Effects of duration and formant movement on vowel perception.

James R. Sawusch

Department of Psychology
State University of New York at Buffalo

ABSTRACT

Acoustical analysis of speech and perceptual studies indicate that the dominant acoustic correlates of vowel perception are the frequencies of the first three formants. However, most vowels are not completely steady-state (even in isolation) and formant frequencies change with variation in the surrounding consonantal context, prosodic influences, speaking rate, and vocal tract length of the talker. In the present studies, both natural and synthetic syllables ("head" and "had") were used to explore the relative potency of average formant frequencies, vocalic duration, and formant frequency movement in vowel perception. A male talker was identified whose formant frequencies, at the midpoint of the words "had" and "head", were identical. However, these tokens differed in their voiced duration and movement of the first three formants and were also highly intelligible. Since the formant frequencies at midpoint could not distinguish these two words, listeners were clearly using different/additional information to guide perception. In the first study, vowel duration was varied. Digital waveform editing was used to generate two series, one based on "had" and the other based on "head". Overall, duration had little effect on listeners' classification of the stimuli. The second study employed synthetic series in which the formant movements of the first three formants were varied between those of the natural "had" and those of the natural "head". Here duration played a much larger role in listeners' responses. Together, these data are a step toward uncovering the relative roles of formant frequencies, formant movement, and duration in vowel perception within fluent syllables.

1. INTRODUCTION

The process of mapping acoustic qualities onto phonetic categories seems to involve both a many-to-one and a one-to-many mapping. There is a large number of jointly sufficient, but individually unnecessary "cues" that influence the phonetic percept. For vowels, the frequencies of the first two or three formants appear to be the dominant acoustic correlates (see Syrdal & Gopal, 1986 for discussion). However, there are other factors that influence perception, including coarticulation (Lindblom, 1963), the direction of formant movement (Nearey & Assmann, 1986), vowel duration (Ainsworth, 1972), fundamental frequency (see Nearey, 1989), and other factors (Sawusch, 1992). This posses two problems for our understanding of the process of speech perception. First, what is the most appropriate characterization of the "acoustic cues" in speech. Second, what is the relative influence of these cues (their weighting) in the perception of speech.

With respect to the first question, Syrdal and Gopal (1986) proposed that the relationships among the formant frequencies, rather than the individual formant frequencies, constitute the primary dimensions for vowel recognition. Recent results of Fahey, Diehl & Traunmüller (1996) seem to support this characterization of the speech signal. The picture on vowel classification is complicated, however, by variations in talker, speaking rate, and the phonetic contexts in which a vowel is produced. Each of these has been shown to alter the acoustic realization of vowels and influence listeners' perception (see Ladefoged & Broadbent, 1957; Lindblom & Studdert-Kennedy, 1967 for examples or Nearey, 1989 for a review). In attempts to understand the nature of vowel perception, individual acoustic correlates are often manipulated while attempting to eliminate or hold constant other factors. For example, to examine the influence of the shape of the spectrum on vowel recognition, Sawusch (1992) used single, steady-state formants. To examine the influence of intrinsic formant movement, Nearey & Assmann (1986) used natural vowels spoken in isolation. Finally, in order to examine the influence of duration, Ainsworth (1972) used isolated, synthetic vowels. Each of these studies demonstrates that a particular acoustic correlate (shape of the spectrum, formant movements, duration) can influence vowel perception.

The primary purpose of the present study was to explore our second question: How are the acoustic correlates combined in perception. In order to answer this question, there are two basic approaches. The first is to start with combinations of cues and map the listeners responses. This is the approach exemplified in various studies of trading relations (see Repp, 1982 for a review). A second approach is to start with natural speech and systematically alter or eliminate acoustic correlates. It is this second approach that will be used here. Our starting point is the words "head" (/hEd/) and "had" (/haed/) spoken by an adult male. Spectrographic measurements showed that the formant frequencies for the two vowels were virtually identical in these two words at the midpoint of the voiced part of each syllable. The fundamental and first three formant frequencies were 152, 607, 1740, and 2340 Hz for /E/ and 132, 608, 1745, and 2271 Hz for /ae/ for one pair of tokens. Consequently, for these tokens by this talker, models which rely exclusively on the formant frequencies (Syrdal & Gopal, 1986) would classify both vowels as the same. Listeners, however, were perfectly consistent in classifying the two words as "head" and "had", as intended by the talker.

Even though the formant frequencies were very similar, the vowels in the two words did differ. The voiced portion of /haed/ was 235 msec in duration while the equivalent portion of /hEd/ was only 157 msec in duration. The patterns of formant movements in the two vowels were not identical, with a greater change in F2 in /ae/ but a greater change in F1 in /E/. The fundamental frequencies of the two tokens were slightly different, with the /ae/ about 20-25 Hz lower than the /E/ over most of its duration. Finally, the change in the shape of the spectrum (the relative amplitudes of the formants) was different for the two vowels. Consequently, this pair of tokens could be distinguished perceptually based on any or all of these acoustic correlates.

2. EXPERIMENT 1

The primary focus of this study is whether listeners will use the duration difference between the natural /hEd/ and /haed/ as a cue to vowel and word identity. If vowel duration is an effective perceptual cue, and if the vowel of /haed/ were shortened, listeners should report hearing "head" and if the vowel of /hEd/ ("head") were lengthened, listeners should report hearing "had". If duration is one of a set of cues, then altering the duration might not be sufficient to alter the word that listeners report. However, some evidence of an effect of duration should be observable, such as changes in ratings of the words by listeners. Finally, if duration is not an effective perceptual cue (in spite of its correlation with the word), then we should see no changes in listeners ratings of tokens that vary in the vowel duration. These three alternatives were tested by editing the natural /haed/ and /hEd/ tokens to create two series that varied in the duration of the voiced portions.

2.1. Method

Listeners. The listeners were 16 undergraduates at the University at Buffalo who participated to fulfill a course requirement. All were native speakers of English with no reported history of a speech or hearing disorder.

Stimuli. The natural /haed/ and /hEd/ tokens were spoken in a carrier sentence, amplified, low-pass filtered at 4.7 kHz, digitized at a 10 kHz sampling rate and stored on computer disk. In the natural /hEd/, there were 25 vocal pulses during the voiced portion of the word (157 msec duration) following the aperiodic /h/ and ending in the closure for the final, unreleased /d/. A series varying in vocalic duration was created from this token by digitally reduplicating pitch pulse length segments of the waveform to make seven new stimuli. With the natural /hEd/ serving as the first series stimulus (the short end), the second stimulus was created by reduplicating the 12th and 13th pitch pulses. Each successive stimulus involved reduplicating two additional pitch pulses. This resulted in a series where the vocalic duration varied from 157 to 241 msec in duration. The second series was created from the natural /haed/ (long end) by deleting non-adjacent pitch pulses, starting in the center of the vowel. This resulted in a seven member series that varied from 142 to 235 msec in duration for the vocalic segment.

Procedure. Each listener participated individually and heard a total of 18 presentations of all 15 stimuli, presented in random order. There was a brief break after the first 32 presentations and after each succeeding 64 trials. On each trial, listeners identified each item as the word "head" or the word "had" using a six point rating scale. A response of 1 indicated a clear example of "head", a 3 indicated a guess of "head", a 4 was a "had" guess, and a 6 was a clear example of "had".

2.2. Results and Discussion

The average rating to each syllable was determined for each listener. The group data for all 16 listeners are shown in Figures 1 and 2. There was little effect of the vocalic duration on listeners' responses. All of the stimuli derived from /hEd/ were labeled as "head" and there was no consistent effect of duration on the listeners' responses. This is easily seen in the flat rating function in Figure 1. There was a small effect of vocalic duration for the stimuli derived from /haed/. Listeners gave reliably lower (but still "had") ratings to the shorter stimuli. This relatively subtle change in the ratings over the /haed/ based series is shown in Figure 2.

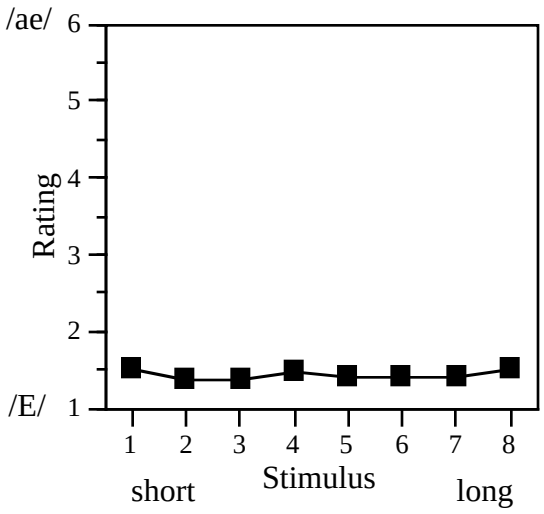

Figure 1. Group ratings of the natural "head" based series. Stimuli varied from 157 (short) to 241 (long) msec in duration.

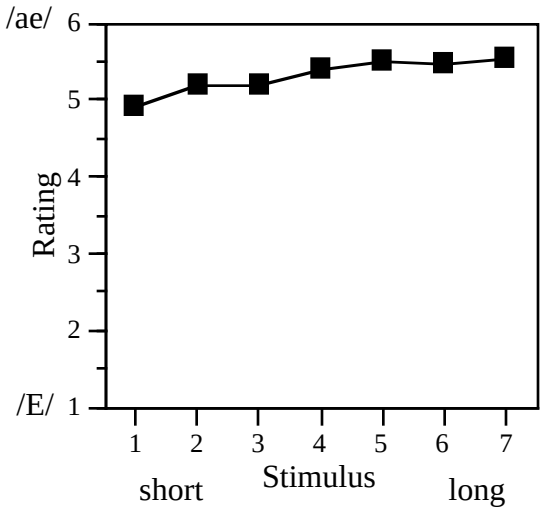

Figure 2. Group ratings of the natural "had" based series. Stimuli varied from 142 (short) to 235 (long) msec in duration.

The data for individual listeners were virtually identical to the group data. None of the 16 listeners consistently gave ratings to a stimulus that differed from the label of the natural token that it was derived from. Consequently, it appears that listeners did not treat the vocalic duration as a potent perceptual cue to word (and vowel) identity in these stimuli.

In considering these data, two further points are worth noting. First, for the /haed/ based series, there is a small influence of

vocalic duration on listeners' ratings. Thus, these results are not incompatible with the earlier results that show listeners can use duration as a cue to vowel identity (Ainsworth, 1972). Rather, the present results show that in a highly natural syllable with multiple acoustic correlates to vowel identity, the vocalic duration may be given relatively little weight as a perceptual correlate of vowel identity.

Second, there are other aspects of the natural tokens that do distinguish them. One is that the pattern and degree of formant movement from the offset of the /h/ through the final /d/ was not the same for the /ae/ and /E/ vowels. Nearey and Assmann, 1986 have previously shown that the direction and extent of formant movement is a potent cue to vowel identity. Second, there was a small, but consistent difference in the fundamental frequency of the two tokens. If the effective perceptual cues to vowels include the F1 - F0 difference (see Fahey et. al., 1996), then this difference was larger for the /ae/ and may have influenced perception. Finally, the two vowels also differed in the shape of their spectra over time.

3. EXPERIMENT 2

In the second experiment, synthetic series were created that were based on the natural tokens. Within each series, the frequencies of the first three formants during the vocalic portion (vowel-consonant) were manipulated to change from those appropriate for /haed/ to those appropriate for /hEd/. Four variants of this series were generated. In two of the series, the vocalic duration was set to reflect the longer, /haed/ token. In the other two series, the vowel duration reflected the shorter, /hEd/ token. For one of the long series and one of the short series, the natural /h/ from /haed/ was spliced onto the beginning of the synthetic VC portion. For the other long and short series, the /h/ from /hEd/ was spliced onto the beginning of the synthetic VCs.

If the dominant cue to vowel identity in these stimuli (and the original tokens) is the pattern of formant movement, then listeners should show a change in labeling within each of these four series. The Stimulus 1 end of each series should be labeled as "had" and the Stimulus 8 end of each series as "head". An effect of vocalic duration would show up as a change in the locus of the category boundary and/or an overall change in the ratings for the long stimuli relative to the short stimuli. Any influence of coarticulatory information from the natural /h/ should similarly show up either as a change in the locus of the category boundary or the overall ratings across series.

Alternatively, if the dominant cue that distinguishes /haed/ and /hEd/ in our natural tokens is either the F1-F0 difference or the shape of the spectrum and its change over the syllable, then we do not expect a simple, clean change in ratings over these series. Across all of the stimuli here, the fundamental frequency was set to a value that was intermediate between that of the natural /haed/ and /hEd/. The shape of the spectrum represents the profile imparted by the Klatt synthesizer and did not vary substantially within each series. In this case, because these other cues have been neutralized, we might see a larger influence of vocalic duration than had been observed in Experiment 1.

3.1. Method

Listeners. The listeners were 10 undergraduates at the University at Buffalo who met the same requirements as in Experiment 1.

Stimuli. The formant frequencies and bandwidths for the first three formants of the vocalic portion (vowel and final consonant) of the natural /haed/ and /hEd/ tokens were extracted using LPC analysis. These values were used with the cascade mode of the Klatt (1980) synthesizer to create base synthetic versions of /aed/ and /Ed/. A new, shorter version of the /aed/ was created by removing non-adjacent synthesis parameter frames to match the duration of the shorter /Ed/. Similarly, synthesis frames were added to the /Ed/ to create a longer VC that matched the duration of the long /aed/. The long and short series were then generated by interpolating the frequencies and bandwidths of the first three formants from those of the /aed/ to those of the /Ed/ in seven equal steps. This resulted in two VC series of different duration. The natural /h/ from the /haed/ token was spliced to the beginning of each stimulus in each series. Similarly, the natural /h/ from /hEd/ was spliced to the beginning of each stimulus. This resulted in four series: long with /h/ from /haed/, long with /h/ from /hEd/, short with /h/ from /haed/ and short with /h/ from /hEd/.

Procedure. Each listener participated individually and heard a total of 14 presentations of all 32 stimuli, presented in random order. There was a brief break after the first 64 presentations and after each succeeding 64 trials. Listeners identified each item using the same 6 point rating scale used in Experiment 1.

3.2. Results and Discussion

The average rating to each syllable was determined for each listener. The group data for all 10 listeners are shown in Figures 3 and 4.

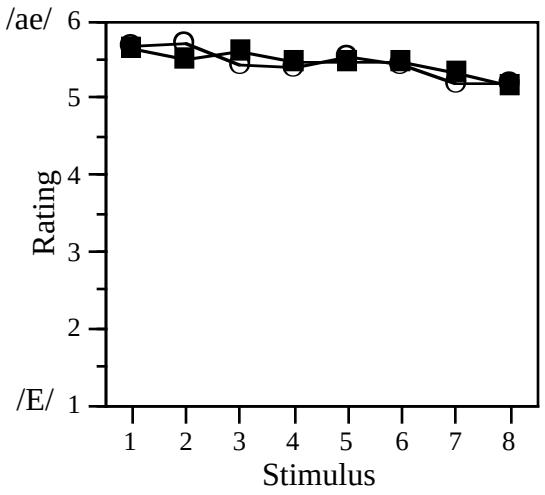

Figure 3. Group ratings of the long duration synthetic series. Stimulus 1 is based on natural "had" formants and Stimulus 8 is based on natural "head" formants. The open circles are stimuli with the "head" /h/ and filled squares are for the "had" /h/.

There was little effect of formant frequencies on listeners' responses. Unlike the first experiment, all of the long stimuli were labeled "had" by listeners, as shown in Figure 3. A small, but consistent effect of the variation in formant frequencies on listeners' ratings was found. Stimuli with formant frequencies based on the natural /hEd/ yielded slightly lower (less extreme "had") ratings, as can be seen for Stimuli 7 and 8 in Figure 3.

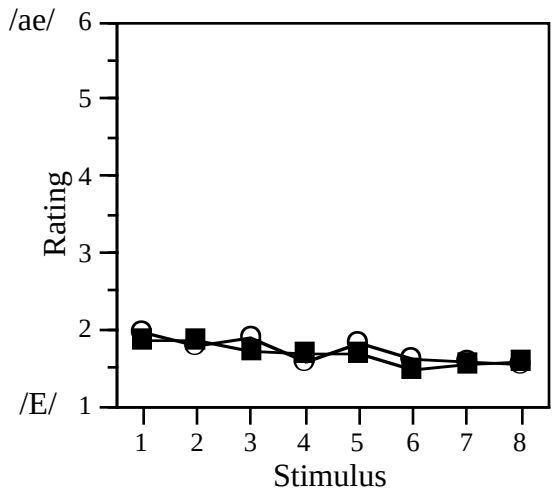

Figure 4. Group ratings of the short duration synthetic series. Stimulus 1 is based on natural "had" formants and Stimulus 8 is based on natural "head" formants. The open circles are stimuli with the "head" /h/ and filled squares are for the "had" /h/.

A substantially similar picture emerged for the short series stimuli. All of these syllables were classified by listeners as "head". A small, but significant, effect or formant frequencies was found with stimuli based on /haed/ yielding slightly higher (less extreme "head") ratings. This can be seen for Stimuli 1 and 2 on the left side of Figure 4.

The data for individual listeners were virtually identical to the group data. Consequently, it appears that listeners did not use the vowel formant frequencies or their movement as the dominant perceptual cue to word (and vowel) identity in these stimuli. Rather, in direct contrast to the results of Experiment 1, listeners appear to have based their judgments predominantly on vocalic duration.

4. GENERAL DISCUSSION

Taken together, the data from Experiment 1 with edited natural tokens and Experiment 2 with synthetic tokens seem to contradict one another. In the first case, little effect of vocalic duration was found on vowel perception while in the second case, vocalic duration dominated perception. This difference, however, is likely to be the result of other differences between the natural and synthetic stimuli. In particular, the natural tokens differed from one another in their fundamental frequency and the shape of their spectrum in addition to their pattern of formant movements and duration. However, the synthetic series contained an intermediate fundamental and the spectral shape largely reflected the Klatt synthesizer, rather than the natural tokens. Consequently, if these other attributes of the natural tokens were incorporated into the synthetic stimuli, we might find a different pattern of results. Experiments designed to test this are currently in progress.

In summary, vowel duration was not a strong perceptual cue to vowel identity for highly natural tokens but was used by listeners when other sources of information were made ambiguous. In fluent, natural English, other acoustic correlates to vowel identity, besides vowel duration, seem to carry a greater perceptual load in determining vowel quality.

5. ACKNOWLEDGMENTS

This research was supported by research grant R01 DC00219 from the National Institute on Deafness and Other Communication Disorders. The author would like to thank Nancy Palmer for her assistance and discussion on this work. Comments may be sent to the author at the Department of Psychology, Park Hall, SUNY at Buffalo, Buffalo, NY, 14260 or via e-mail to jsawusch@acsu.buffalo.edu.

6. REFERENCES

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America, 51,* 648-651.

Fahey, R. P., Diehl, R. L., & Traunmüller, H. Perception of back vowels: Effects of varying F1-F0 Bark distance. *Journal of the Acoustical Society of America, 99,* 2350-2357.

Ladefoged, P. & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98-104.

Lindblom, B. E. F. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35,* 1773-1781.

Lindblom, B. E. F. & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America, 42,* 830-843.

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

Nearey, T. M. & Assmann, P. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustic Society of America, 80,* 1297-1308.

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

Sawusch, J. R. (1992). Auditory metrics for vowel recognition. In M.E.H. Schouten (Ed.), *The auditory processing of speech.* Berlin: Mouton de Gruyter.

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