Ann. Bull. RILP, No. 11, 77-84 (1977)

PERCEPTION OF SEGMENTAL AND SUPRASEGMENTAL UNITS IN TIME-VARYING FORMANT AND FUNDAMENTAL FREQUENCIES*

Sōtarō Sekimoto and Hiroya Fujisaki

Introduction

As it is quite well known, the acoustic manifestations of discrete segmental units, i. e., phonemes, are scarcely discrete but occur in continuous transitions. It may thus be inappropriate to define segmental durations purely in terms of their acoustic characteristics. At least for a certain class of phonemes like vowels and semivowels, however, it is possible to define segmental duration on the basis of their perceptual presence or absence.[1-3] The situation is essentially the same for larger suprasegmental units like syllable or mora which is a suprasegmental unit of duration in spoken Japanese, since they cannot be exempt from coarticulation.[4, 5]

In many languages, segmental units and suprasegmental units are generally different in their temporal extent. The spoken Japanese present rather interesting instances, however, in that a vowel may often constitute a "mora", and thus serve as a segmental and as a suprasegmental unit at the same time. There are many examples of two-mora words which consist of two vowels as in the 2-mora noun /āi/ ('love'), the 2-mora verb /āu/ ('meet'), and the 2-mora noun /iē/ ('house'), and so on. There also exist cases where the first and the second segments are identical vowels and thus appear as one elongated vowel acting as two morae. Examples are: the 2-mora adjective /ii/ ('good'), the 3-mora adjective /ō:i/ ('many'), the 3-mora noun /o:i/ ('cover'), and so on.

In these and many other examples, the segmental information is carried by the time-varying formant frequencies (formant trajectories), while the suprasegmental information is carried mainly by the trajectory of the voice fundamental frequency, or F_0-contour. Figure 1 shows examples of such a trajectory of the fundamental frequency in the upper half, and those of formant frequencies in the lower half as extracted from an utterance of a 2-mora noun /āi/ by a digital computer.

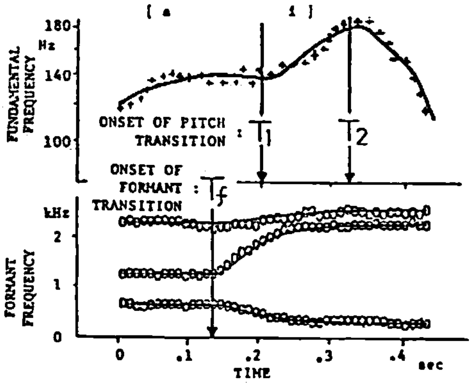

Fig. 1. Comparison of formant and pitch patterns and their Analysis-by-Synthesis for the word accent type B of [ai].

* Paper presented at the IX International Congress on Acoustics, Madrid, July, 1977.

If we look at these trajectories, we can easily notice that it may be possible to locate the onset of successive transitions. In fact, by introducing appropriate models both for the control characteristics of the fundamental frequency and for those of formant frequencies, it has been shown that it is possible to detect the instants of the underlying phonatory and articulatory controls. These instants, indicated by T_2 for the downward transition of F_0-trajectory for the second mora, and by T_f for the formant transition toward the second vowel /i/, respectively mark the boundary between segmental units and the boundary between suprasegmental units at the level of production. It should also be pointed out that a consistent delay of 50-70 msec was found in the timing of F_0-control as compared to that of formant control.[6, 7]

It is apparent, however, that these timing of production may not necessarily coincide with the timing of perception. For example, it can easily be expected that perception of the second vowel [i] may not occur at the onset of formant transition (T_f), but will occur only after a major portion of the formant transition is produced. The purpose of the present study is to obtain quantitative understanding of the perceptual process of segmental and suprasegmental units, and their possible interactions. Two-mora words consisting of two vowels are adopted as stimuli, since they are apparently the minimal material for investigating perception of these units in the context of connected speech. These words were synthesized by digital computer simulation to allow free and precise control of the timing parameters.

Experimental Method

The method of constant stimuli was adopted for determinig perceptual boundaries between segments. The stimuli were synthetic words truncated at various points. One series of experiments used only the "head" or the initial portion of the truncated word, while the other series of experiments used only the "tail" or the final portion of the truncated word, as shown in Figure 2.

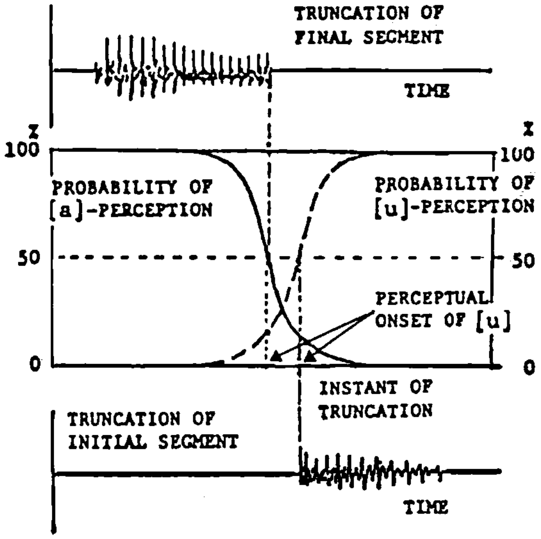

Fig. 2. Determination of perceptual segment boundary in [au] by waveform truncation.

In each series of experiment, a set of 10 points of truncation were seclected at equal intervals on the time axis, producing 10 different "head" stimuli or "tail" stimuli. The 10 head stimuli were presented in random order at interval of 4 sec, and the subjects were instructed to answer, by forced judgement, whether they heard each stimulus as consisting of one unit or two units. The probability of response for one unit, for example, plotted against the time of truncation, can be approximated by a normal distribution, and the point of 50% judgement can be defined as the perceptual boundary between the two units, or the perceptual onset of the second unit, on the time axis. Boundaries obtained by using "head" stimuli and "tail" stimuli were generally different. In the following analysis, both boundaries were examined, though the "head" stimuli generally produced higher consistency of results.

The stimuli were generated by computer simulation of a terminal analog speech synthesizer with five formants, in which the fourth and the fifth formant frequencies were held constant at 3.5 kHz and 4.5 kHz, respectively. The study consisted of three experiments:

Experiment 1:	perception of segmental units,
Experiment 2:	perception of morae (suprasegmental units),
Experiment 3:	their interactions.

Experiment 1

Figure 3 shows patterns of fundamental and formant frequencies used in Experiment 1. All the formants, as well as the fundamental frequency, were held constant except for the first formant frequency, producing four different versions of synthetic /au/ with different rates of transition (20, 30, 40, and 60 msec). The first formant trajectory was calculated as the

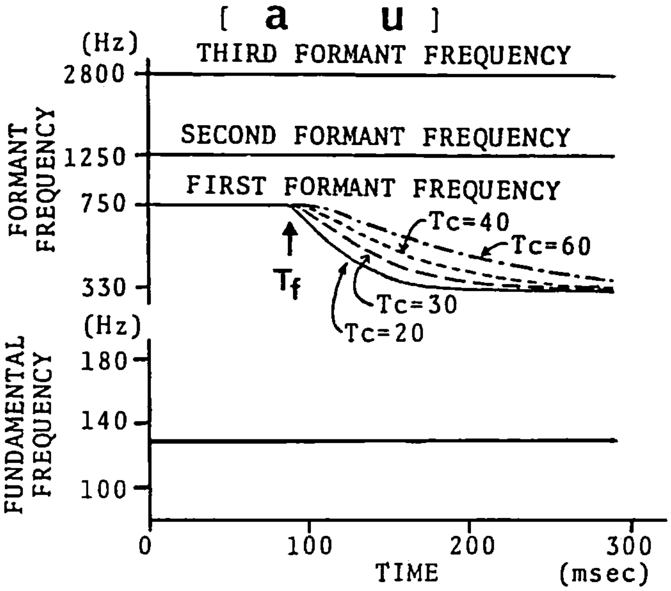

Fig. 3. Formant frequency transitions for four different values of time constant (T_c).

response of a second order linear system to a stepwise input, which was shown to be capable of closely approximating the formant trajectory found in normal utterances.[8] In these stimuli, the onset of transition started at 90 msec after the voice onset, and the total stimulus duration was always kept at 290 msec.

An example of results of Experiment 1 is shown in Figure 4, where the perceptual onset of the second vowel, relative to the onset of first formant transition, is plotted against the time constant of formant transition, denoted by T_c. Though the results obtained by the two kinds of stimuli are appreciably different, i. e., the onset is earlier for "head" stimuli than for "tail" stimuli, they both vary linearly with T_c. At 30 msec of T_c, which is close to the values found in natural speech, the delay of perceptual onset is about 60-70 msec.

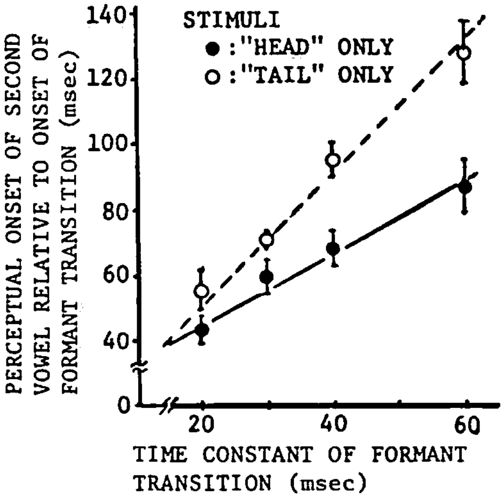

Fig. 4. Delay of perceptual onset of second vowel relative to onset of formant transition as function of T_c.

It may also be of interest to see how much of the total excursion in the formant transition is needed before it elicits the perception of the second vowel segment. From this point of view, the results of Experiment 1 can be replotted as shown in Figure 5. It indicates that "head"

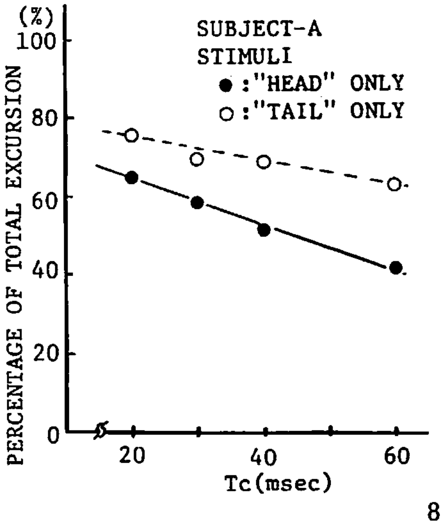

Fig. 5. Percentage of total excursion of formant transition at perceptual onset of $/V_2/$ as function of T_c.

stimuli are again more effective in signaling the presence of V_2 than "tail" stimuli are. It also indicates, at a time constant (T_c) of 30 msec and with the "head" stimuli, for example, the perception of the second vowel starts when 60% of the total formant transition has been traversed, and this value gets smaller for a larger T_c, in other words, for a slower formant transition.

Experiment 2

The second experiment was aimed at finding the relationship between F_0-pattern and perception of the suprasegmental unit "mora" using stimulus parameters shown in Figure 6. Here, the formant frequencies were held constant for the entire stimulus, and were combined with either one of the four F_0-patterns. Two of the F_0-contours, i. e. A-1 and A-2, were simulations of F_0-contour found in "high-low" word accent (Type A), while the remaining two, i. e. B-1 and B-2, simulated those found in the "low-high" word accent (Type B). Contours A-1 and A-2 differed by 50 msec in the timing of the downward F_0-transition. Since the F_0-model has been described elsewhere,[9] it may be sufficient just to mention that it provides very close approximations to those F_0-contours found in natural utterances, and yet allows complete control of the timing parameters, which were found to be quite essential for the perception of accent types.

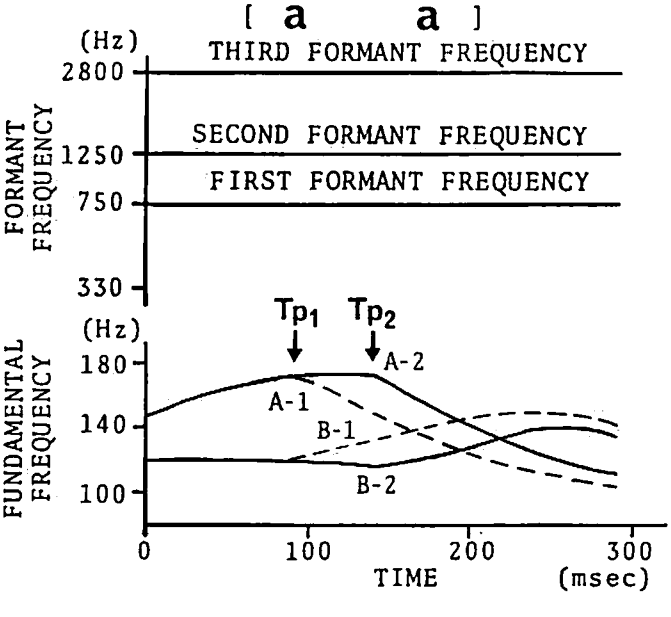

Fig. 6. Formant and pitch patterns for the word accent type A-1, A-2, B-1, and B-2 of [aa].

The procedures of Experiment 2 was the same as in Experiment 1, except that the subjects were instructed to answer whether they heard the truncated word as one mora or two marae. Figure 7 shows the results

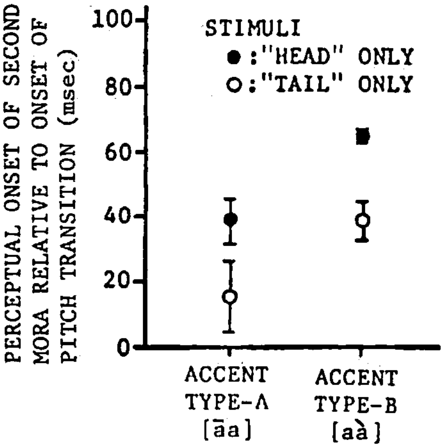

Fig. 7. Delay of perceptual onset of the second mora in two accent types of [aa].

for F_0-contours of A-2, and B-2 only, and indicates that there also exists a perceptual delay between the onset of relatively rapid pitch transition and the perceptual onset of the second mora, being greater for "head" stimuli than for "tail" stimuli. Compared to the perceptual delay of 60 msec in the case of perceptual onset of the second vowel, the perceptual delay here seems to be significantly smaller for Type A (with rapid downward transition), but may be of the same order as that found in the case of perceptual onset of segmental units in Type B.

Experiment 3

Having thus examined perception of segmental and suprasegmental units separately, we intended to find out, in Experiment 3, if there existed any significant interactions between the two. Figure 8 shows parameters

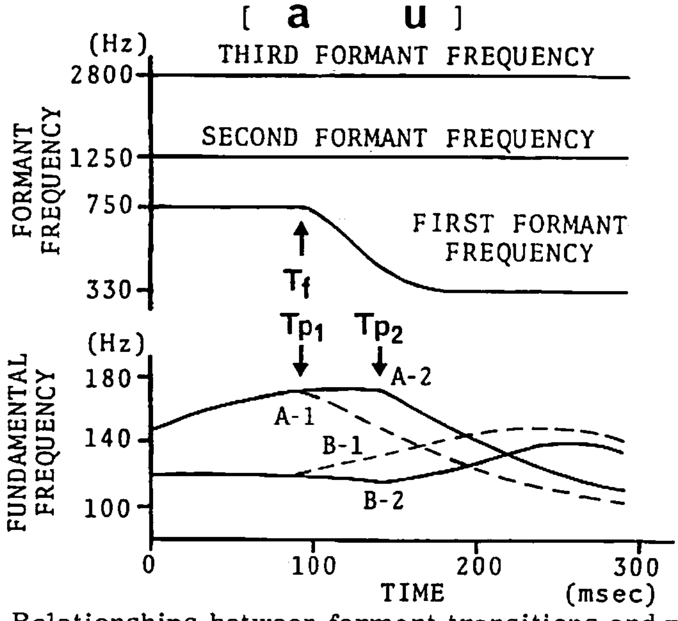

Fig. 8. Relationships between formant transitions and pitch contours for accent type A-1, A-2, B-1, and B-2.

of the four stimulus words [au] consisting of two different versions each of Type A and Type B accent. These stimuli were used both for examining possible effects of F_0-contour on boundary of segmental units, and for examining effects of formant pattern on boundary of suprasegmental units.

Figure 9 shows the influence of suprasegmental features (represented by F_0-contour characteristics) upon perceptual onset of a segmental unit, i. e., the second vowel. Considering the fact that A-1 and A-2 as well as B-1 and B-2 differed by 50 msec in their crucial timing of accent control, the effects of accent types as well as of their timing are seen to be rather small. In fact, a shift of F_0-contour timing of 50 msec is seen to cause a shift of only several msec in the perceptual onset of the second vowel.

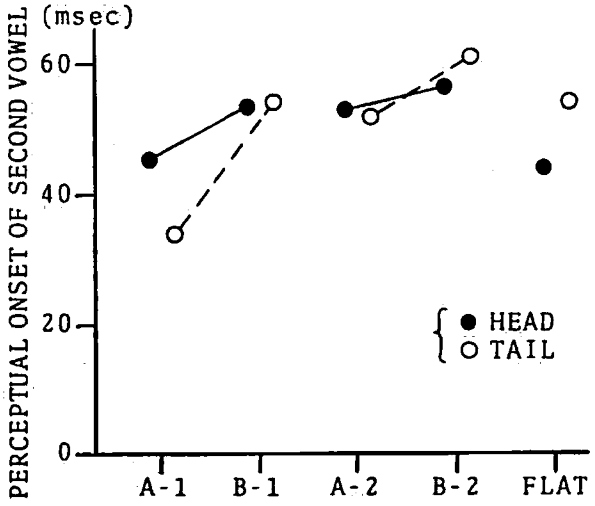

Fig. 9. Perceptual onset of second vowel in [au] vs. pitch contour characteristics.

Similar experiments were also conducted using the same stimuli but with different instructions. This time, the subjects were instructed to report the number of morae they heard. The results indicated still smaller effects of formant pattern on the perceptual onset of suprasegmental units.

Concluding Remarks

Temporal relationships between continuous trajectories of fundamental and formant frequencies and their corresponding underlying discrete linguistic units, both segmental and suprasegmental, have been investigated using synthetic speech stimuli with acoustic characteristics closely approximating those found in natural utterances. The results revealed the existence of considerable delay between the onset of acoustic characteristics, defined as the timing of the stepwise command estimated from observed formant or fundamental frequency trajectories, and the perceptual onset of the corresponding linguistic unit. A delay of about 60 msec was observed between the onset of formant transition and the perceptual onset of the second vowel for the T_c value found in natural speech. A similar delay was also observed between the onset of the pitch transition and the perceptual onset of the second mora. Mutual interaction between perception of segmental and suprasegmental units, however, seems to be rather small.

References

1. Fujisaki, H., H. Morikawa, and M. Sugito (1976); Temporal Organization of Articulatory and Phonatory Controls in Realization of Japanese Word Accent, Ann. Bull. RILP, No. 10, 191-198.
2. Itahashi, S. and S. Chiba (1976); On the Relation Between Auditory Phonemic Segment and Second Order Model, Record of Spring Meeting, Acoust. Soc. Japan, 333-334.
3. Sekimoto, S. and H. Fujisaki (1976); Influences of Formant and Pitch Transitions on Perceptual Onset of a Phonemic Segment in Connected Vowels, Record of Autumn Meeting, Acoust. Soc. Japan, 355-356.
4. Fujisaki, H. and S. Sekimoto (1976); Perception of Phonemes and Morae in Time-Varying Fundamental and Formant Frequencies, Transactions of the Committee on Speech Research, Acoust. Soc. Japan, S76-16.
5. Sekimoto, S. and H. Fujisaki (1977); Perception of Phonemes and Morae in Fundamental and Formant Frequency Transitions, Record of Spring Meeting, Acoust. Soc. Japan, 119-120. .
6. Fujisaki, H., H. Morikawa and M. Sugito (1976); Temporal Organization of Articulatory and Phonatory Controls in Realization of Word Accent, Record of Spring Meeting, Acoust. Soc. Japan, 229-230.
7. Fujisaki, H., H. Hirose and M. Sugito (1976); Acoustic and Electromyographic Observations of Temporal Relationships Between Articulatory and Phonatory Controls, The Japan Journal of Logopedics and Phoniatrics, 17, 134.
8. 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.
9. Fujisaki, H. and H. Sudo (1971); A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent, J. Acoust. Soc. Japan, 27, 445-453.