Ann. Bull. RILP, No. 9, 127-136 (1975)

PERCEPTION OF TIME-VARYING RESONANCE FREQUENCIES IN SPEECH AND NON-SPEECH STIMULI *

Hiroya Fujisaki and Sotaro Sekimoto

Time-varying formant frequencies constitute essential cues for the perception of phoneme sequences in connected speech. This paper investigates the conditions under which such formant transitions elicit perception of two contiguous phonemic segments in a syllable-like unit. Based on the analysis of formant transitions in natural speech, synthetic speech stimuli were generated with various values of magnitude, rate, and duration of formant transitions. Discrimination tests of dynamic and static stimuli indicated the existence of perceptual extrapolation of targets that underlie formant transitions. Results of discrimination tests on non-speech stimuli with similar formant transitions suggested that the extrapolation was to a large extent auditory, and thus was not specific to perception of speech stimuli. On the other hand, identification tests of dynamic and static speech stimuli clearly indicated the short-term context effect in perception of connected segments, which was quantified as the amount of temporary shift in the threshold for phonemic judgment due to perception of the immediately preceding segment. Vowels, semivowels and stop consonants were compared with respect to the magnitude of the context effect which they exert on the perception of the following segment.

1. Introduction

It is a well-known fact that the acoustic characteristics of a vowel in connected speech are often quite different from those of the same vowel uttered in isolation. Namely, formant frequencies of a vowel in the dynamic context display trajectories that almost invariably undershoot their target values. This phenomenon is called vowel reduction since in most cases the influence of the context reduces the distinctive characteristics of individual vowels, and often tends to neutralize them (Lindblom, 1963).

For the most part, the phenomenon is ascribable to the inertial movements of all the articulatory organs, constraining realization of the intended articulatory gestures only in an asymptotic manner. Thus, in connected speech where the commands for a sequence of phonemes follow each other, the intended articulatory gestures and the resulting acoustic characteristics of a phoneme are scarcely complete before they are followed by changes for the next phoneme, or they are truncated by the end of an utterance. The range of such inter-phoneme interference is not necessarily limited to the immediately adjacent phonemes, but can extend over several phonemes (Benguerel and Cowan, 1974). In particular, the influence of a vowel is commonly observed to extend beyond the adjacent consonant to the next vowel.

* Paper to be presented at Symposium on Dynamic Aspects of Speech Perception, Eindhoven, August 4 - 6, 1975.

It is certainly not appropriate to ascribe the entire process of coarticulation to the inertial movements of articulatory organs, and there exist instances where an interpretation in terms of reorganization of articulatory commands seems to be more appropriate. It has been shown, however, that the process of coarticulation between successive phonemes in connected vowels can be approximated in the formant frequency domain in terms of a critically-damped second-order linear system, which represents the combined effects of neural, muscular and motional characteristics that intervene between articulatory commands and the observed acoustic characteristics (Fujisaki et al., 1973).

On the other hand, such variabilities of acoustic characteristics of vowels in dynamic contexts seem to be removed by the perceptual process. Namely, the perceptual invariance of a vowel is maintained by a mechanism which removes the coarticulatory effects, or, more properly, utilizes them to infer the underlying targets which are not attained. In other words, the articulatory undershoot seems to be compensated for by the perceptual process. Although the exact characteristics of such a mechanism have not been precisely formulated, there exist experimental results to indicate that the compensation is almost complete if a vowel is presented along with its immediately neighboring vowels (Kuwahara and Sakai, 1972).

Experimental studies of such perceptual processing of time-varying formant frequencies have been conducted by several investigators. Brady et al. (1961) used test stimuli generated by periodically exciting a tuned circuit whose resonance frequency was varied linearly upwards or downwards over a range of 500 Hz with the duration of 20 msec or 50 msec. They asked subjects to match the test stimuli with comparison stimuli having stationary but adjustable resonant frequencies, and found a consistent tendency for subjects to place the resonant frequency of the comparison stimulus near the terminal value of the time-varying resonance of the test stimulus. Lindblom and Studdert-Kennedy (1967) used synthetic speech stimuli with five formants of which the first three were controlled parabolically to simulate vowels in /w-w/ and /j-j/ contexts, and investigated the influence of such contexts on the phoneme boundary on the /u/-/i/ continuum defined by the formant frequency values at the point of closest approach to the vowel target. In most of the subjects, they found a strong context effect that shifted the phoneme boundary toward the locus of the adjacent semivowel, supporting the view that the undershoot in articulation is compensated for by perception. Hiki et al. (1968) analyzed the context effect in the identification of a vowel pair, and tried to formulate a dynamic model of vowel perception in terms of the conditional probability of identification. Arai et al. (1971) used the method of triad comparison to find a stationary vowel that was perceptually equivalent to the dynamic vowel contained in /bVb/, /bV/, or /Vb/ syllables, and formulated the process in a psychological space constructed by multi-dimensional scaling. More recently, Kanamori et al. (1974) used the method of adjustment to find similar correspondences between stationary and dynamic vowels.

Although these previous studies served to clarify the problem and supplied some useful evidence for further investigation, the process of dynamic perception of vowels is still far from being elucidated. Namely, the following questions are still open:

(1) Is the perceptual processing for time-varying resonances the same for both speech and non-speech stimuli?
(2) How does such processing relate to the so-called context effect that is known to exist between a pair of stationary sounds?

2. Problems

As mentioned above, most of the previous studies have investigated perception of the vowel in symmetrical CVC syllables, which involved two-way transitions. Since, however, the minimal context for a vowel is an open syllable where a one-way transition elicits perception of two contiguous phonemes, the investigation of vowel perception in such an environment seems to be essential. Furthermore, the shapes of formant transitions adopted in previous studies were either linear or parabolic and were not good approximations to those observed in real speech. Consequently, the present study concentrates on the perception of synthetic stimuli characterized by one-way formant transitions which are close approximations to those found in $/V_1V_2/$ or /CV/ combinations.

The first problem is to investigate the perceptual processing of such time-varying characteristics. As indicated by some of the previous studies, subjects are reported to be capable of finding the subjective equality between a dynamic stimulus with time-varying resonances and a static stimulus with stationary resonances both in speech sounds and in non-speech sounds. If the same transition of resonance frequency, shared by speech and non-speech stimuli, is perceived as equivalent to the same stationary resonance frequency in a static speech stimulus and a static non-speech stimulus, the result may be regarded as evidence for the existence of a mechanism for compensating the articulatory undershoot at the level of auditory processing rather than phonetic processing. If, on the other hand, the equivalence between dynamic and static stimuli is found to be different for speech and for non-speech, then at least a part of the mechanism should be operating at the phonetic level. A series of discrimination tests have been conducted to elucidate this problem.

The second problem to be investigated is the so-called context effect in dynamic stimuli. The context effect is generally defined as the change in phonemic identification of a speech sound due to its context (Fry et al., 1962). It is, however, not clear whether the apparent shift in the phonemic identification is caused by a change in the auditory or phonetic representation while the criterion for phonemic judgment remains unchanged by the context, or it is caused by a change in the judgment criterion itself. Identification tests on various dynamic and static stimuli have been conducted to find a solution to this problem.

3. Method

It has been shown that close approximation to formant trajectories of natural utterances of $/V_1V_2/$ can be realized by the step response function of a critically-damped second-order linear system (Fujisaki et al., 1973). If the target values of the nth formant frequency of the first and the second vowels in a vowel pair are denoted respectively by F_{n1} and F_{n2}, and the transition to the second vowel is assumed to be initiated at an instant t_2,

the approximated trajectory of the nth formant $F_n(t)$ is given by

$$F_n(t) = F_{n1} + (F_{n2} - F_{n1}) \left[1-\left\{1+\alpha_{n1}(t-t_2)\right\} \exp\left\{-\alpha(t-t_2)\right\}\right] u_{-1}(t-t_2) ,$$

where we assume that the articulatory gestures for the first vowel are initiated well in advance so that its target formant frequencies are realized at the onset of the utterance. This model has been demonstrated to be capable of approximating formant transitions not only in $/V_1V_2/$ combinations, but also in /CV/ combinations where the consonant phoneme is a semivowel or a voiced stop consonant (Fujisaki et al., 1975).

In order to investigate the perception of the first formant transition in $/V_1\ V_2/$, the vowel /u/ was selected as V_1 and the target value for V_2 was selected on the /u/-/a/ continuum, where only the first formant frequency F_1 was varied, keeping all the higher formant frequencies constant. Corresponding non-speech stimuli were generated by using only the first formant transitions, removing all other formant structures. Likewise, in investigating perception of the second formant transition, the target value for V_2 was selected on the /u/-/i/ continuum, where only the second formant frequency was varied.

On the basis of analysis of natural utterances, the duration of the stationary portion was always kept at 80 msec, while the transition was truncated at points where 95, 80, or 70% of the total excursion was attained, producing three different dynamic stimuli corresponding to durations of the transient portion of 118, 76, or 62 msec, respectively. An example of the first formant transition is shown in Fig. 1. Target formant frequencies of /u/, /a/, and /i/ adopted in the present study are given in Table 1, while the fundamental frequency is held constant at 130 Hz. The speech stimuli were generated by computer simulation of a terminal-analog synthesizer, while non-speech stimuli were generated by using only one formant characteristic of the synthesizer. The synthesized signal was normalized by the overall maximum amplitude and was read out at 10 kHz with an accuracy of 10 bits, converted into analog waveform and recorded for off-line experiments. In the listening tests, the stimuli were presented to subjects in an anechoic room through a loudspeaker. The subjects were three male adults with normal hearing.

Table 1. Target formant frequencies of vowels adopted in synthetic speech stimuli.

	F_1	F_2	F_3	F_4	F_5	
/u/	330	1250	2750	3500	4500	Hz
/a/	750	1250	2750	3500	4500	Hz
/i/	330	2250	2750	3500	4500	Hz

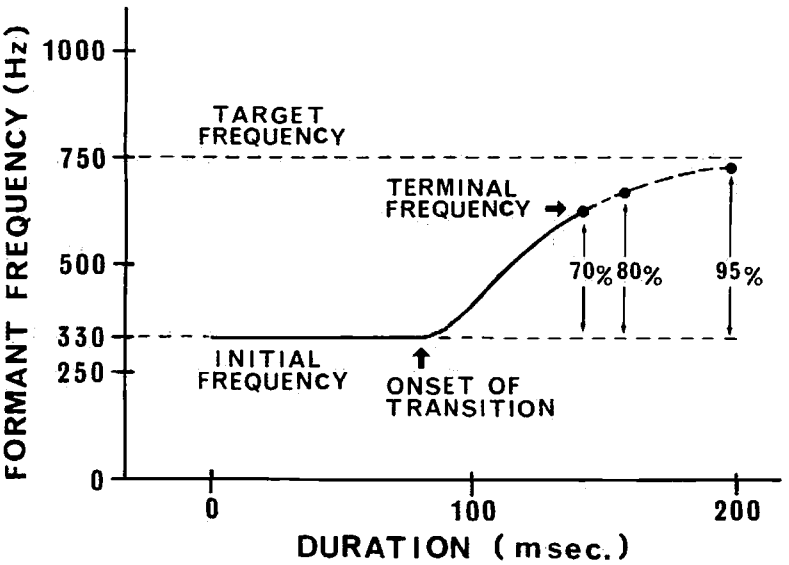

Fig. 1: Formant transition of dynamic speech and non-speech stimuli, approximated by step response of a critically-damped second-order linear system.

4. Perceptual Equivalence of Dynamic and Static Stimuli

Preliminary listening tests of the dynamic stimuli indicated that it might actually be possible to find the point of subjective equality on the continuum of static stimuli generated by replacing the time-varying resonance by a stationary one. In the case of speech stimuli, paired comparison between the second vowel in the dynamic stimulus and a static stimulus can be made to determine which of the two sounds more like /a/. In the case of non-speech stimuli, on the other hand, comparative judgment of dynamic and static stimuli is also possible regarding which of the two sounds higher.

The experiments to find the perceptual equivalence between dynamic and and static stimuli were based on the A-X procedure, in which a dynamic stimulus (A) is followed by one of seven static stimuli (X) whose first formant frequencies are selected in equal steps within the range from 80 Hz below to 40 Hz above the target of the formant transition of the dynamic stimulus. The duration of the static stimuli was 160 msec, and the interval between A and X was 1 sec, while a response interval of 4 sec was inserted between diads. A brief tone marked every 10 diads. The discrimination test was performed for each of the three dynamic speech stimuli and three dynamic non-speech stimuli. Each test consisted of a randomized sequence of 70 diads preceded and followed by 5 dummies, and each subject took at least four tests on each of the dynamic stimuli.

Because of individual differences in subjects' performance, the test results were analyzed individually. In the case of speech stimuli, the percentage of times that a static stimulus is judged to be more like /a/ than the

dynamic stimulus, was obtained for each of the static stimuli, and was approximated by a cumulative Normal distribution, whose mean and standard deviation were calculated by the method of least-mean-squared error with Müller-Urban weighting. The mean corresponds to the point of subjective equality of the dynamic stimulus on the continuum of static stimuli, while the standard deviation serves as an index for the accuracy of discrimination. Results for non-speech stimuli were also analyzed by the same method.

Results of these discrimination tests are summarized in Fig. 2, where the abscissa shows the terminal frequency of the truncated formant transition of dynamic stimuli and the ordinate shows the formant frequency of static stimuli that are judged to be subjectively equal to dynamic stimuli. The solid and broken lines respectively indicate the averaged results of three subjects for speech and non-speech stimuli, and the vertical line indicates the standard deviation due to individual differences. These results indicate that formant transitions, when truncated at 70% or 80% of the total excursion, are perceptually extrapolated and are equated by a stationary formant frequency that is close, if not exactly equal, to the target of the transition. Furthermore, no significant differences are found between the results for speech and non-speech stimuli, suggesting that the processing for such extrapolation occurs at the auditory level. Results obtained for the the first formant transition from /u/ to the phoneme boundary between /u/ and /a/, and for the second formant transition from /u/ to /i/, also indicate the same tendency as found in the /u/-/a/ transition.

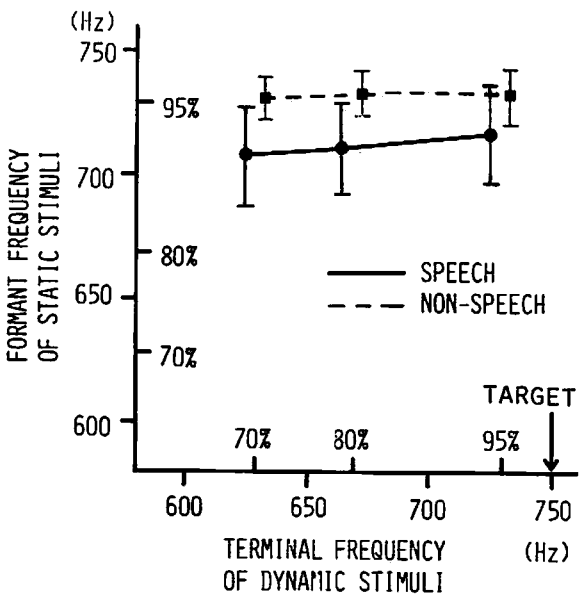

Fig. 2: Subjective equality between dynamic and static stimuli.

5. Identification of Dynamic Speech Stimuli

While the context for a vowel in connected speech inevitably involves formant transitions, the effects of such a dynamic context on vowel identification have scarcely been studied systematically. The following experiments were conducted to obtain quantitative estimates of the effects of magnitude and rate of transition as well as of duration of the preceding segment upon the phonemic identification of a vowel.

The stimuli used in the identification test possessed formant transitions similar to the one shown in Fig. 1, but their target values were selected at points in 15 equal steps on the entire /u/-/a/ continuum.

The first experiment examined the effect of the magnitude and duration of the transition on the identification of a dynamic stimulus. Three sets of stimuli, corresponding to formant trnasitions terminated at 95, 80, and 70% of the total excursion, were generated and used for identification tests. These stimuli could be identified either as /uu/ or as /ua/, and subjects were asked to write down only the second vowel. The stimuli were randomized and presented every 4 sec, and a brief tone was inserted at every 10 stimuli. The test was repeated until each stimulus was identified at least 20 times by one subject, and the results were analyzed individually.

The percentage of times that the second vowel was identified as /a/, was calculated for each stimulus, and was approximated by a cumulative Normal distribution, whose mean and standard deviation were calculated by the same method used in processing the results of the discrimination tests. The mean and the standard deviation are considered as the phoneme boundary and an index for the accuracy of identification, respectively.

Phoneme boundaries thus obtained for dynamic stimuli are compared with those for static stimuli in Fig. 3. The abscissa indicates stimulus conditions, and the ordinate shows the phoneme boundary represented by the target frequency of formant transition. For the three conditions of formant transition, the phoneme boundaries are found to remain almost constant,

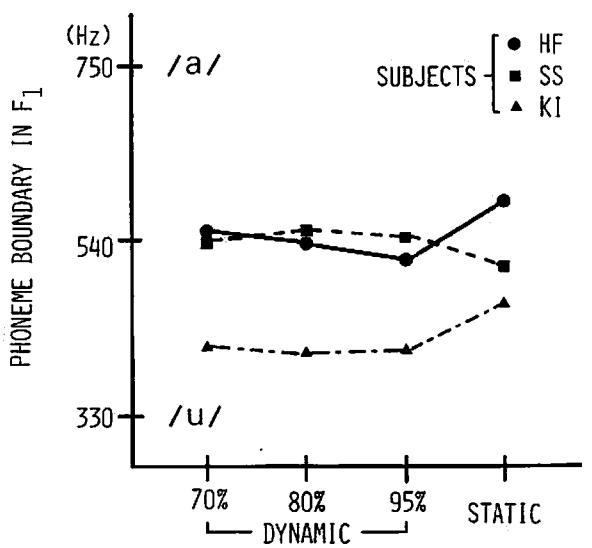

Fig. 3: Phoneme boundary between /u/ and /a/ in target value of F_1 for dynamic and static vowels.

suggesting that perceptual extrapolation of incomplete formant transition in each case produces almost the same auditory representation for the second vowel. Furthermore, comparison with phoneme boundaries for static vowel stimuli indicates that the effect of context is contrast in two subjects (HF and KI), while it is assimilation in one subject (SS). The results of all subjects, however, clearly indicate the short-term shift in categorical judgment for the second vowel which cannot be accounted for by the perceptual compensation found in discrimination tests.

The second experiment examined effects of various changes in the dynamic context on the identification of a vowel. The formant transition was terminated at the point of 95% of the total excursion, while the duration of the stationary initial segment was selected at 120, 80, 40, and 0 msec, resulting in four sets of stimuli for identification. In the first three conditions, the context is identified as /u-/, while it changes into /w-/ in the fourth condition. Furthermore, a fifth set of stimuli was generated by increasing the rate of formant transition by a factor of 5 while removing the initial segment, to produce a /b-/ context.

The results are shown in Fig. 4, again with those for the isolated static vowel. Since little variation was found among phoneme boundaries for the first four sets of stimuli, only results for the two extreme conditions, corresponding respectively to /u-/ and /w-/ contexts, are shown. The fact that such a large change in context duration does not appreciably affect the phoneme boundary suggests that the origin of the context effect observed in these cases may not be auditory, but may rather be phonemic. It is also interesting to note that the semivowel /w/ and the vowel /u/ are almost equivalent in their roles as context in the identification of a dynamic vowel, while a marked difference is found in the contextual role of the voiced stop consonant /b/ in at least two of the subjects.

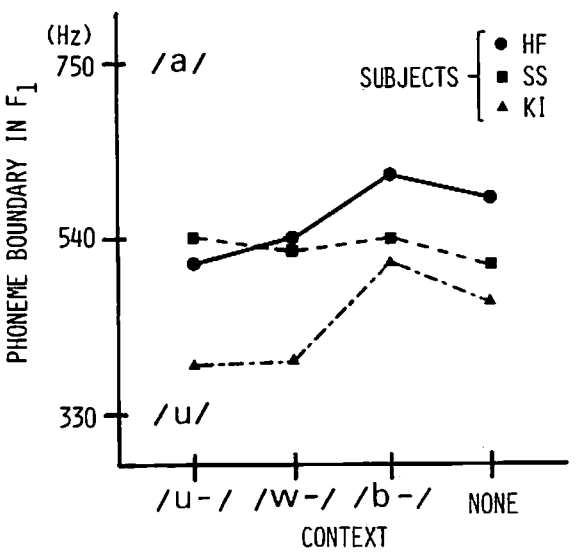

Fig. 4: Influence of various contexts on phoneme boundary between /u/ and /a/.

6. Summary and Discussion

The results of the present study have demonstrated that at least two distinct processes are involved in the perception of a vowel in a dynamic context. Namely, the extrapolation of incomplete formant transitions at the auditory level, and the short-term adaptive change of category boundaries at at the level of phonemic decision.

Our findings on auditory perception of moving formants are generally in line with those of Brady et al. (1961) on non-speech stimuli, but our results made clear that subjects do not necessarily perceive the terminal frequency of a truncated transition, but are capable of approximately extrapolating the target frequency, not only in non-speech stimuli, but also in speech stimuli. The discrepancy may be partly due to the fact that in Brady's experiments, the transition of resonance frequency was not exactly truncated, but was controlled by a trapezoidal function, so that damped oscillation of the terminal frequency might have been continued at the end of the stimuli.

Our results on the effects of dynamic context in vowel identification are also consistent with those of a previous study by Lindblom and Studdert-Kennedy (1967) in that the effect of /w–/ context is contrast in most subjects. The two studies, however, differ with respect to the effect of an increased rate of formant transition; a twofold increase in transition rate did not change the phonemic context, but resulted in an increase in contrast effect in their study, whereas in the present study a fivefold increase in transition rate was found to change the phonemic context from /w-/ to /b-/, and resulted in a decrease of the contrast effect. Whether the discrepancy is due to the change in the phonemic context, or due to the lack of a monotonic relationship between the rate of transition and the magnitude of the context effect, awaits further investigation.

Though both of these processes seem to be responsible for the perceptual compensation of articulatory undershoot in connected speech, their relative roles are found to be subject to individual differences. While the characteristics of auditory extrapolation were quite similar for all three subjects, the effects of dynamic context on categorical judgment were rather dissimilar, being contrast in two subjects and assimilation in another. This fact is also in agreement with the results of Lindblom and Studdert-Kennedy (1967), which indicate that the effect of /w-w/ context was contrast for most of their subjects but assimilation for the rest of them. Whether the observed individual variability in context effect is also observed in perception of speech in a larger context, or it is an artefact produced by the experimental paradigm adopted, also calls for further study.

Finally, it is quite important to know to what extent these perceptual processes are mediated by articulation, although its role was only implicit in the above interpretation of the obtained results. For instance, if reference to articulation or its underlying motor commands is playing a major role in the perceptual compensation of time-varying acoustic characteristics, better performance will be expected along dimensions where articulatory undershoot is more likely to occur. Further work is in progress to elucidate this point.

References

Arai, S., Y. Kanamori, H. Kasuya and K. Kido (1971); Model for the perception of vowels in connected speech, Report of Technical Committee on Electroacoustics, IECE Japan, No. EA 70-27 (1971).

Benguerel, A-P and H. A. Cowan (1974); Coarticulation of upper lip protrusion in French, Phonetica, 30, 41-55.

Brady, P. T., A. S. House and K. N. Stevens (1961); Perception of sounds characterized by a rapidly changing resonant frequency, J. Acoust. Soc. Am., 33, 1357-1362.

Fry, D. B., A. S. Abramson, P. D. Eimas and A. M. Liberman (1962); The identification and discrimination of synthetic vowels, Language and Speech, 5, 171-188.

Fujisaki, H., M. Yoshida, Y. Sato and Y. Tanabe (1973); Automatic recognition of connected vowels using a functional model of the coarticulatory process, J. Acoust. Soc. Japan, 29, 636-638.

Fujisaki, H., Y. Sato, Y. Noguchi and T. Yamakura (1968); Automatic recognition of semivowels in word context, J. Acoust. Soc. Japan, 31.

Hiki, S. H. Sato,, T. Igarashi and J. Oizumi (1968); Dynamic model of vowel perception, Reports of 6th ICA, Tokyo.

Kanamori, Y., Y. Shigeno and K. Kido (Oct. 1974); Dynamic perception of vowels, Reports of Autumn Meeting, Acoust. Soc. Japan.

Kuwahara, H. and H. Sakai(1972); Perception of vowels and C-V syllables segmented from connected speech, J. Acoust. Soc. Japan, 28, 225-234.

_____ (1975); An experiment on the phoneme boudnary locations in the dynamic perception of synthetic vowels, J. Acoust. Soc. Japan, 31, 18-23.

Lindblom, B. (1963); Spectrographic study of vowel reduction, J. Acoust. Soc. America, 35, 1773-1781.

Lindblom, B. and M. Studdert-Kennedy (1967); On the role of formant transition in vowel recognition, J. Acoust. Soc. America, 42, 830-843.

Thompson, C. L. and H. Hollien, (1970); Some contextual effects on the perception of synthetic vowels, Language and Speech 13, 1-13.