these context effects, vowel intelligibility was quite good in all phonetic environments. For example, the full range of variation in average intelligibility across the 42 consonant environments was only ~6 percentage points (~91-97%), with a standard deviation of just 1.7. The listeners, then, were clearly not bothered much by variation in consonant environment. The question is whether spectral movement would continue to aid in the separation of vowel categories in spite of the acoustic complications introduced by variation in phonetic environment. In tests with a quadratic discriminant classifier, we found consistently better category separability for a variety of different combinations of F0, formants, and vowel duration when using two samples of the formant pattern as compared to a single sample at steady-state. We also found that many aspects of the human listener data could be modeled reasonably well with a simple classifier incorporating F0, duration, and two discrete samples of the formant pattern.

Finally, Hillenbrand and Houde (2003) evaluated a spectral-shape model of vowel recognition that is quite different from the discrete cosine model described by Zahorian and Jagharghi (1993). Briefly, each vowel type is represented as a sequence of smoothed spectral-shape templates derived empirically by averaging narrow band spectra of tokens spoken by different talkers at similar times throughout the course of the vowel. The standard version of the model represents each vowel category as a sequence of five spectral-shape templates sampled at equally spaced intervals between 15% and 75% of vowel duration. Input signals are represented by a sequence of five narrow band spectra at these same time points (15%, 30%, 45%, etc.). A simple city-block distance measure is used to compare the sequence of narrow-band input spectra with template sequences for each vowel category (see Figure 6). The vowel is recognized as the template type that produces the smallest accumulated difference over the sequence of templates. The model was trained and tested on the H95 vowels, using separate template sets for men, women, and children. Of special relevance to the present discussion, recognition performance for the standard five-slice version of the model was compared to models using: (a) a single slice (five separate tests at each of the time points; i.e., 15%, 30%, 45%, etc.), (b) two slices (at 15% and 75% of vowel duration), and (c) three slices (at 15%, 45%, and 75% of vowel duration). Vowel recognition accuracy varied between ~75% and 80% for single-slice versions of the model, with performance being somewhat better for slices taken near the center of the vowel rather than at the margins. Performance improved quite sharply to 90.6% for two slices, but little further improvement was seen for three (91.6%) and five slices (91.4%). Consistent with the H95 pattern recognition results, information about vowel onsets and offsets seems to be the most critical. However, both sets of tests were

carried out with citation-form syllables in the fixed and neutral /hVd/ environment. The situation may not be this simple with connected speech and more complex phonetic environments.

Figure 6 (Narrow band model)

Figure 7 (Jenkins et al. silent center)

Listening studies 1: Static targets are not necessary

The evidence discussed to this point is based on data analysis methods of one sort or another. The obvious question is whether listeners make use of these spectral movements in judging vowel identity. Evidence from several sources indicates that static vowel targets are neither necessary nor sufficient for the recognition of vowel identity. Figure 7, which addresses the first of these points, is from Jenkins, Strange, and Edman (1983), one of several silent-center studies carried out at the University of Minnesota. The original, naturally spoken /bVb/ syllables are shown in the top row. For the silent center signals, 50% has been edited out of the center of short vowels, and the middle 65% has been edited out of long vowels. The last row shows the vowel centers alone, with the onglides and offglides edited out. The main finding is that the silent center signals were identified just as well as the full syllables. The vowel centers were also well identified (error rates for the three conditions shown in Figure 7 are statistically indistinguishable), but the main point is that the acoustic segments associated with presumed vowel targets are not needed. A second experiment showed, among other things, that the initial and final segments alone were not nearly as well identified as the silent center stimuli.

Figure 8 (Nearey-Assmann 1986)

Figure 8 is from a follow-up study by Nearey & Assmann (1986), who created test signals from 30 ms Hamming-windowed segments excised from the beginnings and ends of vowels spoken in isolation. Listeners heard these segments played back-to-back, with 10 ms of silence in between. Subjects were asked to identify the full syllables, the two segments played in their original order, the initial segment repeated, and the two segments played in reverse order. They found that the segments played in natural order were as intelligible as the full syllables; however, error rates were more than twice as high for the repeated nucleus and for the segments played in reverse order.

There is a good deal more to the literature on edited signals of these kinds (see especially Andruski & Nearey, 1992; Jenkins & Strange, 1999; Parker & Diehl, 1984; Strange, 1989; Strange, Jenkins, & Johnson, 1983; Strange, Jenkins, & Miranda, 1994), but the primary conclusions to be drawn from this work seem reasonably straightforward. First, static targets are not necessary, which is fortunate since static targets do not exist for most American English vowels. Brief clips taken from the start and end of a vowel seem to be sufficient for good intelligibility. Second, the Nearey and Assmann findings tell us that it is not just spectral movement that is needed. The spectral change needs to match the patterns that are observed in production data.

Listening studies 2: Static targets are not sufficient

Evidence shows that static targets are not merely unnecessary, they are also insufficient to support the very high levels of intelligibility reported in studies such as PB. One of the more compelling pieces of evidence comes from a study by Fairbanks and Grubb (1961) that received relatively little attention even in its day. The authors recorded nine sustained, static monophthongal vowels (/i,ɪ,ε,æ,ʌ,ɑ,ɔ,ʊ,u/) – vowels which they believed could be “... produced without ambiguity in the steady state” (p. 203). The talkers were seven men, all of them faculty in Speech and Hearing Department at the University of Illinois. The authors went to great lengths to record high quality examples of each vowel. Below is a brief excerpt from a lengthy set of instructions to the talkers that extended for more than a full journal page.

“Essentially what we are trying to do is to collect samples of each vowel that are as nearly typical or representative of that vowel as possible. More specifically, we are interested in samples that depict the central tendency of each vowel ...

Another way of putting the problem is to say what we want you to do is to imagine the target on the basis of your experience in listening to speech, and then demonstrate what the target is by producing a vowel of your own that hits the target as you imagine it. You will understand from this that we are trying to get samples that are something more than merely acceptable and identifiable.” (p. 204)

Two examples of each vowel were recorded by each talker. The experimenter and the talker listened to each vowel immediately after it was recorded. Talkers typically made several attempts before accepting the recording. The full set of recordings was then

auditioned and the speaker was invited to make yet another attempt at any recording that was unsatisfactory for any reason. From these recordings, which were initially ~1-2 s in length, 300 ms segments were excised and presented to listeners for identification. The listeners were eight phonetically trained graduate students from the same department.

With that rather detailed setup, the results were simple: in spite of the elaborate steps that were taken to obtain the highest quality and most representative vowel samples, the average intelligibility was just 74%. Intelligibility varied sharply across vowel category. Intelligibility was the highest for /i/ and /u/ (91-92%), but much lower for /ɪ/, /æ/, /ʌ/, and /ɛ/ (53-66%). It is probably not a coincidence that the most intelligible vowels were those that show the least spectral movement among Upper Midwest speakers, and the least intelligible vowels were those that typically show the most spectral movement (see Figure 4). Further, it is likely that the absence of duration variability also played a role.

It is hard to overstate how striking these findings are. The test signals were spoken by just seven talkers, all of them men, and all phonetically trained. The listeners too were phonetically trained. Of greatest importance, prior to the listening tests all of the signals had been certified by two phonetically trained listeners as not merely identifiable but as representative examples of the vowel category. The listening experiment, then, was not really an identification task in the traditional sense. In effect, listeners were being asked to provide confirmation of labeling judgments that had already been made by two trained listeners. Despite all of this several of the vowel types were misidentified on anywhere from about half to a third of the presentations, and *six of the nine vowel types elicited error rates greater than 25%*. By contrast, PB, using untrained listeners and

untrained talkers, asked subjects to identify ten vowel types spoken by 33 men, 28 women, and 15 children, with formants varying *on average* by a factor of about 1.35 from men to children, and fundamental frequencies varying, again on average, approximately an octave from men to children. Yet the highest error rates reported by PB were 13.0% (/ɑ/) and 12.3% (/ε/). Similarly, in H95, with 12 vowel types spoken by 139 talkers, about evenly divided among men, women, and children, the only vowel category to elicit a double-digit error rate was /ɔ/, at 13.0%. The most obvious explanation for the much poorer identifiability of the Fairbanks and Grubb vowels as compared to the vowels recorded in these two /hVd/ studies is the absence of cues related to spectral change and vowel duration.

Although it may be a bit of a stretch, it might be argued that the elaborate instructions asking talkers to focus on representativeness may have inadvertently worked against Fairbanks and Grubb, resulting in an unnatural set of recordings. This does not appear to be the case since there is evidence from other sources indicating that vowels with stationary formant patterns elicit much higher error rates than those with natural spectral change patterns. For example, Hillenbrand and Gayvert (1993) used a formant synthesizer to generate 300 ms static versions of all 1,520 signals in the PB database using the original measurements of F0 and F1-F3. One set of signals was synthesized with monotone pitch and a second set was synthesized with a sigmoid-shaped falling contour. Seventeen speech and hearing undergraduates with some training in phonetics served as listeners. Identification results are summarized in Table 2. For comparison,

Table 2 (Hillenbrand-Gayvert results)

intelligibility figures reported by PB for the original signals are also shown. It can be seen that the flat-formant synthesized signals are roughly as intelligible as the Fairbanks and Grubb vowels, with intelligibility averaging ~73% for the monotone versions and ~75% for the versions with falling F0 contours. The improvement with falling pitch, shown by all 17 listeners, was highly significant but quite small in absolute terms. The main finding, though, is that the intelligibility of the flat-formant vowels is some 20 percentage points below that of the original signals from which they were synthesized. As with Fairbanks and Grubb, there were very large differences in intelligibility across different vowel types. The most intelligible vowels were /i/, /u/, and /æ/ (~87-96%) and the least intelligible were /ɑ/, /ʊ/, /æ/, and /ɛ/ (~55-66%). Again, vowels that typically show the least spectral change were the most intelligible, and vice versa. Although the limited intelligibility of the static resynthesized PB vowels is almost certainly due at least in part to the lack of spectral change cues, it is very likely that the absence of durational information also played a role. <note 1>

Figure 9 (Hillenbrand-Nearey, 1999, spectrograms)

A follow up study by Hillenbrand & Nearey (1999) addressed the same basic question using a 300-signal subset of the H95 /hVd/ syllables. The 300 signals were selected at random from the full database, but with the following constraints: (1) the 12 vowel types were equally represented, and (2) signals were excluded that showed either an unmeasurable formant (e.g., a merger of F1 and F2 for vowels such as /ɔ/ or /u/, or a merger of F2 and F3 for vowels such as /i/), or with an identification error rate of 15% or greater. Listeners identified three versions of each utterance: (1) the naturally spoken syllable, (2) a formant-synthesized version generated from the original measured formant

contours, and (3) a flat-formant version with the formants fixed at their steady-state value (determined visually from broadband spectrograms). Note that, unlike Fairbanks and Grubb and the PB resynthesis study discussed above, duration information was preserved in the resynthesized signals (within the limits of the 5 ms synthesis frame rate).

Figure 10 (Hillenbrand-Nearey, 1999, results)

Labeling results are shown in Figure 10. The ~7 percentage point drop in intelligibility from the original signals to the original formant signals is important and almost certainly reveals, in part at least, the loss of information relevant to vowel quality that occurs when speech signals are reduced to a formant representation (see Bladon, 1982; Bladon & Lindblom, 1981). However, the finding that is most directly relevant to the present topic is the very large drop in intelligibility of nearly 15 percentage points from the original-formant to the flat-formant signals. As was the case with Fairbanks and Grubb (1961) and Hillenbrand and Gayvert (1993), vowels that tend to show relatively large amounts of spectral change were more strongly affected by formant flattening. The main lesson is that spectral change matters a great deal to listeners, even when duration cues are preserved.

Table 3 (Typical durations)

The role of vowel duration

It is difficult to understand the role of VISC without also considering the influence that vowel duration might have on the perception of vowel identity. As shown in Table 3, English has many pairs of spectrally similar vowels that differ in typical duration. The vowel pairs in the table are listed in order of the size of the difference in inherent duration for stressed vowels based on connected speech data from Crystal &

House (1988) (see generally similar connected speech data from van Santen, 1992, and data from isolated syllables in Black, 1949; Peterson & Lehiste, 1960; and Hillenbrand et al., 1995). The obvious question is whether listeners make use of duration in identifying vowels.

Several experiments have examined the role of duration in vowel identification. For example, Ainsworth (1972) synthesized two-formant vowels with static formant frequencies covering the English vowel space. The vowels were generated in isolation and in /hVd/ syllables and varied in duration from 120 to 600 ms. Results indicated that listeners were influenced by duration in a manner that was generally consistent with observed durational differences between vowels; for example, a vowel in the /u/-/ʊ/ region was more likely to be identified as /u/ if long and /ʊ/ if short, and a vowel in the /ɑ/-/ʌ/ region was likely to be heard as /ɑ/ if long and /ʌ/ if short. Ainsworth also reported that high vowels (e.g., /i/-/ɪ/, /u/-/ʊ/) were less affected by duration than other spectrally similar vowel pairs that differ in inherent duration. Similar results were reported by Tiffany (1953), Bennett (1968), and Stevens (1959).

Other experiments, however, have produced more equivocal findings. For example, Huang presented listeners with nine-step continua contrasting a variety of spectrally similar vowel pairs at durations from 40-235 ms. While the expected duration-dependent boundary shifts occurred, duration differences much larger than those observed in natural speech were needed to move the boundaries. For duration differences that approximate those found in natural speech, boundary shifts were small or nonexistent. Somewhat equivocal duration results were also reported by Strange et al. (1983). Listeners were presented with three kinds of silent center stimuli: (1) durational

information retained (i.e., onglides and offglides separated by an amount of silence equal to the duration of the original vowel nucleus), (2) durational information neutralized by setting the silent intervals for all stimuli equal to the shortest vowel, and (3) durational information neutralized by setting the silent intervals for all stimuli equal to the longest vowel. Results were mixed: shortening the silent interval to match the shortest vowels did not increase error rates relative to the natural duration condition, but lengthening the intervals to match the longest vowels produced a significant increase in error rates. Strange et al. speculated that the results for the lengthened signals may have been "... due to the disruption of the integrity of the syllables, rather than misinformation about vowel length; that is, subjects may not have perceived a single syllable with a silent gap in it, but instead, heard the initial and final portions as two discrete utterances" (Strange, 1989, p. 2140).

Figure 11 (Duration 2 - Synthesis)

Hillenbrand, Clark, & Houde (2000) used a synthesizer to linearly stretch or contract 300 /hVd/ signals drawn from H95 (the same 300 signals used in Hillenbrand & Nearey, 1999). A sinusoidal synthesizer similar in conception to McAuley and Quatieri (1986) was used to generate three versions of each syllable: (1) an original-duration version, (2) a short version with duration set two standard deviations below the grand mean across all vowels, and (3) a long version with duration set two standard deviations above the mean (see Figure 11). <note 2> The original-duration signals were highly intelligible (96.0%), showing that the synthesis method did an adequate job of preserving information relevant to vowel identity. Intelligibility dropped to 91.4% for the short signals and to 90.9% for the long signals. This nearly symmetrical drop in intelligibility

resulting from shortening or lengthening is relatively modest overall, but it turns out that this ~5 percentage-point figure is arrived at by averaging some cases in which duration does not matter at all with other cases in which it matters a good deal.

Table 4 (Effects of shortening on individual vowels)

Table 4 shows the effects of vowel shortening (see the column labeled “Listeners”; the “Model” column will be discussed below). Shown toward the top of the table are the vowels that were most likely to change identity from the original-duration to the short-duration version of the same token: /ɔ/ shifting to /ɑ/ or /ʌ/, /æ/ shifting to /ɛ/, and /ɑ/ shifting to /ʌ/ data. In the bottom of the table are some shifts that might have been expected based on production but hardly ever occurred. For example, of the 350 presentations of the short version of each vowel (14 listeners x 25 tokens of each vowel type), there were just seven cases of short /e/ shifting to either /ɪ/ or /ɛ/, only two cases of short /i/ shifting to /ɪ/, and not a single case of short /u/ shifting to /ʊ/. Results for lengthened vowels are not shown in the table, but these effects are generally in line with the findings in Table 4, with the obvious exception that the arrows are facing in the opposite direction.

Why do listeners show sensitivity to duration for some groups and pairs of vowels that differ systematically in production (/ɔ/-/ɑ/-/ʌ/ and /æ/-/ɛ/) but not others (/e/-/ɪ/-/ɛ/, /i/-/ɪ/, and /u/-/ʊ/)? There is no obvious relationship between the size of the perceptual effect and the size of the duration differences that are observed in production data. The /i/-/ɪ/ and /u/-/ʊ/ pairs in particular show production differences that are quite large, yet listeners were almost entirely unaffected by large changes in the durations of these

vowels. By contrast, observed duration differences in production among the /ɔ/-/ɑ/-/ʌ/ cluster are more modest (particularly /ɔ/-/ɑ/), yet listeners showed a good deal of sensitivity to duration for these vowels.

We believe that there is a reasonably straightforward explanation for these apparently counterintuitive findings. In our view, listeners give little weight to duration for vowel contrasts such as /i/-/ɪ/ and /u/-/ʊ/ that can be distinguished with little ambiguity based entirely on spectral properties. On the other hand, vowel contrasts such as /ɔ/-/ɑ/-/ʌ/ and /æ/-/ɛ/ show a greater degree of overlap in their spectral characteristics, causing listeners to rely on duration to a greater degree in identifying these vowels. In support of this idea are the results of pattern recognition tests showing that findings broadly similar to the perceptual results described above can be modeled with a simple discriminant classifier trained on duration, F0, and formant trajectories. Using measurements from the H95 database, a quadratic classifier was trained on duration, F0, and F1-F3 sampled at 20% and 80% of vowel duration. To simulate listening tests using shortened and lengthened signals, the pattern recognizer was then tested on the same spectral measurements that were used in training the model but with duration set to either: (1) the original measured value, (2) the value used to generate the short signals, or (3) the value used to generate the long signals. Overall, the recognition model showed somewhat more sensitivity to duration than the listeners, but there were several important features in common with the perceptual findings. As was the case for the listeners, the most frequent changes in vowel classification for the shortened vowels were /ɔ/ shifting to /ɑ/ or /ʌ/, /æ/ shifting to /ɛ/, and /ɑ/ shifting to /ʌ/ (see right-most column of Table 4),

and the most frequent changes in vowel classification for the lengthened vowels were the mirror image: /ʌ/ shifting to /ɑ/ or /ɔ/ and /ɛ/ shifting to /æ/ (again, in that order). Finally, the classifier output showed no duration-dependent shifts involving either /i/-/ɪ/ or /u/-/ʊ/, and a relatively small number of shifts involving /ɪ/-/e/-/ɛ/.

In summary, evidence from several listening experiments suggests that vowel duration plays a modest role overall in vowel identification. However, the perceptual influence of vowel duration appears to vary substantially across individual vowel categories. The relative weight that listeners give to vowel duration seems to be influenced by the degree to which a given vowel can be distinguished from its neighbors based on spectral characteristics. It should be noted that all of the evidence that has been discussed here is based on vowels either in isolation or in citation-form syllables. In connected speech there is an exceedingly large, diverse, and often competing set of demands imposed on segment durations in addition to the intrinsic duration of the vowel (see Klatt, 1976, for a review). It is not clear what role duration might play in vowel perception when this feature is controlled not just by vowel identity but also by factors such as speaking rate, emphatic stress, word- and phrase-final lengthening, lexical stress, and the phonetic characteristics of neighboring speech sounds.

Conclusions

In a preliminary analysis of the PB vowels, Potter and Steinberg (1950) provided one of the few explicit descriptions of the static view of vowel perception, noting that the acoustic information specifying vowel identity, "... is expressible in two dimensions, frequency and amplitude. When samples of a given vowel are identified by ear as the same vowel, there must be some frequency-amplitude relationship that enables a correct

identification. The problem is one of recognizing these relationships ... for the different vowels.” (p. 809). However, later in the same paper the authors comment, “It should be noted ... that we are representing a vowel by a single spectrum taken during a particular small time interval in its duration. Actually, a vowel in the word situation that has been used undergoes transitional movements from initial to final consonant. Not only does the spectrum of the vowel change with time, but the ear in identifying the word has the benefit of all of the changes.” (p. 815). The primary conclusion to be drawn from the evidence that has been reviewed here is that the cues to vowel identity for North American English vowels are not, in fact, expressible in a single time slice. The transitional movements referred to by Potter and Steinberg do, in fact, play a critical role in the recognition of vowel identity, as does the duration of the vowel. The evidence shows: (1) all but a few nominally monophthongal vowels show a significant amount of spectral movement throughout the course of the vowel, even when those vowels are spoken in isolation; (2) those spectral change patterns aid in the statistical separation of vowels in both fixed and variable phonetic environments; (3) static vowel targets are not necessary for vowel identification, nor are they sufficient to explain the very high levels of vowel intelligibility reported in studies such as PB and H95; and (4) vowel duration also plays an important secondary role in vowel perception, although the influence of this feature appears to be quite uneven across individual vowels. The long tradition of representing vowels in terms of their static cross-sectional characteristics and, more importantly, of conceptualizing vowels as static points in phonetic/acoustic/perceptual space (rather than trajectories through that space), remains in widespread use. This static

view is a convenient simplification that is useful for some purposes. However, it is a simplification with some important liabilities that are not always properly appreciated. It would be easy to get the impression from this review that nearly all problems related to the role of VISC in vowel perception have been worked out. There remain a number of aspects of this problem that are not well understood. For example, an issue that has received attention only recently has to do with the role of spectral change in characterizing dialect differences. It is well known that the most prominent phonetic variations across English dialects have to do with differences in vowel production. As shown in Figure 12, a common approach to characterizing differences in monophthong production across dialects involves plotting F1 and F2 at steady-state. This figure shows measurements from three dialects of American English: the (predominately) Mid-Atlantic

Figures 12 and 13 (static and dynamic dialect comparisons)

dialect from PB, the Michigan vowels from H95, and unpublished data from /hVd/ syllables spoken by 19 talkers from Memphis, Tennessee. (For simplicity measurements are shown for women only. The patterns for men are quite similar.) Differences in vowel production across the three dialect groups are quite easy to see in this static representation. The description in Figure 12 is not wrong or misleading so much as it is incomplete. Figure 13 shows the spectral change patterns for the Michigan and Memphis talkers (measurements for the PB vowels were made at steady-state only). It can be seen that some vowels show quite similar spectral change patterns and appear to be simply shifted in phonetic space relative to one another (e.g., /i/, /ɪ/, /o/, /ʊ/, and /u/). However, other vowels (e.g., /e/, /ɔ/, /ʌ/, /ʊ/, and /ɑ/) appear to show quite different patterns of

spectral movement as well (see also Fox & McGory, 2007). These kinds of issues will be discussed in much greater detail in Chapters 7 and 8.

Perhaps the most pressing remaining problem has to do with understanding how both talkers and listeners negotiate the competing demands of VISC and coarticulation. A few conclusions seem reasonable based on current evidence. First, although coarticulation produces a large number of reliable effects on the spectral characteristics of vowels (most of them modest in size, but a few quite large – Stevens & House, 1963; Hillenbrand et al., 2001), listeners have little difficulty accurately identifying vowels that are spoken in a wide variety of phonetic environments (Hillenbrand et al., 2001). Second, spectral change patterns of the kinds that can be seen in Figures 3 and 4 cannot be attributed entirely to coarticulation since these patterns are observed in isolated vowels as well as CVCs, and because listeners have been found to rely on these spectral change patterns in identifying vowels in both CVCs and in isolation (e.g., Nearey & Assmann, 1986). Third, while coarticulation effects complicate the relationships between vowel type and VISC patterns, they do not serve to entirely obscure these relationships. Vowel category separability is improved with the incorporation of spectral change information even when there is considerable variation in phonetic environment (Zahorian & Jagharghi, 1993; Hillenbrand et al., 2001). Having said all that, it needs to be noted that a very large share of the evidence on these questions comes from citation-form utterances that are much longer and, in all likelihood, show less pronounced coarticulatory effects than are seen in connected speech. Exploring these effects using the far more rapid speaking rates that are observed in conversational speech would be a useful avenue for further work on this problem.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health (R01-DC01661) and by a grant from the National Center for Research Resources (C06RR17533-01). Thanks to Kelly Woods for her extensive work in recording and analyzing the Memphis vowels, and to Michael Clark, Stephen Tasko, and Geoff Morrison for helpful comments on an earlier draft.

ENDNOTES

1. PB did not measure vowel duration, making it impossible to run the obvious follow-up study using static vowels resynthesized at their measured durations.
2. A neutral-duration condition was also run in which duration was fixed at the grand mean calculated across all utterances. Results from this condition do not add a great deal to the story, so to simplify the presentation these findings will be omitted.

REFERENCES

- Ainsworth, W.A. (1972). Duration as a cue in the recognition of synthetic vowels. *J. Acoust. Soc. Am.*, 51, 648-651.
- Andruski J.E., Nearey T.M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *J. Acoust. Soc. Am.*, 91, 390-410.
- Bennett, D.C. (1968). Spectral form and duration as cues in the recognition of English and German vowels. *Lang. & Speech*, 11, 65-85.
- Black, J. W. (1949). Natural frequency, duration, and intensity of vowels in reading. *J. Speech Hear. Dis.*, 14, 216–221.
- Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson and B. Granstrom (Eds.) *The Representation of Speech in the Peripheral Auditory System*, Amsterdam: Elsevier Biomedical Press, 95-102.
- Bladon, A. and Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *J. Acoust. Soc. Am.*, 69, 1414-1422.
- Crystal, T.H., and House, A.S. (1988). Segmental durations in connected-speech signals: Current results. *J. Acoust. Soc. Am.*, 83, 1553-1573.
- Parker, E. M. and Diehl, R.L. (1984) Identifying vowels in CVC syllables: Effects of inserting silence and noise. *Percept. Psychophys.* 36, 369–380.
- Fairbanks, G., and Grubb, P. (1961). A psychophysical investigation of vowel formants. *J. Speech Hear. Res.* 4, 203-219.
- Fox, R., and McGory, J. (2007). Second language acquisition of a regional dialect of American English by native Japanese speakers. In *Language Experience in*

- Second Language Speech Learning*. O.-S. Bohn and M.J. Munro (Eds).
Amsterdam: John Benjamins Publishing Company.
- Hillenbrand, J.M., Clark, M.J., and Houde, R.A. (2000). Some effects of duration on vowel recognition. *J. Acoust. Soc. Am.*, 108, 3013-3022.
- Hillenbrand, J.M., Clark, M.J., and Nearey, T.M. (2001). Effects of consonant environment on vowel formant patterns. *J. Acoust. Soc. Am.*, 109, 748-763.
- Hillenbrand, J., Cleveland, R., and Erickson, R. (1994). Acoustic correlates of breathy vocal quality. *J. Speech Hear. Res.*, 37, 769-778.
- Hillenbrand, J., and Gayvert, R.T. (1993). Identification of steady-state vowels synthesized from the Peterson-Barney measurements. *J. Acoust. Soc. Am.*, 94, 668-674.
- Hillenbrand, J.M., Getty, L.A., Clark, M.J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.*, 97, 3099-3111.
- Hillenbrand, J.M., and Nearey, T.N. (1999). Identification of resynthesized /hVd/ syllables: Effects of formant contour. *J. Acoust. Soc. Am.*, 105, 3509-3523.
- Jenkins, J. J., and Strange, W. (1999). Perception of dynamic information for vowels in syllable onsets and offsets. *Percept. Psychophys.* 61, 1200–1210.
- Jenkins, J.J., Strange, W., and Edman, T.R. (1983). Identification of vowels in 'vowelless' syllables. *Percept. Psychophys.*, 34, 441-450.
- McAuley, R. J., and Quatieri, T. F. (1986). 'Speech analysis/synthesis based on sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Process, ASSP-22*, 330–338.

- Nearey, T.M. (1978). *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club, Bloomington, IN.
- Nearey, T.M., and Assmann, P. (1986). Modeling the role of vowel inherent spectral change in vowel identification. *J. Acoust. Soc. Am.*, 80, 1297-1308.
- Peterson, G., and Barney, H.L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24, 175-184.
- Peterson, G., and Lehiste, I. (1960). Duration of syllable nuclei in English. . *J. Acoust. Soc. Am.*, 32, 693-703.
- Potter, R.K., and Steinberg, J.C. (1950). Toward the specification of speech. *J. Acoust. Soc. Am.*, 22, 807-820.
- Stevens (1959). The role of duration in vowel identification. *Quarterly Progress Report* 52, Research Laboratory of Electronics, MIT.
- Stevens, K.N., and House, A.S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *J. Speech Hear. Res.* 6, 111-128.
- Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *J. Acoust. Soc. Am.*, 85, 2135-2153.
- Strange, W., Jenkins, J.J., and Johnson, T.L. (1983). Dynamic specification of coarticulated vowels *J. Acoust. Soc. Am.*, 74, 695-705.
- Strange, W., Jenkins, J.J., and Miranda, S. (1994). Vowel identification in mixed-speaker silent-center syllables. *J. Acoust. Soc. Am.*, 95, 1030-1043.
- Tiffany W. (1953). Vowel recognition as a function of duration, frequency modulation and phonetic context. *J. Speech Hear. Dis.*, 18, 289-301.

van Santen, J. P. H. (1992). 'Contextual effects on vowel duration. *Speech Commun.*, 11, 513–546.

Zahorian, S.A., and Jagharghi, A.J. (1991). Speaker normalization of static and dynamic vowel spectral features. *J. Acoust. Soc. Am.*, 90, 67-75.

Zahorian, S.A., and Jagharghi, A.J. (1993). Spectral shape versus formants as acoustic correlates for vowels. *J. Acoust. Soc. Am.* 94, 1966-1982.

Zahorian, S.A., and Zhang, Z.-J. (1992). Perception of vowels synthesized from sinusoids that preserve either formant frequencies or global spectral shape. *J. Acoust. Soc. Am.* 92, 2414-2415 (A).

Table 1. Accuracy in categorizing vowels using a quadratic discriminant classifier trained on one, two, or three samples of various combinations of F0 and/or formant frequencies. From Hillenbrand et al. (1995).

<u>Parameters</u>	<u>1 sample</u>	<u>2 samples</u>	<u>3 samples</u>
F1, F2	76.1	90.3	90.4
F1-F3	84.6	92.7	93.1
F0, F1, F2	82.0	92.5	92.6
F0, F1-F3	87.8	94.1	94.8

Table 2. Percent correct identification of flat-formant resynthesized versions of the Peterson and Barney (1952) vowels with either flat or falling F0 contours. For comparison, identification results are shown for the original stimuli. From Hillenbrand and Gayvert, 1993).

<u>Vowel</u>	Resynthesized		<u>Original Signals</u>
	<u>Flat F0</u>	<u>Falling F0</u>	
/i/	96.2	95.4	99.9
/ɪ/	67.0	76.8	92.9
/ɛ/	65.8	60.9	87.7
/æ/	63.2	64.2	96.5
/ɑ/	55.0	51.0	87.0
/ɔ/	67.2	71.6	92.8
/ʊ/	62.0	72.8	96.5
/u/	89.1	84.6	99.2
/ʌ/	74.7	79.0	92.2
/ə/	86.6	91.7	99.7
Mean:	72.7	74.8	94.4

Table 3. Pairs of adjacent American English vowels differing in typical duration
 Shown in parentheses are average duration ratios for vowels in stressed syllables
 based on connected speech data from Crystal and House (1988). (Speakers were
 from the Mid-Atlantic region.)

/e/	>	/ɪ/	(1.81)
/i/	>	/ɪ/	(1.59)
/æ/	>	/ɛ/	(1.50)
/u/	>	/ʊ/	(1.48)
/ɑ/	>	/ʌ/	(1.36)
/e/	>	/ɛ/	(1.28)
/ɔ/	>	/ɑ/	(1.06)

Table 4. Changes in vowel identity resulting from vowel shortening by human listeners and by a pattern recognition model trained on F0, two samples of the formant pattern, and duration (see text). The percentages reflect the number of shifts in vowel identity (original duration to short duration) divided by the number of opportunities for such shifts to occur. The three rows toward the top of the table show the most common shifts in vowel category involving adjacent vowels different in inherent duration, and the three rows toward the bottom show shifts in vowel category that rarely occurred in spite of systematic differences in intrinsic duration. From Hillenbrand et al. (2000).

	Listeners	Model
<i>/ɔ/ → /ɑ/ or /ʌ/</i>	43.0	54.2
<i>/æ/ → /ɛ/</i>	20.7	25.0
<i>/ɑ/ → /ʌ/</i>	9.4	8.0
<hr/>		
<i>/e/ → /ɪ/ or /ɛ/</i>	2.2	4.0
<i>/i/ → /ɪ/</i>	0.6	0.0
<i>/u/ → /ʊ/</i>	0.0	0.0

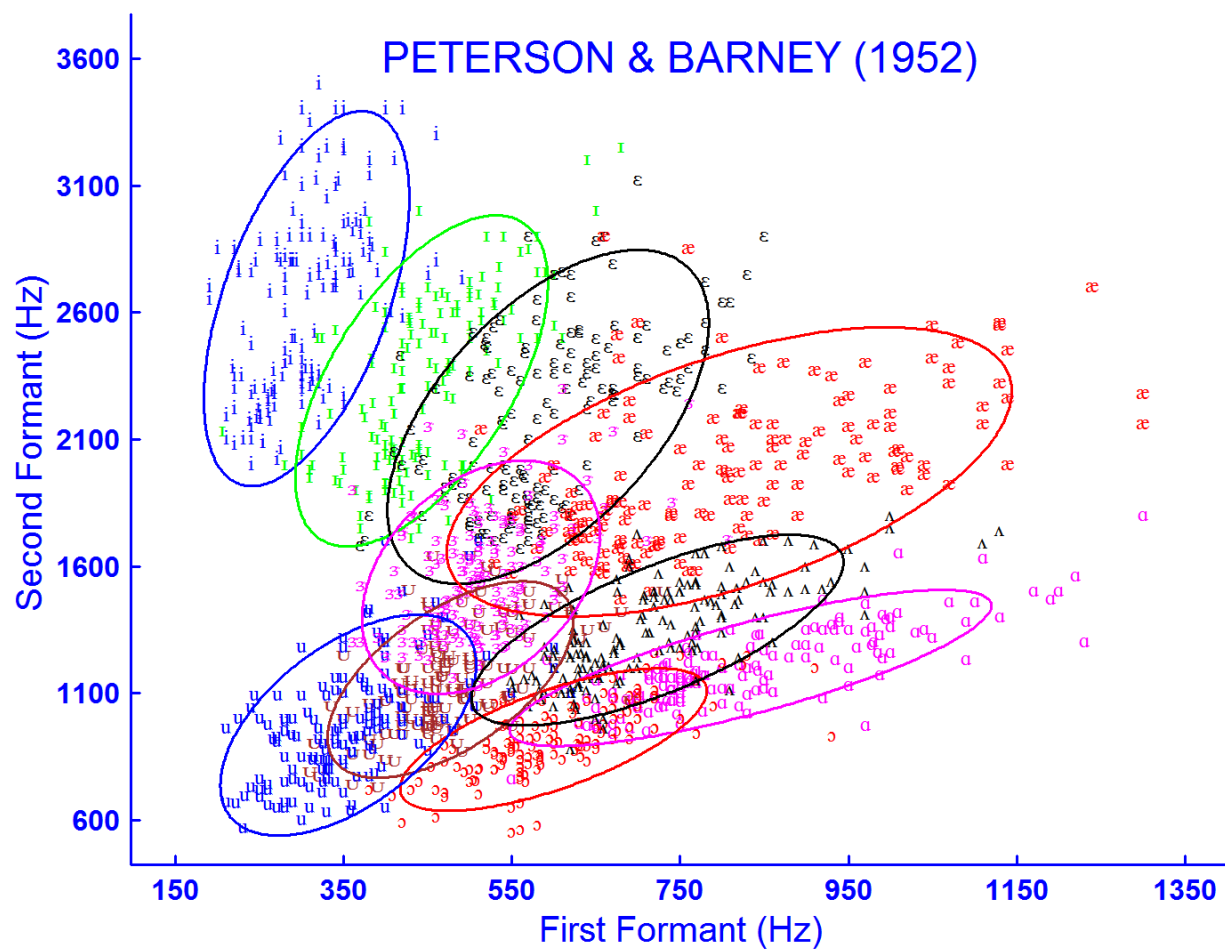


Figure 1. Formant frequency measurements from Peterson and Barney (1952).

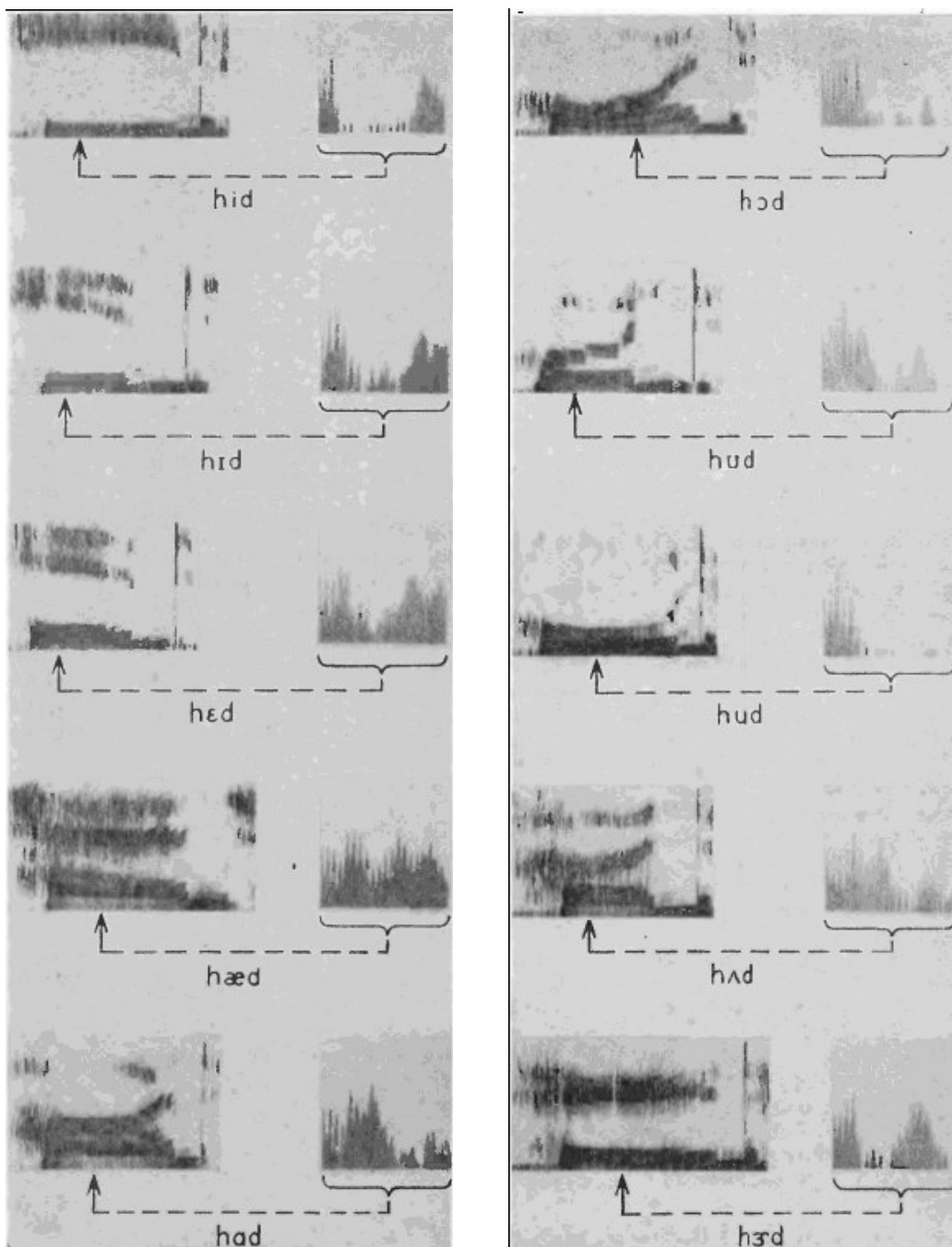


Figure 2. Spectrograms and spectral slices from Peterson and Barney (1952).

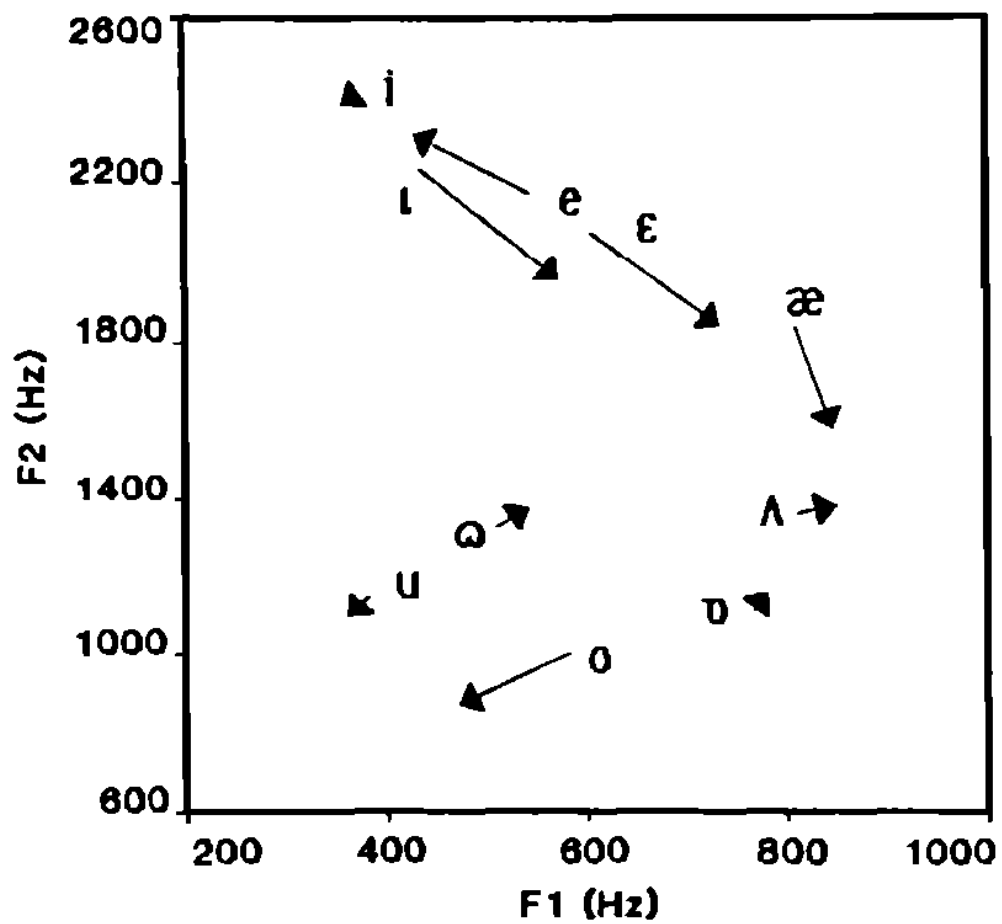


Figure 3. Formant frequencies measured from the beginnings and ends of ten Western Canadian vowels spoken in isolation by five men and five women (Nearey & Assmann, 1986). Arrow heads indicate vowel offsets

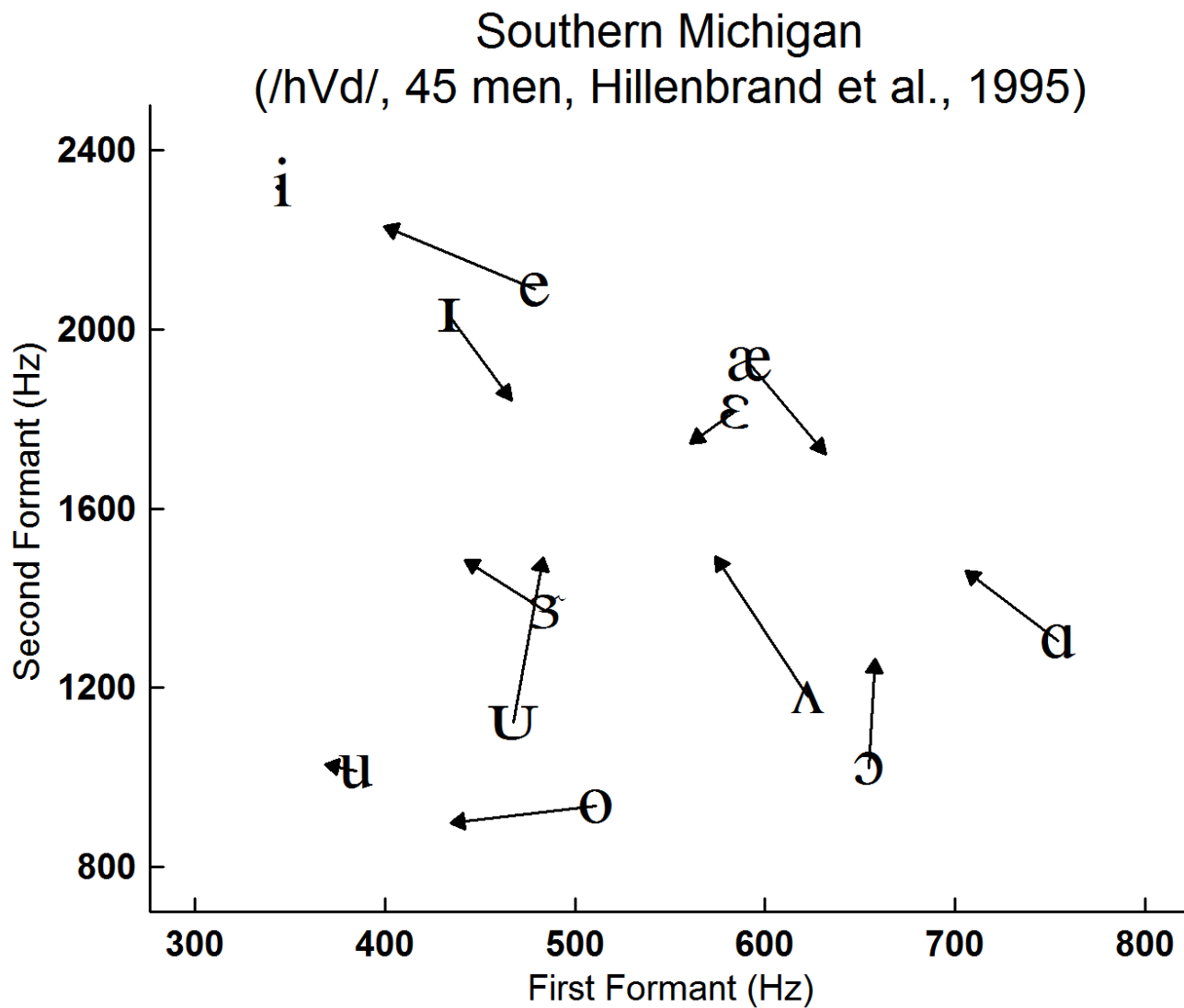


Figure 4. Formant frequency measurements measured at 20% and 80% of vowel duration for vowels in /hVd/ syllables spoken by 45 men. Arrow heads indicate vowel offsets. Speakers are from the Upper Midwest, predominately Southern Michigan. From Hillenbrand et al. (1995).

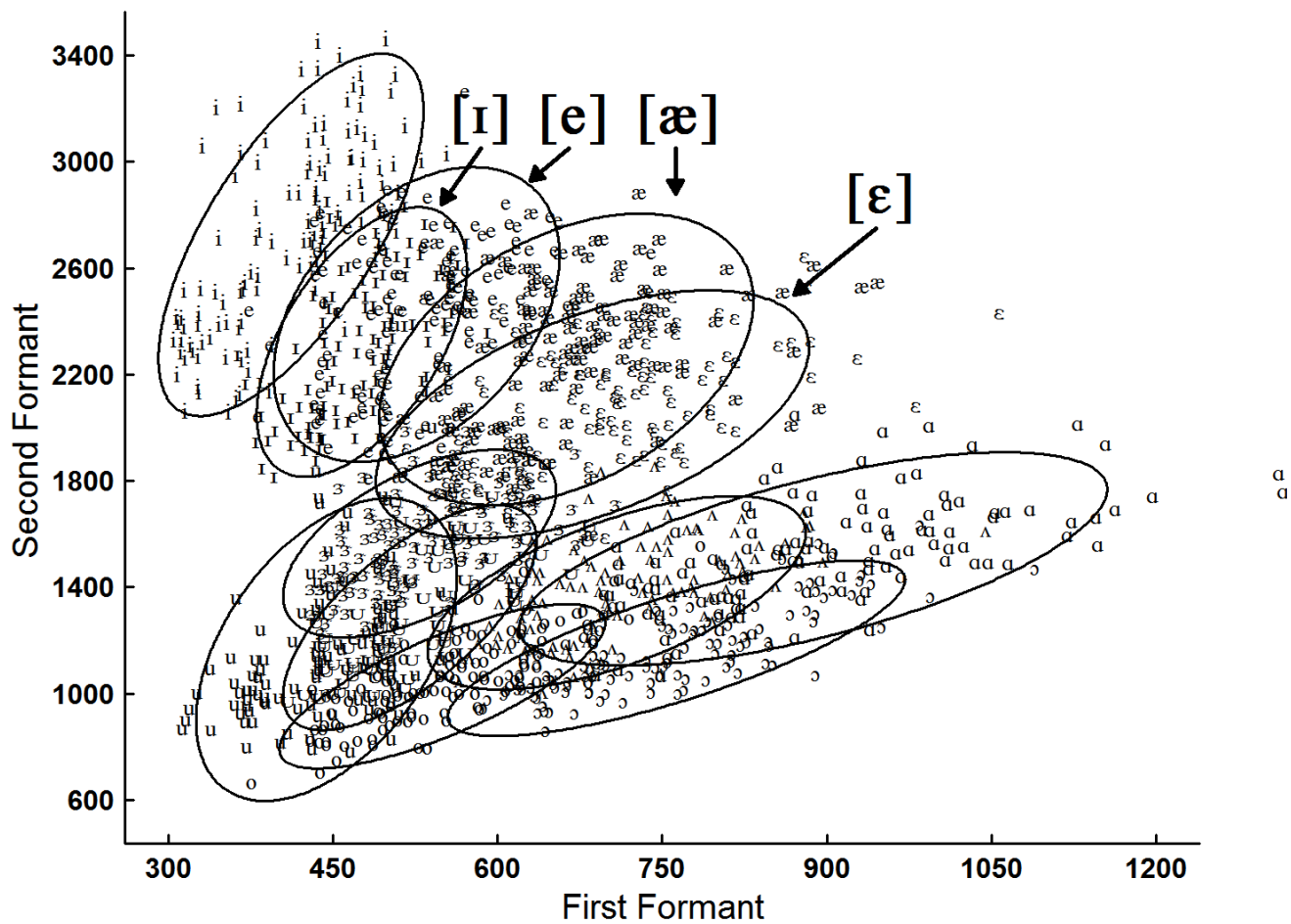


Figure 5. Formant frequencies measured at steady state from Hillenbrand et al. (1995).

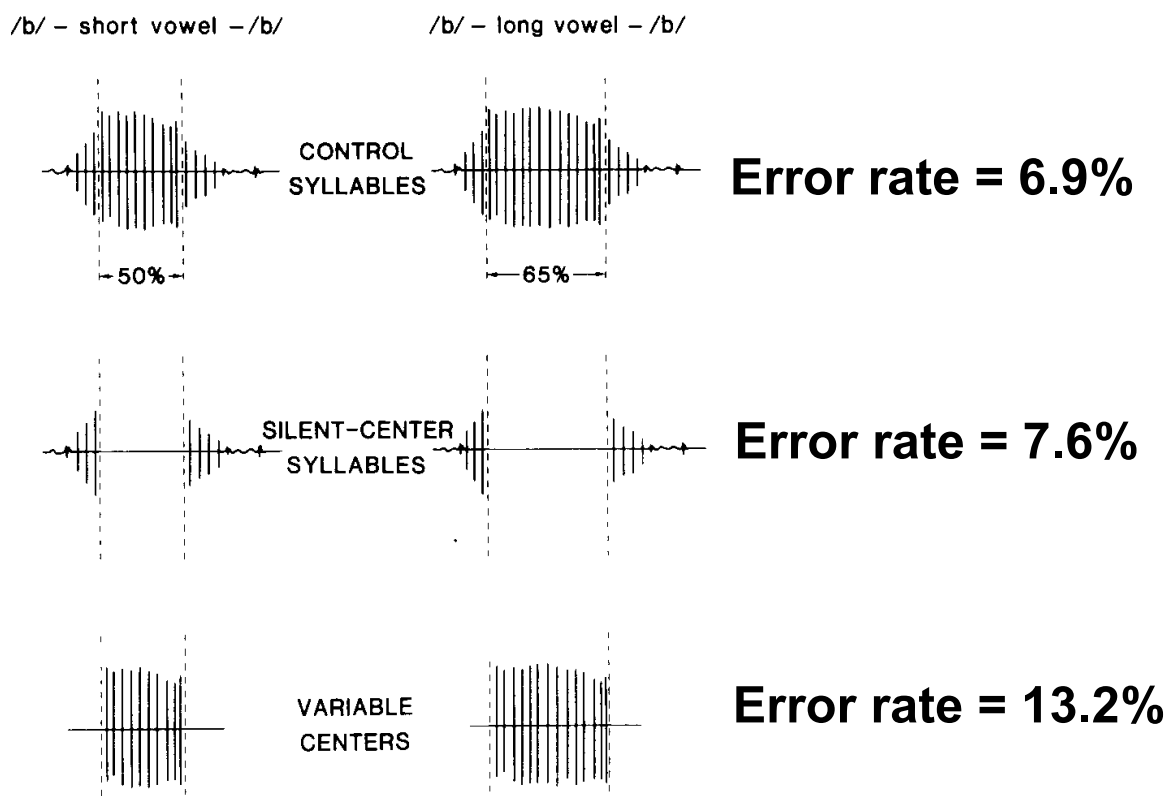


Figure 7. Control, silent-center, and variable-center conditions from Jenkins et. Al. (1983).

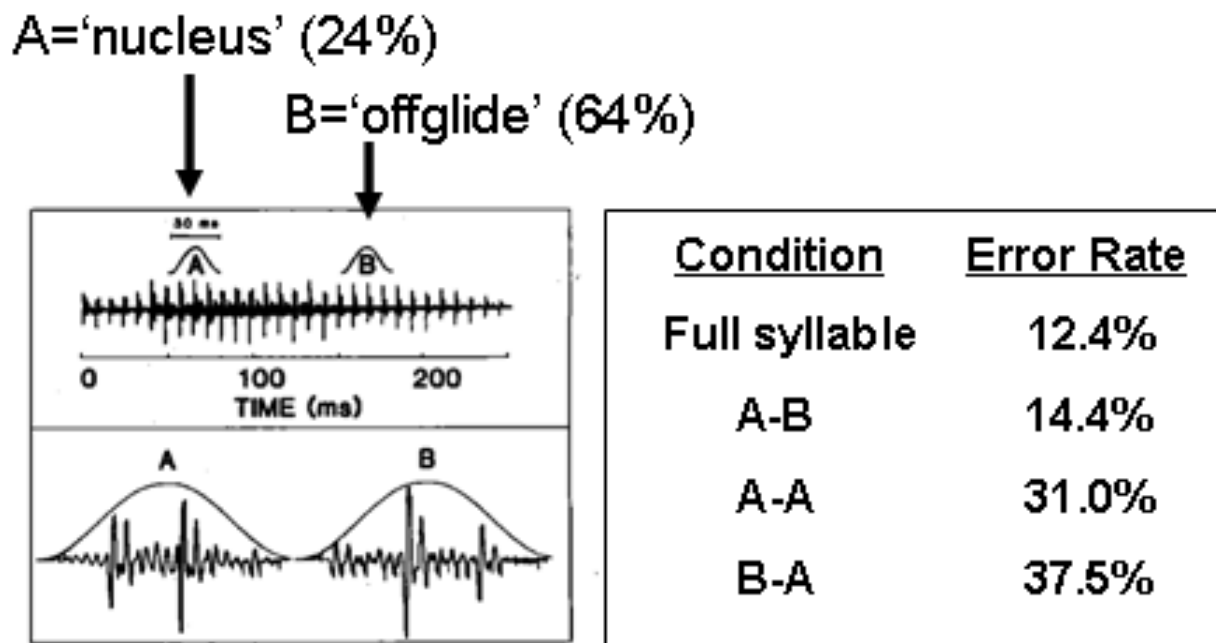


Figure 8. Stimulus conditions and identification error rates from Nearey and Assmann (1986).

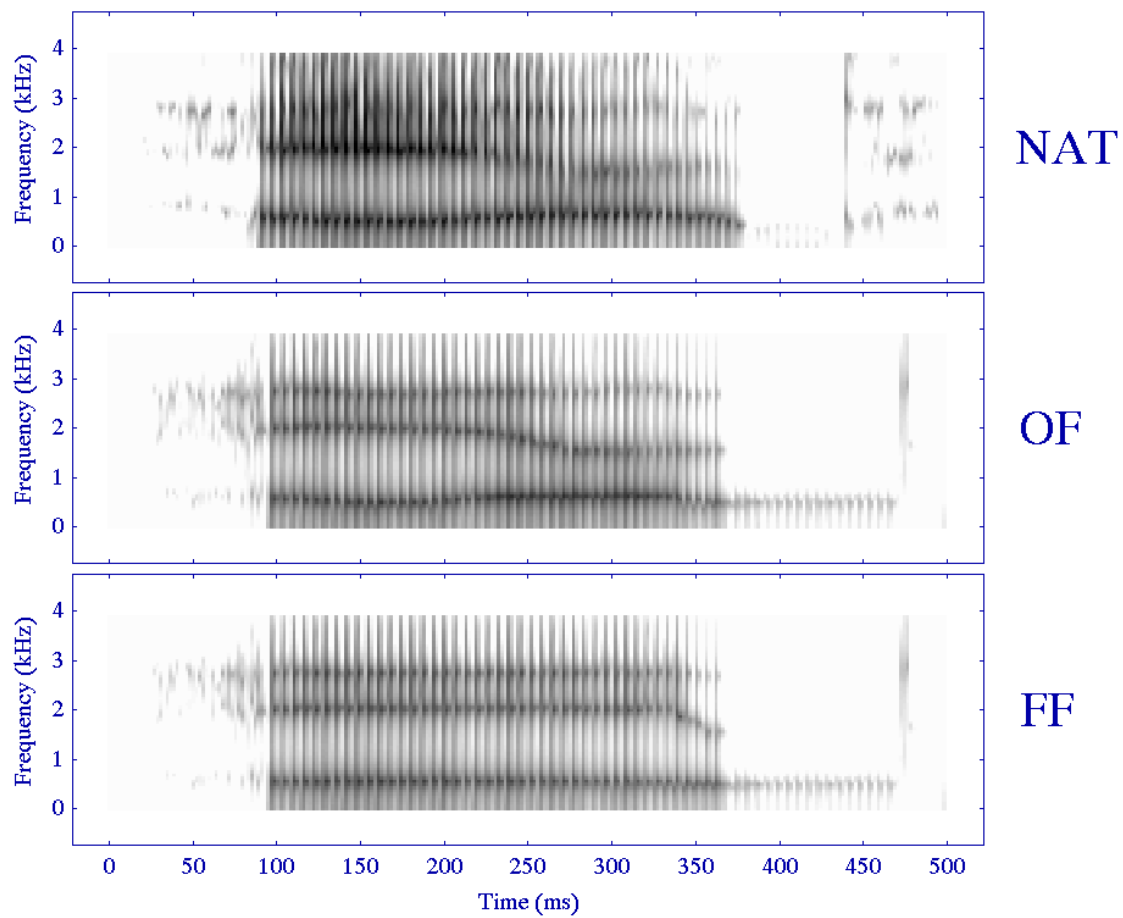


Figure 9. Three stimulus conditions from Hillenbrand and Nearey (1999). NAT= naturally spoken signal, OF = formant synthesized signal generated from the original measured formant contours, FF = formant synthesized signal generated with formants fixed at the value measured at steady-state.

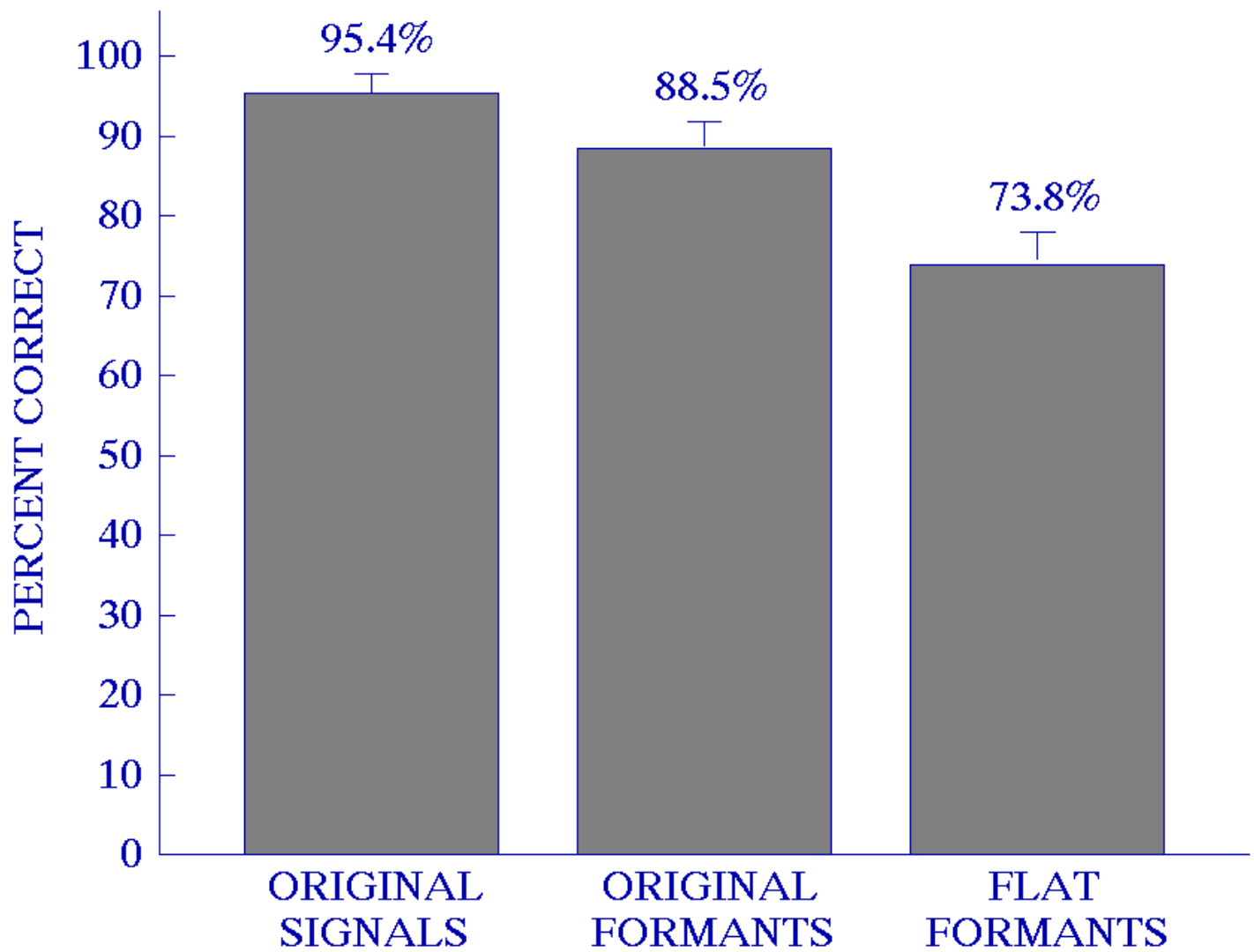


Figure 10. Average intelligibility for the three stimulus conditions from Hillenbrand and Nearey (1999).

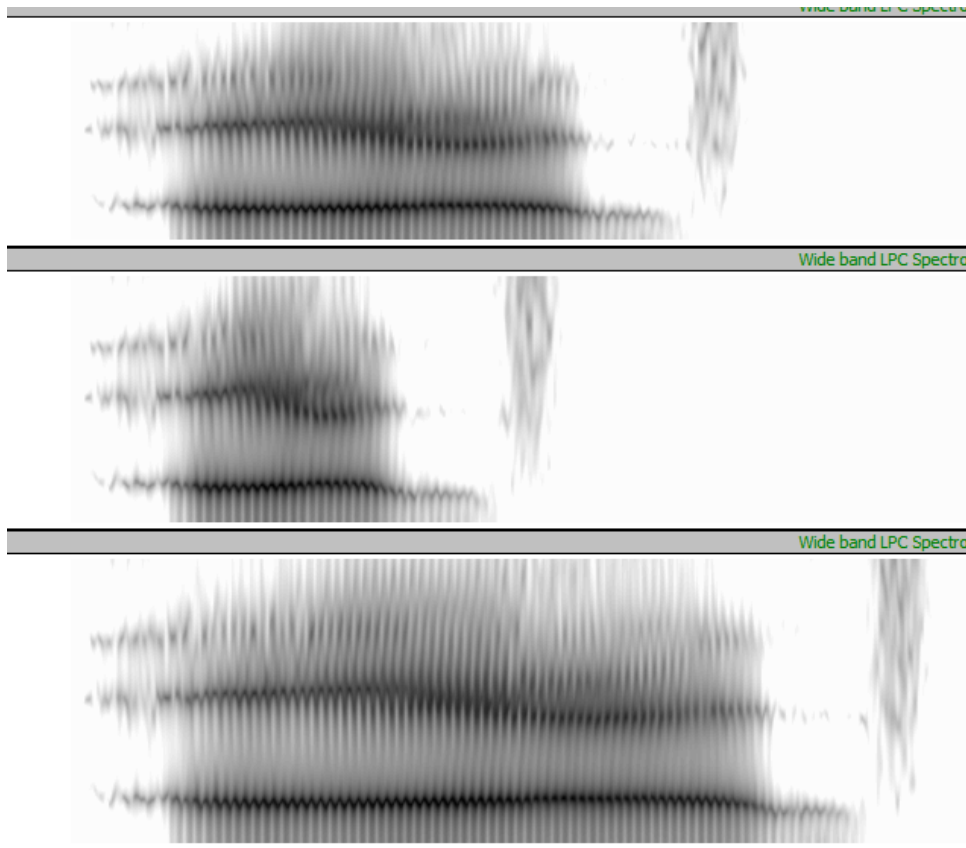


Figure 11. Spectrograms of three synthesized versions of the syllable /hæd/ (from Hillenbrand, Clark, and Houde, 2000).

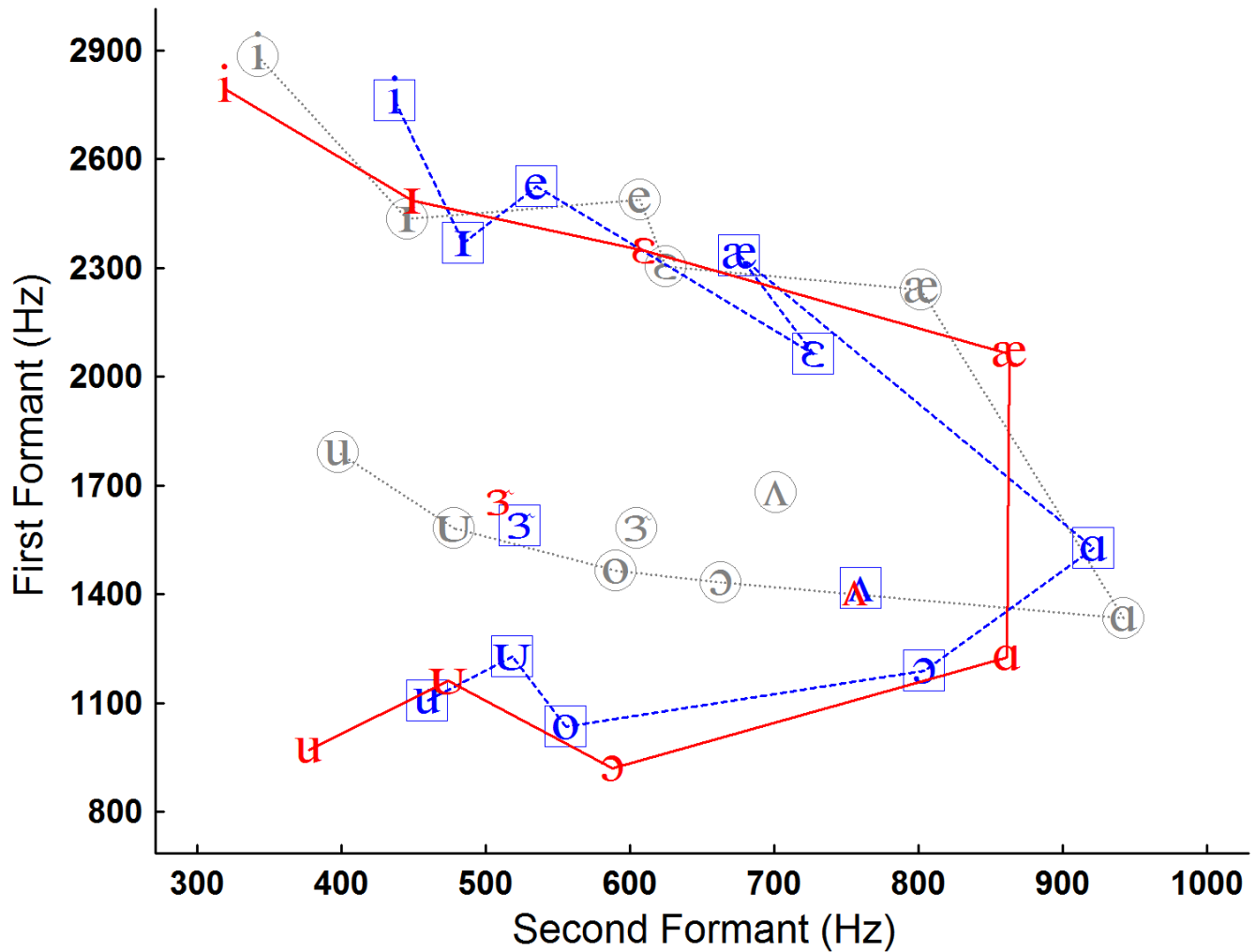


Figure 12. Formant frequencies at steady-state for three dialects of American English: (a) Mid-Atlantic (solid lines), (b) Michigan (dashed lines, phonetic symbols enclosed in squares), and (c) Memphis, Tennessee (dotted lines, phonetic symbols enclosed in circles). For simplicity, measurements are shown for women only.

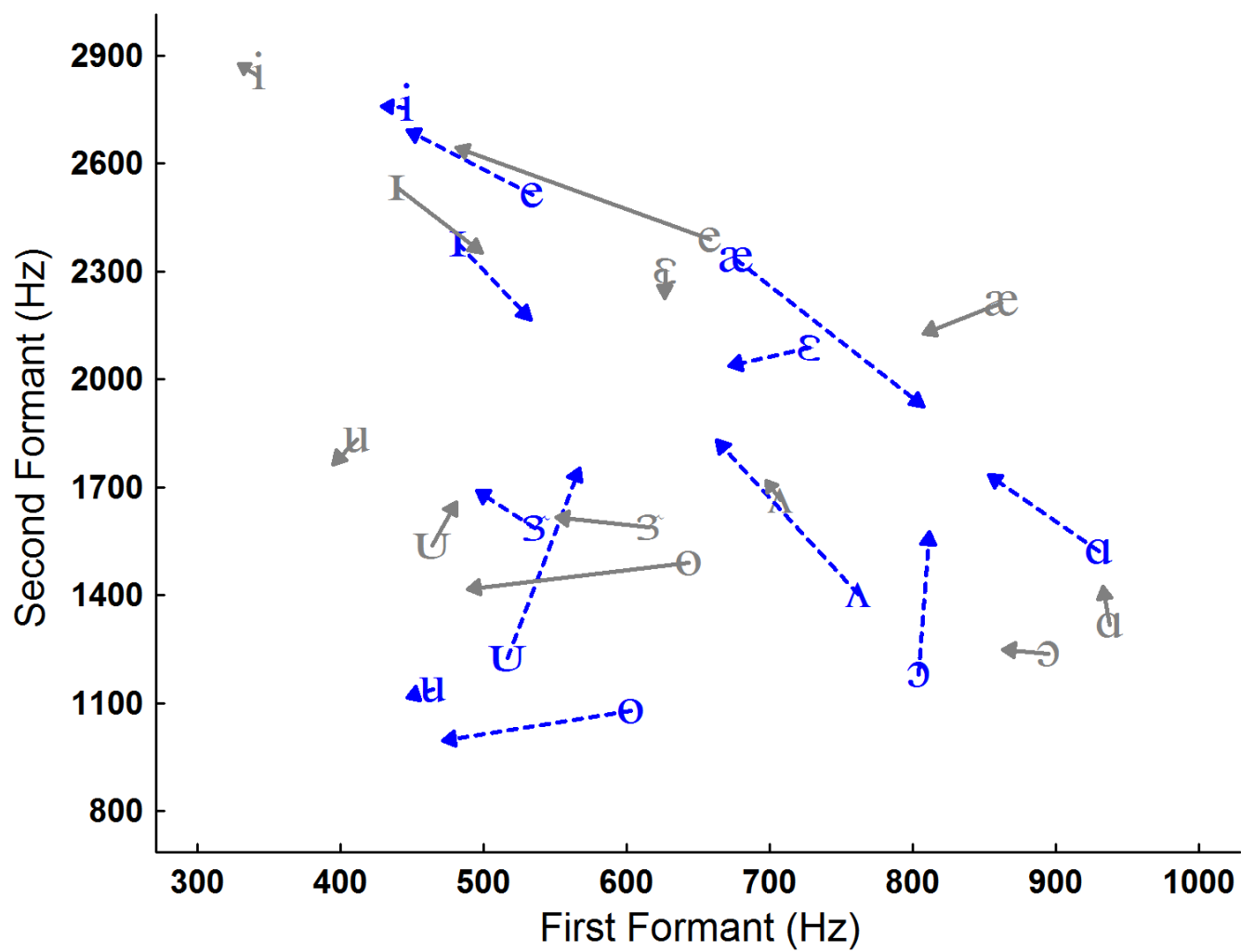


Figure 13. Spectral change patterns for two dialects: Southern Michigan (dashed lines) and Memphis, Tennessee.