

TEMPORAL ORGANIZATION OF ARTICULATORY AND PHONATORY  
CONTROLS IN REALIZATION OF JAPANESE WORD ACCENT\*

Hiroya Fujisaki, Hiroyoshi Morikawa\*\* and Miyoko Sugito\*\*\*

I. Introduction

The phonemic information of a linguistic message is realized predominantly through the medium of the articulatory control of speech organs, and is manifested as time-varying segmental features of speech, which can most precisely be specified in terms of such acoustic parameters as formant frequencies. The prosodic information such as stress and accent, on the other hand, is realized mainly by the phonatory control of the larynx, and is manifested as a set of suprasegmental features whose acoustic correlates are such parameters as the intensity and the fundamental frequency of the glottal source. There exist, of course, certain exceptions to this oversimplified statement of the roles of articulatory and phonatory controls. For example, voicing information of a phoneme is evidently realized by the phonatory control, while prosodic information on quantity is realized mostly by the timing of the articulatory control. It is apparent, however, that a certain kind of synchronism has to be maintained between these manifestations of the phonemic and the prosodic information in order that they should integrate into a meaningful linguistic entity.

Although there now exists a fair amount of quantitative knowledge both on the segmental and on the suprasegmental features of speech, comparatively little is known about the nature of their synchronism, i. e., their temporal organization as well as the coordination of articulatory and phonatory controls underlying their realization. For example, when going from one phonemic segment to another, it is not even known whether these features undergo simultaneous changes or whether changes in one precede those in the other. The paucity of quantitative findings in this respect may possibly be due to the lack of means both for accurately extracting the acoustic correlates of these features and for properly interpreting patterns of their temporal variations in terms of their underlying controls.

The present paper describes the methods and the results of our preliminary investigation into the temporal relationships between segmental and suprasegmental features of speech in the realization of Japanese word accent. The word accent found in many dialects of Japanese is of particular interest in this respect since each accent type is characterized by a specific pattern in the fundamental frequency contour which remains essentially the same irrespective of the phonemic constituents of words as long as they share the same number of morae, and since there exist many instances of words which consist of an identical phonemic sequence and

---

\* Paper to be presented at the Third World Congress of Phoneticians, Tokyo, August 23-28, 1976.

\*\* Department of Electrical Engineering, Faculty of Engineering, University of Tokyo

\*\*\* Department of Japanese Literature, Faculty of Education and Liberal Arts, Osaka Shoin Women's College.





differ only in their accent types. The mutual independence of the phonemic and the accentual information strongly suggests that the corresponding articulatory and phonatory controls are programmed separately yet executed simultaneously in the utterance of a specific word.

Segmental features were extracted as formant frequency trajectories from short-time frequency spectra of speech, while suprasegmental features were extracted as the contour of the fundamental frequency of the glottal source. The temporal relationships between these features were quantified by analyzing the time-varying patterns of these acoustic parameters on the basis of functional models for their control processes, and the instants of initiation of articulatory and phonatory controls were estimated, the results indicating that they were not exactly simultaneous but were possibly brought into an approximate synchronism by a certain mediating process. The possible role of perception as the mediating process was then examined both by identification of truncated utterances and by click location to determine perceptual segment boundaries. It was found that the perceptual segment boundary roughly coincided with the instant of initiation of the phonatory control characteristic of each word accent type, thereby suggesting the role of perception in the coordination of articulatory and phonatory controls.

## II. Speech Materials

The speech materials adopted for the present investigation were two-mora words of the Osaka dialect consisting only of vowels such as [ai], [ao], [ie], [ue], etc. The Osaka dialect was selected because of its richness of accentual types for words with the same phonemic constituents.<sup>1</sup> For example, the same two-mora phoneme sequence may possess as many as four accent types, as compared to two in the Tokyo dialect. Examples of these four types, designated as Types A, B, C, and D, are shown in Table 1. Words consisting of vowel sequences were selected since their

Table 1. Examples of accent types of two-mora words in the Osaka dialect.

Accent type	A	B	C	D
Subjective pitch				
[ai]	'love'	————	'blue'	'interval'
[ue]	'hunger'	————	————	'top'

segmental features could be most precisely specified by trajectories of formant frequencies, and thus lent themselves to accurate estimation of the instants of initiation of the underlying articulatory control.

The speaker was a 50-year-old male who was born, was educated, and spent his entire life in Osaka. The two-mora words were pronounced in two accent types: Type A (high-low) and Type B (low-high-low). For

certain phonemic sequences both types were meaningful words, whereas for other sequences only one of the two was meaningful.

The words were listed in random order and were read repeatedly and recorded in a quiet room. The speaker was asked to produce two series of recordings: one for careful reading and the other for casual reading. Because of greater variabilities in data obtained by casual reading, the following discussions are based mainly on the analysis of materials recorded by careful reading.

### III. Methods and Results of Acoustic Analysis

#### 3.1. Extraction of segmental features

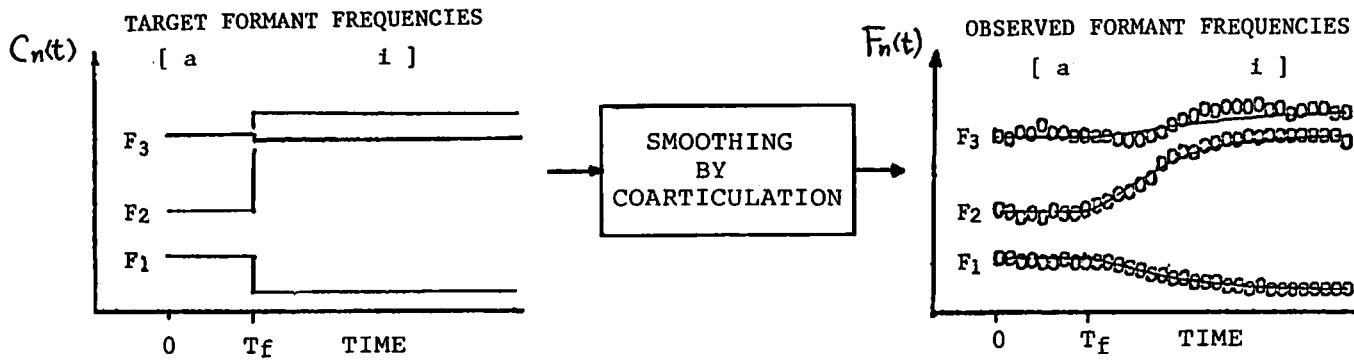
Analysis and extraction of segmental and suprasegmental features were performed by a digital computer. The recorded materials were sampled at 10 kHz upon playback, quantized with an accuracy of 10 bits per sample, and stored in the computer's magnetic tape memory.

As is well known, segmental features of vowels are mainly consequences of the transfer characteristics of the vocal tract, and are thus most accurately specified by their formant frequencies, which are defined as the frequencies of poles of the vocal tract transfer ratio relating the volume velocity at the oral aperture to the one at the glottis.<sup>2</sup>

The transfer characteristics of the vocal tract are reflected on the spectral envelope of speech. For connected speech, whose segmental features are never stationary but vary continuously with time, reliable estimates of the short-time spectral envelopes can be obtained by pitch-synchronous spectral analysis, using a time window whose width is exactly equal to one fundamental period, centered at successive local maxima of the absolute amplitude of the speech signal.<sup>3</sup> In the following analysis, this procedure was slightly simplified by using a hanning window with a constant width of 7.2 msec, which was approximately equal to the mean fundamental period of all utterances by the speaker.

From the short-time spectral envelope, formant frequencies were extracted by the method of Analysis-by-Synthesis in the frequency domain,<sup>4</sup> i. e., by finding the optimum set of formant frequencies of a synthetic spectral envelope which would give the closest approximation to the measured spectral envelope. In the present study, the model for synthesis of spectral envelopes consisted of factors representing contributions of vocal tract transfer function, glottal source, and radiation, and of an additional factor for miscellaneous frequency characteristics not accounted for by other components. Frequencies of the first four formants were thus extracted from a short-time spectral envelope by successive approximation to minimize the mean squared error between the observed and the synthesized envelopes expressed in (dB)<sup>2</sup>. These formant frequencies were determined pitch-synchronously with a quantum step of 50 Hz, but were converted by interpolation into uniformly sampled formant trajectories at intervals of 5 msec.

The timing of the underlying articulatory control was then estimated by Analysis-by-Synthesis of trajectories of the first three formants, i. e., by finding the optimum set of synthetic formant trajectories giving the closest approximation to the observed trajectories.<sup>5</sup> The model for the synthesis of formant trajectories was an approximate formulation of the coarticulatory process between successive vowels, based on an assumption that the



Target formant frequencies :  $C_n(t) = F_{n1} + (F_{n2} - F_{n1}) u(t - T_f),$

Observed formant frequencies :  $F_n(t) = F_{n1} + (F_{n2} - F_{n1}) \left[ 1 - (1 + \delta t - T_f) \exp(\delta t - T_f) \right] u(t - T_f),$

$T_f$  : instant of initiation of the second vowel [i].

Fig. 1. Formulation of the process of coarticulation in the formant frequency domain: conversion of idealized formant targets into actual formant trajectories.

entire process of conversion from a string of vowel phonemes to the formant trajectories can be represented by a hypothetical linear system which accepts stepwise target formant frequencies of each vowel as input and smoothes them into the observed formant trajectories. Analysis of formant trajectories revealed that a fair approximation could be obtained by the step response of a critically damped second-order linear system, if a proper set of parameters, i. e., target formant frequencies, instants of command onset, and the rate of formant transition, were selected. The underlying concept of this model is illustrated in Fig. 1, together with mathematical expressions of the input command  $C_n(t)$  and the smoothed trajectory  $F_n(t)$  of the  $n$ th formant frequency.

The above formulations allow one to estimate target formant frequencies as well as the instant of initiation of the articulatory control for each of the successive vowel phonemes in a sequence of vowels, and also to estimate the rate of transition from one vowel to another which represents the combined effects of various neural, muscular, and motional smoothing characteristics prohibiting instantaneous realization of the linguistic command. An example of analysis and determination of the target formant frequencies and the instant of onset of formant transition is shown in Fig. 2 for the word [ai] of Type A accent.

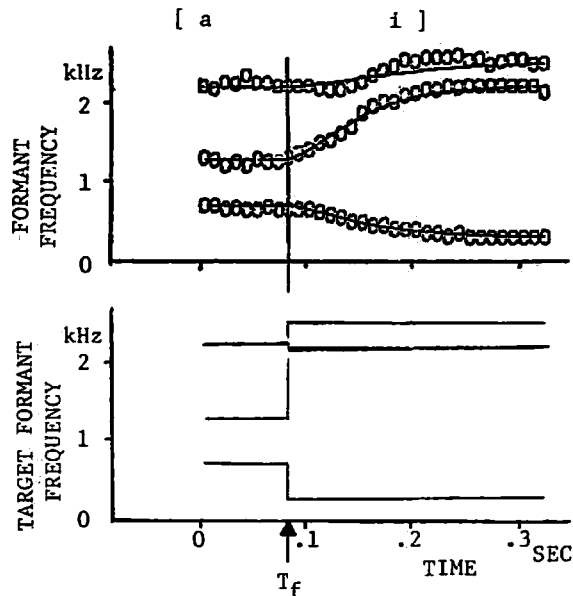


Fig. 2. Extraction of target formant frequencies and the instant of onset ( $T_f$ ) of formant transition by Analysis-by-Synthesis.

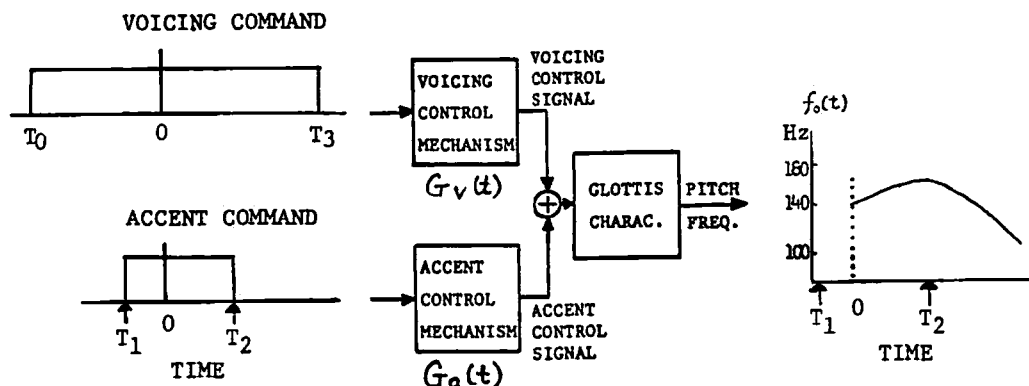
### 3.2. Extraction of suprasegmental features

The word prosody in a number of dialects of Japanese is essentially characterized by binary (high/low) patterns of subjective pitch generally associated with each mora, and its primary acoustic correlates are the systematic temporal variations in the fundamental frequency of the glottal source, i. e., the pitch contour, throughout an utterance. On the other hand, prosodic roles of stress and quantity, respectively manifested as the intensity and the duration of a particular segment relative to those of adjacent segments, are only secondary. For example, concurrent rises in pitch and intensity are generally observed at the accented morae of a word, but the role of a rise in intensity has been proved to be negligible in the perception of word accent types.<sup>6</sup> Extraction of the time-varying fundamental frequency, i. e., pitch extraction, is thus essential as the first step in the analysis of suprasegmental features of Japanese word accent.

The method of pitch extraction used in the present study was based on short-time autocorrelation analysis of the speech signal with a time window whose width was equal to a few fundamental periods.<sup>7</sup> A technique was introduced to suppress subsidiary peaks of the short-time autocorrelation function due to formant structure and to retain only those produced by the periodicity of the glottal source.<sup>8</sup> The value of this modified autocorrelation function at its greatest peak served as an index of signal periodicity, so that detection of periodicity and extraction of the fundamental period could be performed simultaneously by detecting the greatest peak exceeding a pre-determined threshold and by finding the time delay corresponding to the peak. Once the periodicity was detected and the short-time fundamental period extracted, the value was utilized to facilitate detection of the next period directly from the speech waveform. Fundamental periods were thus detected pitch-synchronously but were converted to the corresponding fundamental frequencies, which were further interpolated to produce a pitch contour uniformly sampled at intervals of 12.8 msec for the sake of further processing.

Because of various neural, muscular, and pneumatic factors which intervene between the intended linguistic units and their physical realizations, the above-mentioned binary pitch patterns never manifest themselves as such in the observed pitch contour, which is characterized by a rather smooth rise and decay at the accented morae, superposed on a base line that initially rises and then gradually decays toward the end of an utterance. Quantitative interpretation of suprasegmental features manifested by such a smoothed pitch contour requires a model of the processes that mediate between the linguistic information and its physical realization.

Such a model has been proposed by one of the authors<sup>9</sup> and has proved to be capable of closely approximating observed pitch contours of both isolated words and sentences.<sup>10, 11</sup> The basic concept of the model for the word pitch contour is illustrated in Fig. 3 together with mathematical formulations of its characteristics. Linguistic factors of voicing and accent are both assumed to take the form of stepwise binary commands to the control mechanism of the fundamental frequency. Commands for voicing and accent are smoothed separately by the low-pass characteristics of their respective control mechanisms, each being approximated by a critically damped second-order linear system, and their outputs are combined to



Fundamental frequency :

$$f_0(t) = f_{\min} \exp \{ G_v(t - T_0) + G_a(t - T_1) - G_a(t - T_2) - G_v(t - T_3) \},$$

Step response to voicing command :  $G_v(t) = [A_v dt \exp(-dt)] u(t)$ ,

Step response to accent command :  $G_a(t) = [A_a \{ 1 - (1 + \beta t) \exp(-\beta t) \}] u(t)$ ,

$T_0$  : instant of onset of voicing command,

$T_1$  : instant of onset of accent command,

$T_2$  : instant of offset of accent command,

$T_3$  : instant of offset of voicing command.

Fig. 3. Formulation of control processes of voice fundamental frequency: conversion of voicing and accent commands into a pitch contour.

control the fundamental frequency of glottal oscillations through a nonlinear mechanism. The basic structure and its functional formulation have been validated quantitatively by the analysis of word pitch contours found in the Tokyo and the Osaka dialects.<sup>1</sup>

Given this model for the synthesis of word pitch contours, it is possible, again by Analysis-by-Synthesis, to extract not only parameters that quantitatively characterize the underlying phonatory control, i.e., the timing of voicing and accent commands, but also both amplitude and response rate of the two linear systems producing the respective control signals, and parameters of the nonlinear control characteristics of the fundamental frequency.<sup>10</sup> The timing of the accentual command has been found to be crucial in characterizing various word accent types.<sup>12</sup> An example of Analysis-by-Synthesis of observed pitch contours and extraction of the voicing and the accent commands is shown in Fig. 4 for the same utterance of the word [ai] (Type A) as in Fig. 2.

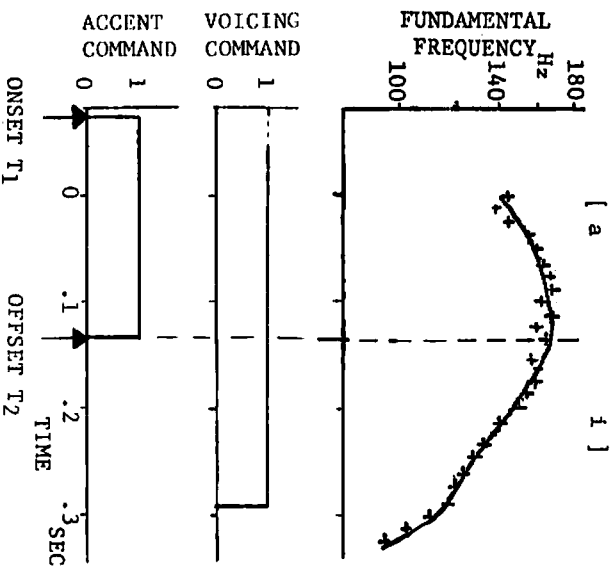


Fig. 4. Extraction of instants of onset and offset of voicing and accent commands by Analysis-by-Synthesis of pitch contour.

### 3. 3. Temporal relationships between articulatory and phonatory controls

The methods for analyzing segmental and suprasegmental features and for extracting the respective underlying controls can be combined to investigate the temporal relationships in the realization of these features. An example of such analyses is shown in Fig. 5 for the word [ai] of Type A, where the result for the pitch contour and that for the formant trajectories are shown respectively in the upper and the lower figures. The origin of the time axis is set at the onset of the utterance. In this case, the phonatory control for the word accent Type A is characterized by a negative value of the onset  $T_1$  of the accent command (not shown in the figure) and a positive value of its offset  $T_2$ , i. e., the onset of downward pitch transition. In the lower figure, the articulatory control for the transition from the initial vowel [a] to the final vowel [i] is characterized by the onset  $T_f$  of the command for the latter. Especially to be noted here is the fact that the onset of phonatory control lags behind that of articulatory control by approximately 45 msec.



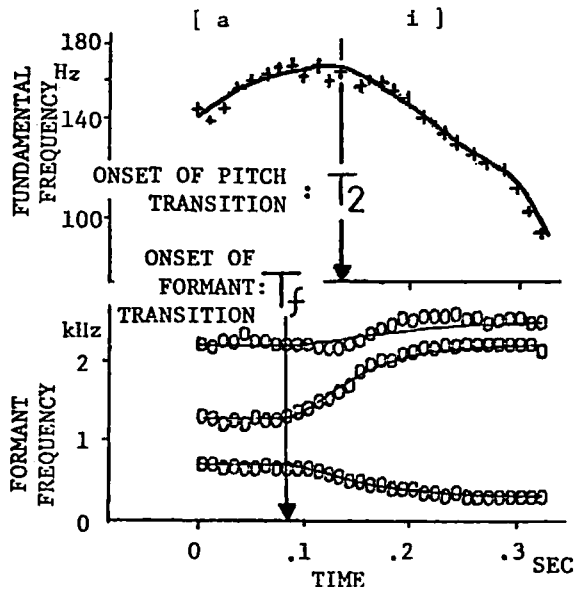


Fig. 5. Comparison of formant and pitch patterns and their Analysis-by-Synthesis for the word accent type A of [ai].

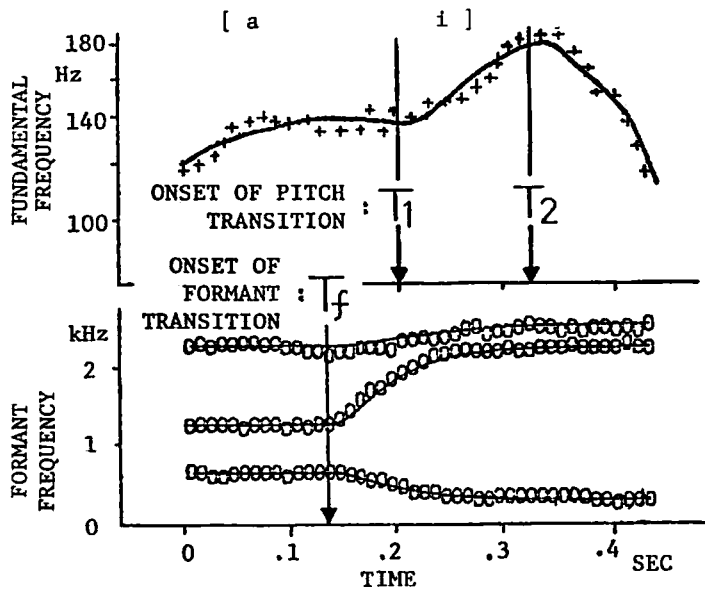


Fig. 6. Comparison of formant and pitch patterns and their Analysis-by-Synthesis for the word accent type B of [ai].

Figure 6 shows another example of analyses for the word [ai] of Type B, where the pitch contour is primarily characterized by a positive value of  $T_1$  approximately at the beginning of the second mora, followed by the offset  $T_2$  initiating the downward pitch transition within the same mora. The temporal order of onset of articulatory and phonatory controls is the same as in Type A, but the time lag is as large as 70 msec.

Table 2. Parameters of formant and pitch patterns of word accent type A and type B obtained by Analysis-by-Synthesis.

Type of Accent	A	B
Onset of formant transition, $T_f$ (msec)	89	131
Rate of formant transition, $\gamma$ ( $\text{sec}^{-1}$ )	25	24
Onset of accent command, $T_1$ (msec)	-75	195
Offset of accent command, $T_2$ (msec)	143	317
Rate of pitch transition, $\beta$ ( $\text{sec}^{-1}$ )	22	22

Table 2 lists mean values of parameters characterizing the articulatory and phonatory controls, averaged over four utterance samples each of Type A and Type B of [ai]. The mean values of delay between the onsets of articulatory and phonatory controls in Type A and Type B are 54 msec and 64 msec, respectively. Analysis of variance indicates that the difference between instants of onset for the articulatory and the phonatory controls is highly significant in both accent types. The difference is too large to be accounted for by any characteristics that might be involved in the phonatory process itself, which suggests that the two control processes are not exactly simultaneous, but are brought into an approximate synchronism by a certain mediating process, which detects consequences of the articulatory control and thereby initiates the phonatory control. Table 2 also indicates that the time constants of formant and pitch transitions, once they are initiated, are roughly equal, at least as far as the present data are concerned.

#### IV. Perceptual Segment Boundary and Onset of Phonatory Control

While the foregoing analysis has indicated that the articulatory control for production of the second vowel in a two-vowel sequence is initiated well ahead of the phonatory control for the second mora, the resulting changes in segmental features are only gradually realized, and hence a certain time is necessary before they are large enough to elicit perception of the second vowel. It is therefore possible that perception of segmental features of the second vowel in one's own utterance might initiate the phonatory control, and thus play a major role in maintaining an approximate synchronism between segmental and suprasegmental features.

In order to examine the possible role of perception in coordinating the articulatory and the phonatory controls, two sets of experiments were designed and conducted to determine the instant of perceptual onset of the second vowel constituting the second mora of the utterances under study. The first set of experiments utilized truncated natural utterances to determine the perceptual segment boundary between the first and the second vowels. The stimuli were prepared from a typical utterance of each of the words [ai] of Type A and Type B. Sixteen points were selected at intervals of 10 msec to cover the range of formant transitions. Two series of stimuli, each consisting of 16 tokens, were generated by truncating the same utterance at these points: one consisting only of "heads" (H-series), the other consisting only of "tails" (T-series) of the same utterance.

The stimuli in each series were arranged both in the ascending and in the descending order and were presented to subjects through a loud-speaker at intervals of 4 sec. The subjects were three adult males with normal hearing, who were instructed to answer, by forced judgment, whether they identified the stimulus as a single phoneme (i. e., /a/ in the case of H-series and /i/ in the case of T-series) or as two phonemes. Results obtained from the three subjects were pooled and plotted on a normal paper, and the point of 50% judgment was defined as the perceptual segment boundary. Boundaries obtained for the two series of stimuli were further averaged to eliminate possible response bias. The tests were conducted both for Type A and Type B of [ai]. Examples of the stimuli and subjects' responses are illustrated in Fig. 7.

The second set of experiments was also conducted in order to locate the perceptual segment boundary between the two vowels, but by inserting a click in the same natural utterance as used in the first set of experiments. Thirty points were selected at intervals of 10 msec to cover approximately the entire utterance, and a total of 30 stimuli were generated by superposing a click on the same natural utterance at each one of these points. These stimuli were arranged both in the ascending and in the descending order and were presented, at intervals of 4 sec, to the same subjects as in the first set of experiments, who were instructed to answer, by forced judgment, whether they heard the click in the first vowel or in the second. Results of the three subjects were again pooled, and the probability of perceiving the click in one of the two vowels, say [i], was plotted on a normal paper, the point of 50% probability being determined as the perceptual segment boundary. The click location tests were conducted both for Type A and Type B of [ai]. Figure 8 shows one example each of the stimuli and of the subjects' responses in the click location tests. The perceptual segment boundary obtained by click location generally tended to occur slightly

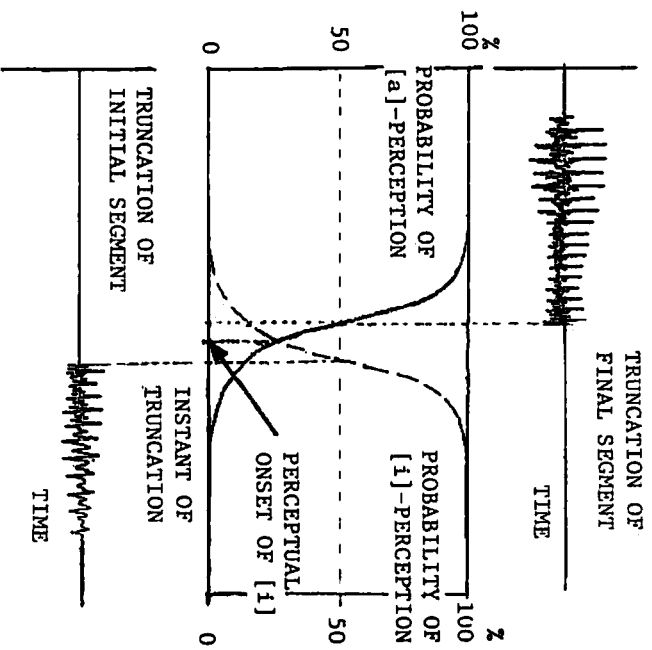


Fig. 7. Determination of perceptual segment boundary in [ai] by waveform truncation.

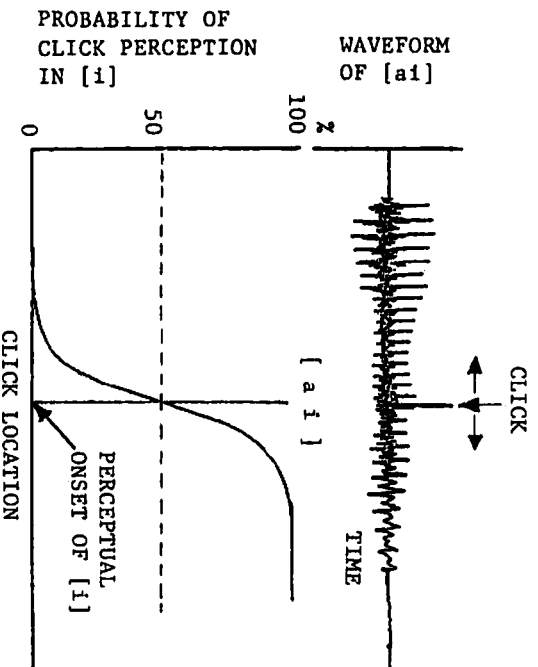


Fig. 8. Determination of perceptual segment boundary in [ai] by click location.

earlier in time than that obtained by waveform truncation.

The perceptual onsets of the second vowel, obtained by the two sets of experiments, are compared with the estimated onset or offset of the phonatory control characterizing the respective accent types, and are plotted in Fig. 9 against the estimated onset of the articulatory control for the second vowel of the same utterance. Although the perceptual experiments are still preliminary, and the amount of data is not sufficient to confirm the exact temporal order of the perceptual segment boundary and the estimated instant of initiation of the phonatory control, the results clearly indicate their approximate concurrence regardless of accent types, viz., regardless of direction of pitch transition and of duration of the interval between onset of the first vowel and that of articulatory control for the second vowel. These results strongly suggest the role of perception as the mediating process between the articulatory and the phonatory controls.

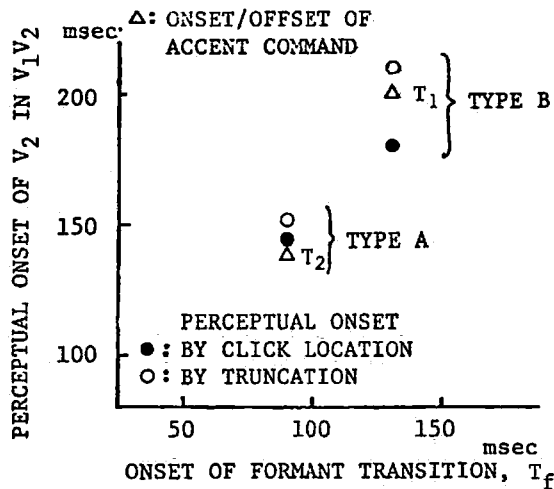


Fig. 9. Perceptual onset of [i] in [ai] vs. onset of formant transition,  $T_f$ .

## V. Conclusions

Temporal organization of segmental and suprasegmental features of speech in the word accent of Japanese has been investigated for the purpose of elucidating the underlying mechanism that maintains their synchronism. Methods for accurate extraction of these features as well as for estimation of crucial temporal parameters characterizing the underlying articulatory and phonatory controls have been described and applied to the analysis of some two-mora words of the Osaka dialect: words consisting of a sequence of two vowels. The results disproved exact simultaneity of these features and revealed the existence of a systematic delay in the onset of the phonatory control relative to that of the articulatory control, thereby suggesting the existence of a process mediating between the articulatory and the phonatory controls.

The role of perception as a possible mediating process has been examined by two sets of perceptual experiments designed to determine the perceptual boundary between the first and the second vowels. The perceptual segment boundaries, obtained both by identification of truncated utterances and by location of a click superposed on an utterance, indicated their approximate concurrence with the instants of initiation of the phonatory control estimated from the suprasegmental features, thereby suggesting that the phonatory control is initiated upon perception of acoustic consequences of the articulatory control.

#### References

1. Fujisaki, H., Y. Mitsui and M. Sugito (1974), "Analysis, Synthesis, and Perception of Accent Types of Two-Mora Words in Tokyo and Osaka Dialects," Transactions of the Committee on Speech Research, Acoust. Soc. Japan, 573-51.
2. Fant, G. (1960), Acoustic Theory of Speech Production, Mouton & Co., 's-Gravenhage.
3. David, E. E., Jr. and H. S. McDonald (1956), "Note on Pitch Synchronous Processing of Speech," J. Acoust. Soc. Am., 28, 586-589.
4. Bell, C. G., H. Fujisaki, et. al. (1961), "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Am. 33, 1725-1736.
5. Fujisaki, H., M. Yoshida et al. (1973), "Automatic Recognition of Connected Vowels Using a Functional Model of the Coarticulatory Process," J. Acoust. Soc. Japan, 29, 636-637.
6. Sugito, M., H. Fujisaki, and H. Morikawa (1974), "Characteristics of Accent Types and Their Perception," Transactions of the Committee on Speech Research, Acoust. Soc. Japan, 574-15.
7. Fujisaki, H. (1960), "Automatic Extraction of Fundamental Period of Speech by Autocorrelation Analysis and Peak Detection," J. Acoust. Soc. Am., 32, 1518.
8. Fujisaki, H. and Y. Tanabe (1973), "A Time-Domain Technique for Pitch Extraction of Speech," J. Acoust. Soc. Japan, 29, 418-419.
9. Fujisaki, H. and S. Nagashima (1969), "A Model for the Synthesis of Pitch Contours of Connected Speech," Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, 28, 53-60.
10. Fujisaki, H. and H. Sudo (1970), "Models for the Word and Sentence Pitch Contours of Japanese," Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, 29, 215-221.
11. 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.
12. Fujisaki, H., H. Hirose and M. Sugito (1975), "Analysis, Synthesis, and Perception of Word Accent Types of Japanese," Paper 99 presented at the 9th International Congress of Phonetic Sciences, Leeds, August 17-23.
13. 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.