PERCEPTION OF FIRST AND SECOND FORMANT FREQUENCY TRAJECTORIES IN VOWELS*

Caroline B. Huang

Department of Electrical Engineering and Computer Science and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
USA

ABSTRACT

Previous studies suggest that the first formant trajectory in vowels is perceived differently from the second formant trajectory. F1 may be perceived as a weighted time-average of its time-varying frequency values (Huang, 1985, Di Benedetto, 1987). F2 in high vowels may be perceived with an overshoot (Lindblom and Studdert-Kennedy, 1967). The present study examines F2 in the low vowel region using synthesized utterances. Results from identification tests suggest that F2 in low vowels is perceived with an overshoot of 60 Hz in some contexts. However, results from preliminary experiments in which subjects matched vowels in nonsense words to steady state vowels seem to conflict with the perceptual overshoot theory for F2.

INTRODUCTION

The present study addresses the question: Is the first formant trajectory in a vowel perceived in a different manner from the second formant trajectory? Does a person listening to a vowel with time-varying formant frequencies use one strategy to determine a single value for the vowel's height, which is related to F1, and another strategy to determine the vowel's backness, which is related to F2? Evidence from perceptual tests suggests that the strategies for F1 and F2 perception are indeed different.

There are also theoretical reasons which suggest that F1 and F2 could be perceived differently. F1 and F2 correspond to independent phonological features, high-low and front-back, respectively. The phonological features high-low and front-back (and therefore F1 and F2) have independent articulatory correlates, tongue body height and tongue body backness. Tongue movements in running speech may result in different coarticulation effects for F1 and F2 trajectories. In the vowel spectrum, the spectral prominence corresponding to F1 may be widened or obscured by nasalization, which introduces a pole-zero pair to the spectrum (Stevens et al. [8]) or breathiness, which increases the amplitude of the fundamental harmonic (Bickley [1]). The F2 spectral prominence is not subjected to such effects. The different acoustic characteristics of F1 and F2 could be mirrored in their perception.

The properties of the peripheral auditory system form the basis of an alternative reasoning for the possibility that F1 and F2 are perceived differently. F1 and F2 occupy different frequency bands in the vowel spectrum. The peripheral auditory system processes low frequency and high frequency sounds differently, as shown by the differences in the shapes of the tuning curves for auditory nerves which respond most strongly to low frequency sounds when compared to those for auditory nerves responding most strongly to high frequency sounds. By this reasoning, it may be hypothesized not only that the F1 and F2 trajectories are perceived differently from each other, but that any formant trajectory is perceived differently depending on whether it is high or low in frequency.

PREVIOUS STUDIES

Studies by Huang [3], Di Benedetto [2] and Lindblom and Studdert-Kennedy [6] can be interpreted as evidence for F1 and F2 trajectories being perceived differently. Each study consisted of a series of tests in which subjects were presented with the synthesized vowels in nonsense words and asked to identify the synthesized vowel by making a forced choice between two vowels or two classes of vowels.

In Figure 1, the F1 trajectories for equivalent stimuli in Huang's study are shown. On the basis of identification data from five subjects, each of the stimuli would be called /ɪ/ half of the time and /ɛ/ half of the time. Results for the /ʊ,ʌ/ continuum (not shown) were very similar. The F1 target frequencies of the equivalent stimuli differ by up to about 20 Hz in both vowel continua. The stimulus with the longer onglide and offglide had to attain a higher F1 target value to be perceived to be equivalent to the stimulus with the shorter onglide and offglide. These results are consistent with a theory of perceptual averaging of F1. Subjects seem to perceive an effective F1 frequency which is between the maximum and minimum frequencies attained in the formant trajectory. Unfortunately, in this

*Supported by grants from NINCDS (Nos. NS-04332 and NS-07040)

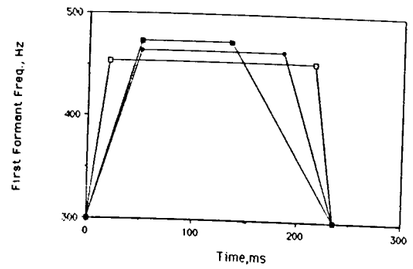

Figure 1: F1 trajectories for three equivalent stimuli in Huang's study.

study F2 was also varied, but only by half the change in F1 frequency on the Bark frequency scale (Schroeder et al. [7]). It may be argued that the change in F1 was perceptually more important.

In Figure 2, two vowel trajectories from Di Benedetto's study are shown. The F1 trajectory shape was different for two types of stimuli. The trajectories for F2 and all higher formants were the same and symmetric for both types of stimuli. Although the two F1 trajectories have the same average (defined as the area under the trajectory vs. time curve divided by the duration of the curve), they are perceived to be different vowels. The trajectory shape with the early steady state caused each of four subjects to identify the vowel as /ɛ/ more than half the time, and the trajectory shape with the later steady state was identified as /ɪ/ or /i/ more than half the time. The tendency was the same for three other subjects who were native speakers of languages other than American English, although the target value of the fifty-percent crossover stimulus was different. These results can be accounted for if a weighted average in which the early portion of the vowel is given more importance than the later portion is hypothesized. The later portion must be given non-zero weight, however, since it was shown in Huang's study that stimuli with trajectories as in Figure 1 with the same onglide duration and target frequency are not equivalent.

Lindblom and Studdert-Kennedy's study suggests that F2 is perceived with an overshoot. For example, for an F2 trajectory which rises to a target and falls again, subjects seem to hear an effective F2 frequency which is higher than the frequency actually attained. Note that if it is not hypothesized that F1 and F2 are perceived differently, Lindblom and Studdert-Kennedy's study would be seen to be in conflict with the studies described above. The vowel formants in their study had parabolic trajectories. The F1 trajectory was the same for all stimuli, while the F2 and F3 trajectories were either concave upward, resulting in a nonsense word of the form /jVj/ or concave downward, resulting in a nonsense word of the form /wVw/. The targets for F2 and F3 were varied while the target for F1 remained fixed for all stimuli, yielding a continuum between the vowels /ʊ/ and /ɪ/. Subjects' identification of the vowels with parabolic formant trajectories were compared to their identification of steady state vowels. The equivalent stimuli shown in Figure 3 are derived from the median fifty-percent crossover points in the identification curves for the steady-state vowels from ten subjects and median fifty-percent crossover shifts for the two contexts relative to the steady state vowels. (An identification curve shows percentage identification of a stimulus as /ɪ/, for example, versus the stimulus' position in the continuum.) The targets of the equivalent steady state and /wVw/ stimuli differ by 185 Hz. The targets of the equivalent steady state and /jVj/ stimuli differ by 75 Hz.

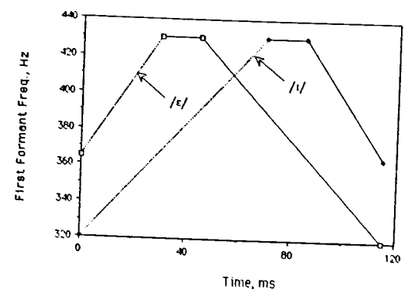

Figure 2: F1 trajectories from Di Benedetto's study. The vowels are perceived as ɛ and ɪ.

There was much inter- and intra-subject variation in Lindblom and Studdert-Kennedy's data which probably arose from subjects' difficulty in hearing the /wVw/ and /jVj/ stimuli as words. Huang [4] did a similar but smaller study using the nonsense words /əwVwə/ and obtained more consistent data which confirm Lindblom and Studdert-Kennedy's results.

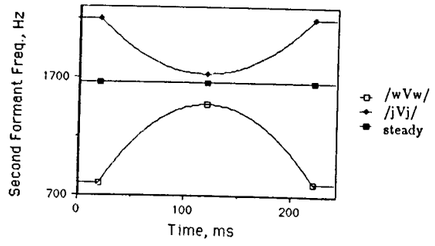

Figure 3: F2 trajectories for equivalent stimuli in Lindblom and Studdert-Kennedy's study.

F2 PERCEPTION IN THE LOW-VOWEL REGION

Lindblom and Studdert-Kennedy's study investigated F2 in the high-vowel region. The present study examines F2 in the low vowel region. Utterances of the form /əwVwə/ and /əjVjə/ were synthesized using the Klatt cascade formant synthesizer [5]. The target for the second formant of the vowel /V/ was varied in 57 Hz steps from 1090 Hz to 1720 Hz, a range of values appropriate for the vowel continuum /æ,ɑ/. The vowel had four formants, and the first, third, and fourth formant targets were 695 Hz, 2425 Hz, and 3500 Hz, respectively, for all stimuli. Two vowel durations were studied, 100 ms and 200 ms. The vowel trajectories in the nonsense words were parabolic and were concave upward for the /j/ context and concave downward for the /w/ context. Steady-state vowels with formant frequencies at the targets of the parabolic trajectories were also synthesized. The utterances were presented to five subjects in forced-choice identification tests in an order which ensured a balanced context. Each stimulus was repeated twelve times. Nonsense words of the same type and duration were presented together.

Fifty-percent crossover points were obtained by hand-fitting smooth curves to the identification curves. The fifty-percent crossover points for each subject and for the averaged identification curves are shown in Table 1. In Figure 4, the F2 trajectories of equivalent stimuli derived from the crossover points from the averaged data are shown. There is a shift in fifty-percent crossover point of 60 Hz

Subjects	/w/, 200ms /w/, 100ms	Context steady, 200ms steady, 100ms	/j/, 200ms /j/, 100ms
nd	5.2	5.5	6.8
	4.5	5.5	6.5
ms	6.8	6.5	8.2
	6.1	5.8	—
aw	5.8	7.0	7.5
	5.2	7.0	8.5
th	6.2	5.1	6.8
	5.5	4.6	8.1
cb	6.1	6.5	7.2
	5.5	6.8	7.2
Average	6.1	6.2	7.2
	5.5	5.8	7.6

Table 1: Results of the present study: 50% crossover points from identification curves. The numbers refer to the scale of stimulus numbers. Stimulus 1 was the most /ɑ/-like; Stimulus 12 was the most /æ/-like. The lower the crossover point, the more stimuli in the vowel continuum were called /æ/. The step-size was 57 Hz in F2. A dash (—) means data was unusable.

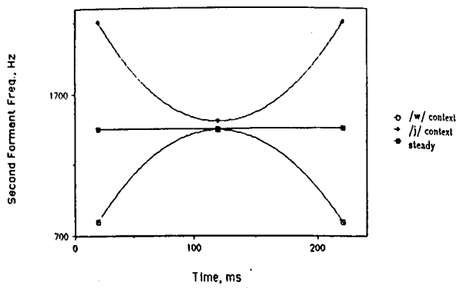

Figure 4: F2 trajectories for equivalent stimuli in the low vowel study.

when comparing the vowels in the /j/ context to the steady state vowels. The shift is in a direction consistent with a hypothesis of perceptual overshoot. On the average, there is no shift in the crossover point when comparing vowels in the /w/ context to the steady state vowels, since individual subjects showed shifts in both directions. There are small shifts in crossover point when comparing the 200 ms vowels to the 100 ms vowels in both the /w/ and /j/ contexts in directions indicating that the perceptual overshoot effect increases for shorter duration stimuli.

PRELIMINARY RESULTS FROM VOWEL MATCHING EXPERIMENTS

The same five subjects were then asked to match the vowels in the nonsense words to steady state vowels. F1, F3, and F4 of the steady state vowels for matching were at the target frequencies of those formants in the vowels in the nonsense words. The F2 of adjacent steady state vowels differed by about 30 Hz, and subjects knew the relative position of each matching stimulus on the steady state vowel continuum. Nonsense word stimuli were chosen in which subjects had unambiguously identified the vowels. The subjects matched the vowels by playing any desired vowels in sequence on a computer as often as they wished. As shown in Figure 5, subjects tended to match a vowel in the /w/ context to a steady state vowel with a lower F2 than actually attained in the parabolic trajectory, suggesting that F2 is averaged. Subjects also tended to match a vowel in the /j/ context to a steady state vowel whose target was lower than actually attained in the parabola, which is consistent with the original hypothesis of perceptual overshoot.

If F2 were perceived with averaging in the /w/ context, the vowel in /əwVwə/ should be equivalent to a steady state vowel whose F2 frequency is below that actually attained in the parabolic trajectory. That is, to be consistent with the trends seen in the preliminary vowel matching experiment, the fifty-percent crossover stimulus on the identification curve should be closest to the most extreme /ɑ/ stimulus for the steady state vowel continuum and closest to the most extreme /æ/ stimulus for the vowels in the /j/ context. The identification data for subjects TH and MS are consistent with the trends seen in the preliminary matching data. Identification data for the other subjects seem to conflict with this trend.

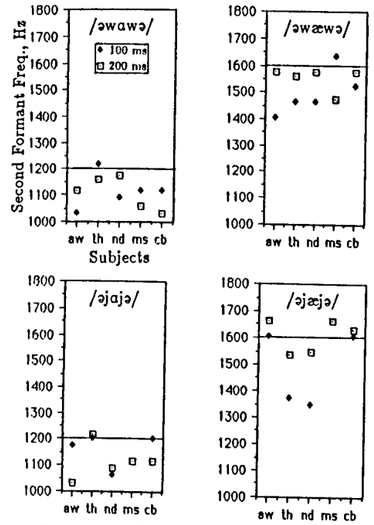

Figure 5: Data from the preliminary matching experiment. Horizontal lines show the F2 target value of the vowel in the nonsense word. Points show the F2 of the steady state vowel matched to the 100 ms and 200 ms parabolic vowels.

DISCUSSION

Apparent conflicts between the identification test results and the vowel matching results must be explained. The two kinds of experiments may be yielding information about different aspects of vowel perception. In this study, the matching experiment only investigated vowels which had been unambiguously identified by the subjects, while identification tests only yielded information about the vowels at the perceptual boundaries. A new vowel matching experiment must be done using the entire continuum of vowels in nonsense word contexts. The tasks of vowel identification and vowel matching are different, and this may also explain the apparent conflict. In vowel identification, a subject labels the vowel, possibly comparing it to an internal idea of how the vowel should sound. This "internal idea" may change depending on the context of the vowel. In vowel matching, a subject compares two "external" vowels and is not required to label. A subject may label the vowels before matching them, however. Subjects may listen to the vowels more analytically in the matching test than in the identification test, especially since they were allowed to play the vowels as often as they wished in this matching experiment. Subjects may use more language knowledge to perform the identification task than the vowel matching task. Individual subjects' strategies may account for the individual differences seen in the data.

Trying to determine whether the effects observed are a result of language learning or of properties of the peripheral auditory system is essential to understanding these effects. A starting point could be to see which of the observed effects can be reproduced using a model incorporating current knowledge of the peripheral auditory system.

Data from identification tests in previous studies and the present study are consistent with the hypothesis that the F1 and F2 trajectories are perceived differently. However, the original hypothesis that F1 is perceived with averaging and F2 with an overshoot does not account for all the effects observed in different types of experiments. Further work needs to be done to understand the relationship between the identification experiments and matching experiments for both F1 and F2 trajectories.

REFERENCES

[1] Bickley, C.A. Acoustic Analysis and Perception of Breathy Vowels. Working Papers, Speech Communication Group 1, Research Laboratory of Electronics, MIT, 1982.

[2] Di Benedetto, M.-G. An Acoustical and Perceptual Study on Vowel Height. Doctoral dissertation, Università degli Studi di Roma 'La Sapienza,' 1987.

[3] Huang, C.B., Perceptual Correlates of the Tense/Lax Distinction in General American English. Master's thesis, MIT, 1985.

[4] Huang, C.B. The Effect of Formant Trajectory and Spectral Shape on the Tense/Lax Distinction in American Vowels, *Proceedings ICASSP 86*, 1986.

[5] Klatt, D. Software for a Cascade/Parallel Formant Synthesizer. *JASA* 67(3):971-995, 1980.

[6] Lindblom B. and M. Studdert-Kennedy. On the Role of Formant Transitions in Vowel Recognition. *JASA* 42(4):830-843, 1967.

[7] Schroeder, M.R., Atal, B.S. and J. L. Hall. Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception. In B. Lindblom and S. Öhman (editors), *Frontiers of Speech Communication Research*, pages 217-229. Academic Press, London, 1979.

[8] Stevens, K.N., Fant, C.G.M. and S. Hawkins. Some Acoustical and Perceptual Correlates of Nasal Vowels. In R. Channon and L. Shockey (editors), *Ilse Lehiste Festschrift*, Foris Publications, Dordrecht, Holland, (in press).