TEMPORAL ORGANIZATION OF SEGMENTAL FEATURES IN JAPANESE DISYLLABLES

Hiroya Fujisaki and Norio Higuchi, Faculty of Engineering, University of Tokyo, Tokyo, Japan

Introduction

While it is apparent that the realization of successive units in connected speech is based on the proper timing of articulatory and phonatory events, much remains to be investigated regarding the nature of the timing mechanism. It is not even agreed whether the regularity of timing (isochrony) resides in speech production or in speech perception, as pointed out by Lehiste (1977). The lack of our knowledge on this issue may primarily be due to the fact that the speech signal often fails to display marked segment boundaries, and that even the apparent boundaries do not directly reveal the timing of production nor the timing of perception. Elucidation of the mechanism underlying the isochrony thus requires experimental techniques for extracting, from the speech signal, the indices for the timing of production as well as the indices for the timing of perception of each of the successive units.

The present paper deals with both the productive and the perceptual aspects of the segmental timing in Japanese disyllabic words consisting only of vowels. Disyllabic words were selected since they display the characteristics of connected speech on the smallest scale. Vowel sequences were chosen since their acoustic characteristics can be most clearly defined in terms of formant frequencies, and the articulatory transition from the initial vowel to the second vowel can be traced in the trajectories of their formant frequencies.

The Speech Material

The speech material consisted of 20 disyllables, i.e., all the possible pairs of the five Japanese vowels (/i/, /e/, /a/, /o/, and /u/), pronounced with the "flat-type" word accent. Among these disyllables, nine were meaningful with the given accent type, four were meaningful when pronounced with a different accent type, and the rest were nonsense words. A randomized list of 100 words, containing five tokens each of the 20 disyllables, was read by a male speaker of the Tokyo dialect of Japanese. These disyllables were pronounced in isolation at an interval of three seconds. The speech signal was sampled at 10 kHz with an accuracy of 11 bits/sample and stored in the magnetic tape memory of a digital computer.

Analysis of Segmental Timing at the Level of Speech Production

An LPC analysis was made of all the utterances to extract the frequencies and bandwidths of 11 poles, from which the first three formant frequencies were selected on the basis of bandwidth and the continuity of the trajectories. These trajectories were then used to estimate the onset of the transition from the initial to the second vowel.

The estimation was based on the model of the coarticulation process in connected vowels previously proposed by Fujisaki et al. (1974, 1977). As shown in Fig. 1, the entire production process for connected vowels is represented by a hypothetical linear system which converts the stepwise target formant frequencies of each vowel into actual formant trajectories. An analysis of observed formant trajectories has indicated that a good approximation can be obtained by a critically-damped second-order linear system.

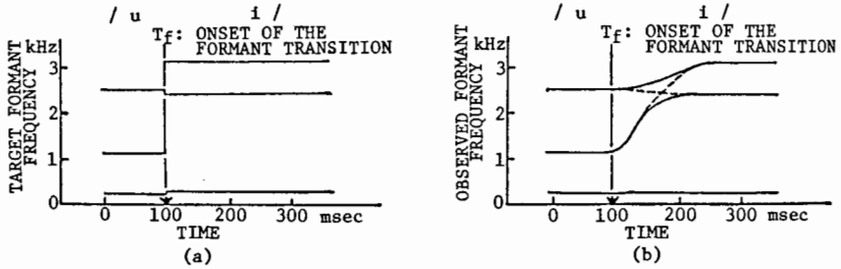

Fig. 1. Formulation of the process of coarticulation in the formant frequency domain: conversion of idealized formant target (a) into actual formant trajectories (b).

In the case of the disyllables under study here, we may assume a target frequency for the nth formant as

$$C_n(t) = F_{n1} + (F_{n2} - F_{n1})\,u(t - T_f),$$

where F_{ni} denotes the target frequency of the nth formant of the ith vowel, and T_f denotes the onset of the transition measured from the voice onset of the initial vowel as the origin of the time axis. Then the actual formant frequency can be given by

$$F_n(t) = F_{n1} + (F_{n2} - F_{n1})\{1 - (1 + \frac{t - T_f}{\tau_n})\exp(-\frac{t - T_f}{\tau_n})\}u(t - T_f),$$

where τ_n denotes the time constant for the transition of the nth formant. Further considerations regarding the continuity and coupling of the resonance modes lead to good approximations of the formant trajectories for all of the vowel combinations. When a set of observed formant trajectories ($F_1(t)$, $F_2(t)$, and $F_3(t)$) is given, it is possible, by the method of Analysis-by-Synthesis, to determine the common onset of the formant transition and the time constants of individual formant trajectories. In the following analysis, a common time constant τ_2 was assumed for the second and third formants. Examples of the observed formant trajectories and their best approximations by the above-mentioned model are shown in Fig. 2 for /ui/ and /iu/, where the estimated onset T_f of the formant transition is also indicated.

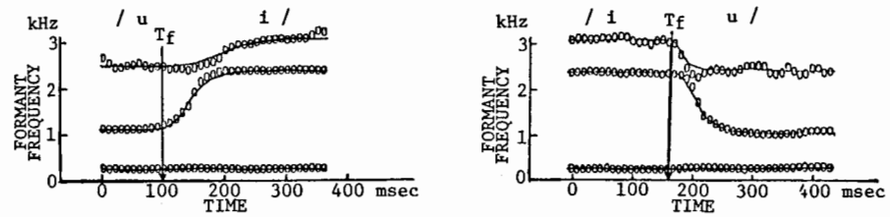

Fig. 2. Observed formant frequency trajectories (⋯⋯), their best approximations (——), and the estimated onset of the formant transition (T_f) for /ui/ (left) and /iu/ (right).

Table 1 summarizes the results for all the utterance samples and lists the mean values of T_f and τ_2 for five tokens of each disyllable. The following comments can be drawn from a comparison of the results for pairs of disyllables having the same vowel combination in a different order.

first vowel		second vowel /i/	/e/	/a/	/o/	/u/
/i/	T_f	–	155	131	136	134
	τ_2		21	25	22	27
/e/	T_f	90	–	132	134	149
	τ_2	59		40	26	20
/a/	T_f	119	125	–	124	125
	τ_2	48	39		41	37
/o/	T_f	101	94	92	–	113
	τ_2	44	58	51		-
/u/	T_f	108	117	126	146	–
	τ_2	39	44	28	-	

Table 1. Mean values for the interval (T_f[msec]) from voice onset to the onset of formant transition and for the time constant (τ_2[msec]) of the second formant trajectory for the 20 disyllabic words.

(1) In disyllables involving jaw movement without a change in lip articulation (i. e., /ie/ vs. /ei/, /ea/ vs. /ae/, /ia/ vs. /ai/, and /uo/ vs. /ou/), T_f is always larger for the disyllable produced by an opening movement of the jaw than for that produced by a closing movement. Analysis of variance indicates that the difference is highly significant (0.1% level) in /ie/ vs. /ei/, and is also significant (1% level) in /uo/ vs. /ou/, as well as in /ia/ vs. /ai/.

2) In disyllables involving changes in lip articulation with or without minor jaw movement (i. e., /iu/ vs. /ui/, /eu/ vs. /ue/, /io/ vs. /oi/, /eo/ vs. /oe/, and /ao/ vs. /oa/), T_f is always larger for the disyllable produced by a rounding of the lips than for that produced by an unrounding of lips. The difference is significant in /eu/ vs. /ue/ (1% level); /ao/ vs. /oa/ (2% level); /eo/ vs. /oe/ (2% level); /io/ vs. /oi/ (5% level); and /iu/ vs. /ui/ (5% level).

(3) No significant difference in T_f was found for /au/ vs. /ua/, which involve both major jaw movement and changes in lip articulation in the transition from the initial to the second vowel. The effects of these two articulatory factors are considered to counteract and cancel each other.

These points may be easily observed in Fig. 3, which schematically shows the regions of the vowel target on the F_1 - F_2 plane. An arrow from one vowel target to another corresponds to a disyllable,

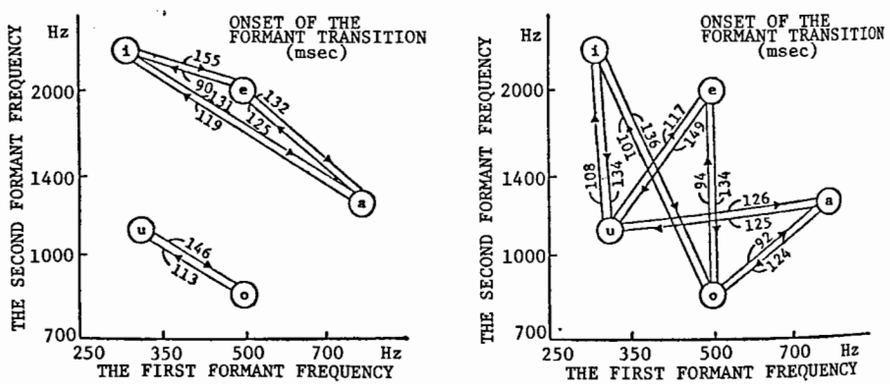

Fig. 3. Direction of the formant transition in the first and the second formant frequency plane and the onset of the formant transition (T_f).

and the number associated with the arrow indicates the mean value of T_f (in msec) for that disyllable.

Furthermore, there exists a very high negative correlation between T_f and τ_2 ($r = -0.91$) as shown in Fig. 4. Hence,

(4) Differences in the onset of transition (T_f) tend to compensate for the differences in the rate of transition; a slower transition is initiated earlier and vice versa.

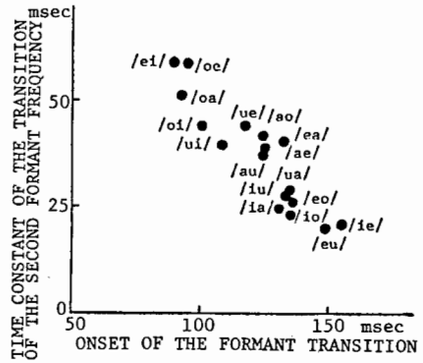

Fig. 4. Relationship between the onset of the formant transition (T_f) and the time constant (τ_2) of the second formant trajectory.

Analysis of Segmental Timing at the Level of Speech Perception

The last finding of the preceding section suggests the possibility that the apparent diversity in the onset of transition in various disyllables is introduced to maintain the uniformity of the perceived duration of segments. The following experiment was designed to investigate this possibility, using the same utterance samples as in the above analysis to find the instant of the perceptual onset of the second vowel within a disyllable.

A set of 20 points were selected at intervals of 5 msec to cover the range of the major formant transitions in the waveform of each disyllabic utterance. Twenty tokens of truncated disyllables were then prepared by curtailing the original speech waveform at these 20 points. These tokens were arranged in serial order at an interval of 3.5 sec as stimuli in an identification test using the method of limits. The subject was asked to answer whether he heard one vowel segment or two in a truncated disyllable. The test was repeated to obtain the response probability, and the perceptual onset of the second vowel was defined as the point corresponding to an equal probability for the two alternatives. An example of the stimuli and the subject's response probability is schematically il-

lustrated in Fig. 5. The test was conducted using one utterance of each of the twenty disyllables. The subjects were two male speakers of the Tokyo dialect.

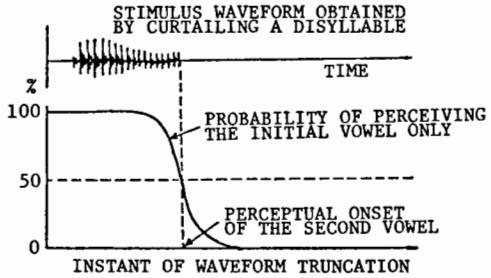

Fig. 5. Determination of the perceptual onset of the second vowel in a disyllable by waveform truncation.

Figure 6 shows the relationship between the perceptual onset (T_p) of the second vowel and the onset of formant transition (T_f) for each of the disyllables. Both T_p and T_f are expressed by their values relative to the total duration of an utterance. While the T_f's for the various disyllables are distributed over a very wide range (22% - 42%), the Tp's are found to be concentrated within a rather narrow range around the center of each utterance (48% - 53%). These findings suggest that the apparent diversity in the onset of the second vowel at the level of speech production may be the consequence of the speaker's effort to maintain the uniformity of perceived syllabic durations regardless of vowel combinations.

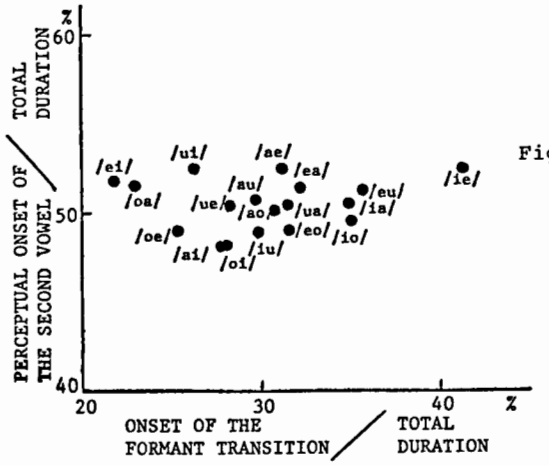

Fig. 6. Relationship between the perceptual onset (Tp) of the second vowel and the onset of the formant transition (T_f) for one sample of each of the disyllables.

Discussion

Two models of the possible mechanisms underlying the temporal organization of speech have been presented by Kozhevnikov and Chistovich (1965) and have since been widely discussed, e. g. by Ohala (1970), Leanderson and Lindblom (1972), and others. One is the so-called "chain model" based on the hypothesis of a closed-loop control of the speech production process. The other is the so-called "comb model" based on the hypothesis of an open-loop control. From our present knowledge concerning the motor organization of skilled behaviors, the chain model may be discarded, although it may certainly be true that various modes of feedback are necessary for the formation of the motor program. The findings of our present study suggest, however, that the comb model requires further elaboration. Our findings suggest that the formulation of the temporal relationship between the motor control and the articulatory/acoustic realizations of speech units is not complete without a consideration of their relationship to perceptual timing. From this point of view, two possible models can be distinguished under the open-loop (or "comb") hypothesis, as shown in Fig. 7.

In model (a), successive segments are produced with an isochronism at the level of the motor commands, so that their articulatory/acoustic realizations are not necessarily isochronous because of differences in the physiological and physical properties of the various articulators, as well as in the manner of articulation. In model (b), on the other hand, the motor commands and the articulatory/acoustic realizations of successive segments are programmed in such a way that the perceptual onsets of successive segments occur

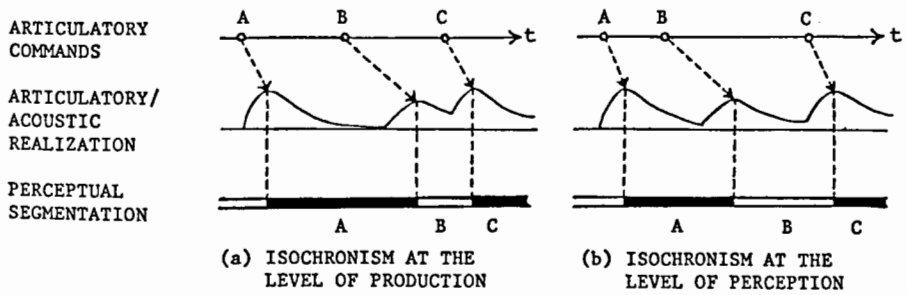

Fig. 7. Two models of the mechanisms underlying the temporal organization of speech units under the open-loop control hypothesis.

with an isochronism, viz., the perceived durations of these segments are kept equal. The results of the present study may be considered as corroborating model (b) as far as the Japanese disyllables are concerned.

Conclusion

Temporal organization of speech segments was investigated using disyllabic Japanese words consisting only of vowels. An acoustic analysis of their formant trajectories has indicated that the onset of the transition to the second vowel in various disyllables is distributed over a relatively wide range. This variation tends to compensate for the differences in the rate of transition due to differences in the articulator(s) involved and the direction of movement. On the other hand, a perceptual analysis of the onset of the second vowel has indicated that the perceptual onset of the second vowel in utterance samples of the same disyllable is concentrated within a relatively narrow range regardless of the particular vowel combination or the order of the vowels in the disyllable. The implication of these findings for the possible mechanisms underlying the temporal organization of speech units was discussed in connection with two models already proposed with regard to these mechanisms.

References

Fujisaki, H. et al. (1974): "Formulation of the coarticulatory process in the formant frequency domain and its application to automatic recognition of connected vowels," Proc. SCS-74 3, 385-392.

Fujisaki, H. (1977): "Functional models of articulatory and phonatory dynamics," in Articulatory Modeling and Phonetics, R. Carré, R. Descout, and M. Wajskop (eds.), 127-136, G. A. L. F. Group de la Communication Parlée.

Kozhevnikov, V. A. and L. A. Chistovich (1965): Speech: Articulation and Perception, Moscow: Nauka.

Leanderson, R. and B. E. F. Lindblom (1972): "Muscle activation for labial speech gestures," Acta Otolaryng. 73, 362-373.

Lehiste, I. (1977): "Isochrony reconsidered," Journal of Phonetics 5, 253-263.

Ohala, J. (1970): "Aspects of the control and production of speech," Working Papers in Phonetics 15, UCLA.