Ann. Bull. RILP, No. 11, 1-10 (1977)

SPEECH INFORMATION PROCESSING BY MAN*

Hiroya Fujisaki

1. Introduction[1]

The human abilities of thought and communication owe very much to the use of language as a unique system of codes by which the information is expressed as messages. Communication by language, however, is achieved only when the messages are converted into physical signals, i.e., either into speech as their acoustic manifestations or into characters as their optical manifestations. The primary importance of the spoken language as compared to the written language is apparent from its exclusive use by many of the primitive cultures as well as from the order of acquisition by infants.

The reason for its primary importance resides partly in its simplicity as a means of communication; it requires only the respiratory air flow for the source of energy, nothing but a part of our body as the instrument for production, nothing but the ambient atmosphere for transmission, and requires very little effort both from the speaker and from the listener. And yet, speech can provide by far the highest rate of information emission among all the output modes available to man. The rate of 40 bit/sec is achieved by simultaneous control of several articulators, each of which can operate only at much slower rates. Every normal human being beyond infancy establishes, through continual training in his daily life, his ability as an expert both in emission and in reception of speech.

On the other hand, communication through speech by itself suffers from both spatial and temporal restrictions in that the signal can be received only at the immediate vicinity of its sender and at the instant of its generation. It is thus quite natural, that a large amount of scientific effort has been expended for the alleviation of these restrictions. These efforts have been first materialized by the invention of telephone and gramophone exactly one century ago, and the techniques have since developed to such an extent that the use of trans-oceanic cables and communication satellites allows almost instantaneous telephone conversation without restrictions on the distance, and the use of magnetic recording permits storage and reproduction of spoken messages at any desired instant. Thus it may be said that the spatial and the temporal restrictions inherent to speech communication have now been almost completely eliminated by these techniques.

For the purpose of man-to-man communication for which these techniques have been developed, accurate transmission of the speech signal itself was sufficient. For the purpose of man-machine speech communication, however, analysis of the relationship between the speech signal and the underlying linguistic code is mandatory. Although the past three decades have seen a great progress both in techniques of speech synthesis and of automatic speech recognition, we still seem to lack fundamental knowledge concerning mutual conversion between speech and language. A thorough

---

* Invited lecture given at the IX International Congress on Acoustics, Madrid, July, 1977

understanding of the human processes of speech production and perception seems to be a prerequisite for the realization of true man-machine speech communication, which I regard as one of the major goals of acoustics for the coming century. From this point of view I shall describe some of our recent studies on speech production and perception, in order to clarify the unsolved problems and to suggest possible ways of approach, rather than to present an overall state of the art of man-machine speech communication.

2. The Speech Communication Process

The essential function of a language as a coding system is quantization in the broadest sense, by which continuous or discrete information is represented in terms of discrete units for the sake of reliable communication. These discrete units, however, undergo mutual interaction and smoothing, and are superposed with various factors of both systematic and statistical nature in the process of their actualization as speech, and are observed as a continuous acoustic signal. As shown in Figure 1, which illustrates major steps in generation, transmission, and reception of speech in man-to-man communication, the phoneme strings organized as codes at the center of speech are converted into neuro-motor commands at the motor speech area, and are sent out in parallel to various phonatory and articulatory organs such as vocal chords, mandible, tongue body and tongue tip, lips as well as velum. These commands undergo several stages of mutual interaction and smoothing during neural transmission and muscular contraction, and cause individual or concurrent inertial motions of the respective organs. The resultant deformation, both global and local, of the vocal tract in turn affect the acoustic characteristics of the speech signal in such a way that successive phonemes heavily overlap with each other and generally leave no trace of discreteness in the speech signal itself or in its acoustic characteristics. Analysis and formulation of dynamic characteristics of the articulatory and the phonatory systems are thus seen to be crucial for the quantitative understanding of the conversion from the linguistic code into the speech signal. [2]

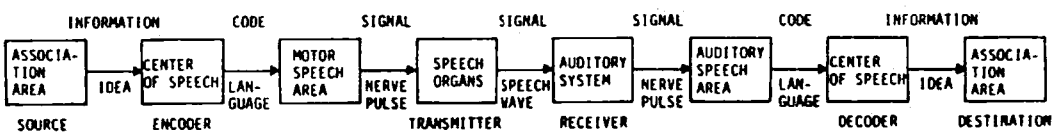

Fig. 1. Steps in generation, transmission and reception of information through speech.

On the other hand, the speech signal as pressure fluctuations on the eardrum of the listener undergoes various stages of acoustic, mechanical, electrical and neural transduction. The resulting neural signal is then processed by the auditory nervous system, restored as the linguistic code at the auditory speech area, and is eventually decoded at the center of speech. Although recent developments in the auditory neurophysiology are truly impressive, the complexity of the entire neural mechanism, responsible for converting speech signals into linguistic codes, still seems to defy

quantitative analysis of the total process. Alternatively, the whole process can be approached by methods of psychoacoustics, which allow one to quantify the input-output characteristics of the listener. The use of synthetic speech, initiated more than a quarter of a century ago, [3] still remains to be quite powerful in finding and evaluating the cues that play important roles in speech perception. The major problem here is the formulation of the perceptual process by which speech signals in dynamic context is converted into discrete units. [4]

3. Speech Production — From Code to Signal

3.1. Formulation of the Coarticulatory Process [5-7]

In view of our present lack of knowledge concerning dynamic characterisiics of some of the physiological and physical stages that intervene between the linguistic code and the speech signal, quantitative formulation of the coarticulation process is almost impossible if we try to build a structural model. Alternatively, the whole process can be formulated on the basis of a functional model, i. e., by approximating characteristics of the whole system by a transfer function from the linguistic code to the observed acoustic properties of the speech signal. Such a model has been shown at first to be applicable to connected vowels, but has since been shown to apply also for other classes of speech sounds, i. e., for semivowels, liquids, and voiced stops.

The basic idea of the model is illustrated by Fig.2 for connected vowels. Each vowel phoneme is assumed to possess a set of target formant frequencies, and the whole process of coarticulation is functionally approximated by a hypothetical linear system which converts the target formant frequencies into observed formant trajectories. It has been shown that formant transitions between a pair of front vowels as well as between a pair of back vowels can be approximated quite closely by step response functions of a critically-damped second order system. In an utterance of a sequence of

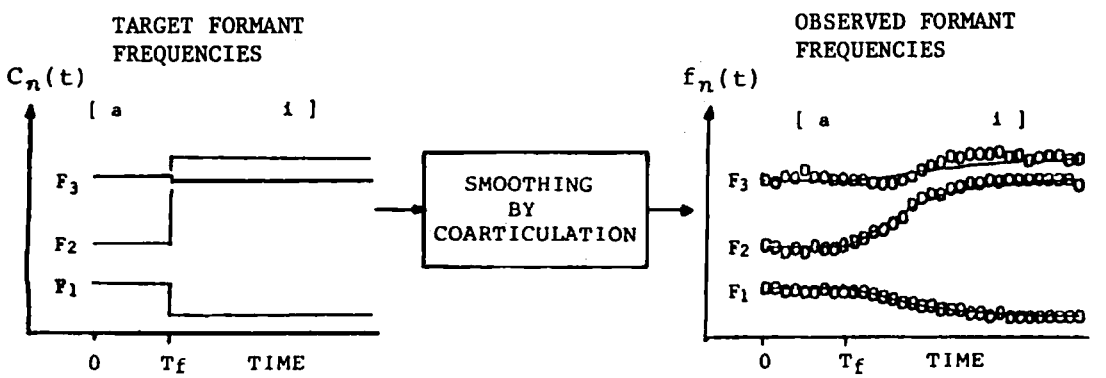

Fig. 2. Formulation of the process of coarticulation in the formant frequency domain.

m vowels, the trajectory of the nth formant frequency, $f_n(t)$, is given by

$$f_n(t) = F_{n,1} + \sum_{j=2}^{m} (F_{n,j} - F_{n,j-1}) [1 - \{1 + \gamma_{n,j}(t-\tau_j)\} \exp\{-\gamma_{n,j}(t-\tau_j)\}] u(t-\tau_j),$$

where $F_{n,j}$ and τ_j respectively denote the target value of the nth formant frequency and the time of onset of the command for the jth vowel. For formant transitions between front and back vowels, continuity of resonance modes requires reversal of formant numbering, but crossing of formant trajectories can be avoided by taking into account the coupling between two resonance modes.

The validity of such a functional formulation has to be tested by its capability of approximating formant trajectories actually observed. The formant trajectories shown on the right-hand side of Fig. 2 illustrate the degree of approximation obtained by the model, which immediately suggests the model's applicability to speech synthesis. When a set of formant trajectory is given, it is also possible to estimate, by the method of Analysis-by-Synthesis, the target formant frequencies and the onset of command for each vowel in the sequence. These parameters have been applied to segmentation and recognition of connected vowels, semivowels, and voiced stops.

3.2 Formulation of Control Process of Voice Fundamental Frequency[8-10]

Realization of suprasegmental features also involves processes of concatenation and smoothing due to various neural, muscular and pneumatic factors, and the present state of our knowledge concerning these factors are quite similar to that for the coarticulation process. A functional model has thus been presented to approximate the whole process of converting the linguistic code tinto contour of the voice fundamental frequency, which is a major acoustic correlate of the word accent in many dialects of spoken Japanese. In these dialects, all content words as well as some function words are associated with binary patterns of subjective pitch, which serve to eliminate certain lexical and syntactic ambiguities. These binary patterns of subjective pitch, however, never manifest as such in the contour or the fundamental frequency. The latter is invariably 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.

The basic idea of the model is illustratee by Fig. 3. Linguistic factors of voicing and accent are both assumed to be stepwise binary commands to the control mechanism of the fundamental frequency. These commands are smoothed separately by the low-pass characteristics of their respective control mechanism, each being approximated by a critically-damped second-order linear system, and their outputs are combined to control the fundamental frequency of glottal oscillations through a nonlinear mechanism. Thus the fundamental frequency $f_0(t)$ is given by the following equation:

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

where

$$G_v(t) = [A_v \alpha t \exp(-\alpha t)]u(t),$$

$$G_a(t) = [A_a\{1-(1+\beta t)\exp(-\beta t)\}]u(t),$$

and

T_0 : onset of voicing command,
T_1 : onset of accent command,
T_2 : offset of accent command,
T_3 : offset of voicing command.

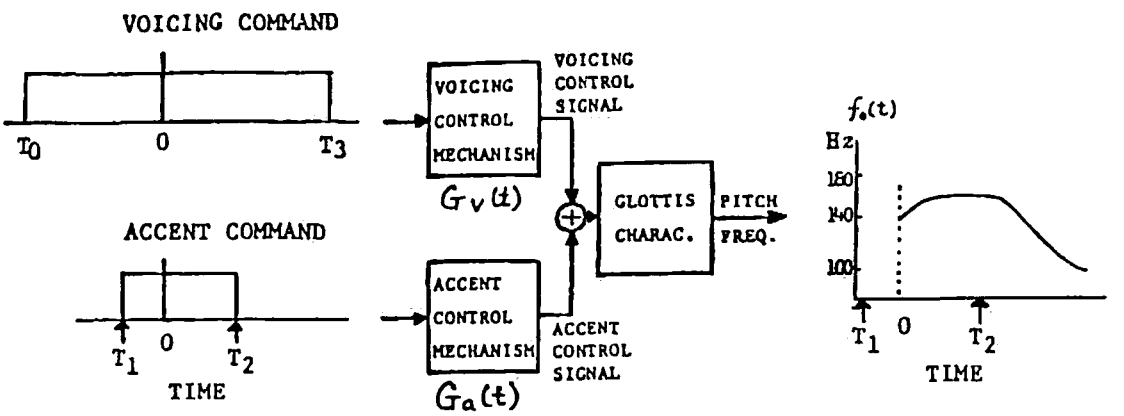

Fig. 3. Formulation of control processes of voice fundamental frequency.

The validity of the basic structure and its functional formulation have been tested by analysis of fundamental frequency contours of words both in the Tokyo and Osaka dialects. For a given fundamental frequency contour, it is possible to extract, also by the method of Analysis-by-Synthesis, parameters that characterize the underlying phonatory control, i.e., the timing of voicing and accent commands, as well as those which characterize individual phonatory mechanism.

4. Speech Perception — From Signal to Code

4.1 Perception of Non-stationary Vowels

Although most vowels can be produced in isolation and can be sustained by static articulation, they are invariably produced by dynamic articulation in connected speech, so that their acoustic characteristics are seldom stationary, and are different from those of their static counterparts even at points where they seem to be most stationary. On the other hand, such variabilities seem to be removed and the invariance of vowel is maintained by the perceptual process. Although several investigations have

studied this phenomenon of perceptual nomalization, it has not been made clear whether such normalization occurs only for speech sounds or for non-speech sounds as well. Perceptual experiments have therefore been conducted using synthetic stimuli with formant transitions closely approximating those in natural utterances of two-vowel sequences as well as CV-syllables. These synthetic syllables were adopted since they constitute minimal perceptual units of connected speech in Japanese.

Figure 4 shows an example of the first formant transition of stimuli used for perceptual study of two-vowel sequences /$V_1 V_2$/, where the vowel /u/ was selected as V_1 and the target value for V_2 was selected on the /u/-/a/ continuum. All other formant frequencies were held constant. The transition was truncated at points where 95, 80, or 70% of the total excursion was attained, producting three different dynamic stimuli. Each of these dynamic stimuli was paired with one of seven static stimuli for paired comparison, and the point of subjective equality was determined on the frequency scale of static stimuli. A similar test was also conducted using non-speech stimuli which consisted only of the first formant.

Results of these tests are shown in Fig. 5, where the abscissa shows the terminal frequency of the truncated formant transition in the dynamic stimuli and the ordinate shows the point of subjective equality on the formant frequency scale of the static stimuli. The solid and broken lines respectively indicate the averaged results of three subjects for speech and non-speech stimuli. These results show that formant transitions, when truncated at 70 or 80% of the total excursion, are perceptually extrapolated both in speech and in non-speech stimuli, though the extrapolation does not overshoot the original target of the transition.

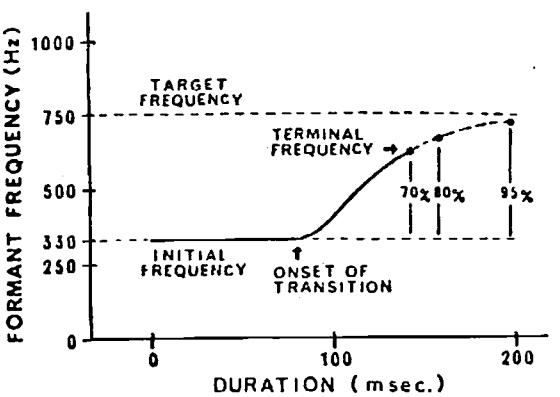

Fig. 4. Formant transition of dynamic speech and non-speech stimuli, approximated by step response of a critically-damped second-order linear system.

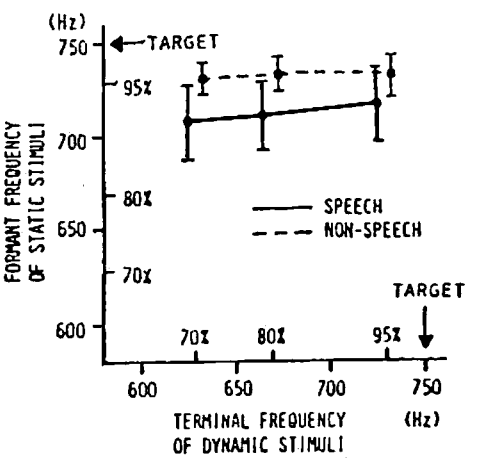

Fig. 5. Subjective equality between static and dynamic stimuli in the region of typical vowel /a/.

4.2 Perception of Segmental Duration in Context.

While in many languages sound duration carries primarily prosodic information, there exist certain languages, including Japanese, in which it carries segmental information, i.e., distinction between short/long vowels, single/geminate consonants, etc. In this respect, the sound pattern of Japanese presents examples of considerable interest since all the vowels, as well as nasal consonants and voiceless consonants in intervocalic positions, possess their longer counterparts that can be discriminated primarily by duration. Perception of these sound involves varieties of spectral features of the interval in question; i.e., periodic in vowels and nasal consonants, aperiodic in voiceless fricatives, and nil in voiceless plosives and affricates. They are also peculiar in that the criteria for their identification may be highly adaptive to the tempo of their immediate context.

In order to investigate identification of these speech sounds, synthetic words containing either one of the four categories were generated by a digital computer, producing stimuli along the four continua: [oi] – [o:i], [ama] – [amma], [ise] – [isse], and [ita] – [itta], each constituting a minimal pair of meaningful words. For the sake of comparison, non-speech sounds with characteristics somewhat similar to those speech segments in question, i.e., 500 Hz tone burst, white noise, band-pass noise, and pause between two tone bursts were also generated and used as stimuli in discrimination tests of duration.

Accuracies of discrimination of filled and empty non-speech intervals were found to be significantly different as seen from Table 1, which lists averaged results of five subjects. Identification tests of speech segments in word and sentence contexts, on the other hand, showed marked uniformity of judgment criteria as well as accuracies of identification for vowels, nasals and voiceless plosives, as shown in Table 2. These results suggest

Table 1. Accuracy of discrimination for duration of various non-speech stimuli.

Stimuli	500 Hz tones					white noise	filtered noise	pause between tones
	50 msec	100 msec	150 msec	200 msec	300 msec	100 msec	100 msec	100 msec
Accuracy of discrimination	7.6 msec	9.6	11.6	14.5	23.1	9.1	6.7	21.5
Standard deviation	1.7 msec	1.5	1.9	1.3	1.1	1.9	0.86	4.9

Table 2. Phoneme boundaries and accuracy of identification for various synthetic speech.

Stimuli	vowel /oi/-/ooi/		fricative /ise/-/isse/		plosive /ita/-/itta/		nasal /ama/-/amma/	
Context	word	sentence	word	sentence	word	sentence	word	sentence
Phoneme boundary	156 msec	168	166	165	169	164	141	152
Accuracy of identification	9.5 msec	7.1	16	10	11	8.9	10	8.5

that the durational cue of these speech stimuli are processed by the same mechanism in spite of the differences in their acoustical characteristics. Figure 6 shows results of further identification tests at various speech rates of the immediate context, being just one vowel in the case of word, and a 5-mora phrase in the case of sentence context, and clearly indicates the short-term adaptivity of the judgment criterion.

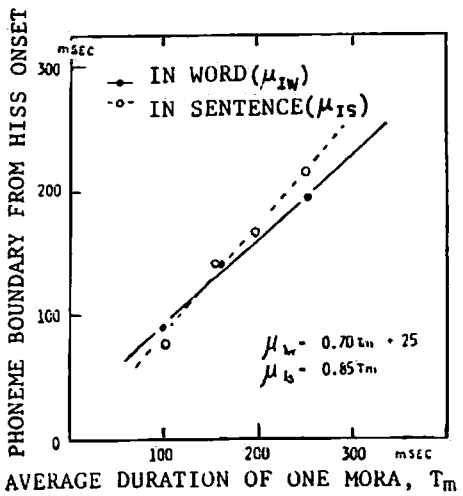

Fig. 6. The influence of speech rate on the identification of a geminate fricative consonant in word and sentence context.

4.3 Perception of Word Accent Types[10]

As already mentioned in 3.2, the fundamental frequency contours of Japanese word accent types are characterized mainly by the timing parameters of the accent command. Figure 7 illustrates analysis of four word accent types of two-mora [ame], where '+' symbols indicate measured fundamental frequencies, the solid curves indicate their best approximations based on the proposed model, and the stepwise waveform indicate the timing of the extracted accent command. The close agreement between the measured and the predicted contours of the fundamental frequency suggests that the timing parameters of the accent component is, not only acoustically, but also perceptually, important. This was verified by identification tests of word accent types, in which 40 synthetic versions of [ame] were used including typical stimuli of the four types, indicated by stimuli No. 1, 11, 21, and 31 in Fig. 8 on the $T_1 - T_2$ plane, as well as intermediate ones selected on the four sides of the quadrangle. The stability of categorical judgment of word accent types is indicated by the small range of variations in judgment criteria on the four continua of Fig. 8.

On the other hand, these judgment criteria must necessarily be adaptive to the tempo of the segmental features in order that word accent types should be correctly recognized at various speech rates. This was verified by further identification tests in which durations of various segments were systematically controlled and their effects on specific category boundaries were quantified. The results of these experiments demonstrated that identification of a specific accent type may be based predominantly upon the temporal relationship between the accent command and a specific phonemic segment within the word.

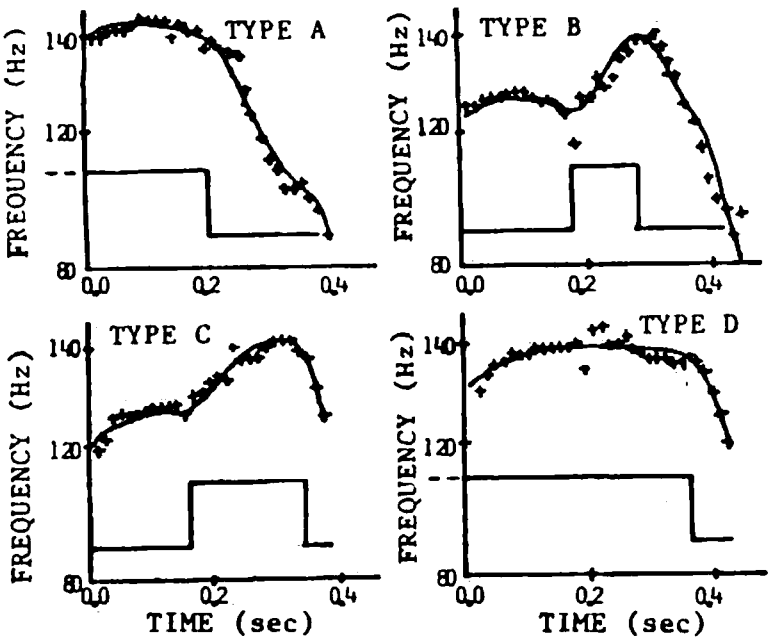

Fig. 7. Analysis-by-Synthesis of fundamental frequency contours in four accent types of two-mora [ame].

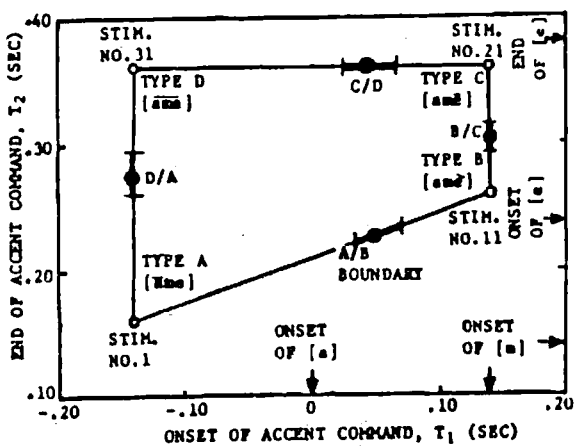

Fig. 8. Category boundaries and their variations between word accent types of [ame].

5. Conclusion

Processes of speech production and perception, regarded as interconversion between the linguistic code and the speech signal, have been reviewed. The studies described here are by no means meant to be representative of all the important works recently accomplished in the field of speech communication, but are meant merely to reveal some of the crucial points and the unsolved problems, and to suggest possible ways of approach. It is the belief of the author that a deeper understanding of the human processes of speech production and perception will lead to a better man-

machine communication through speech, which is by far the most important means of human communication.

References

1. H. Fujisaki (1972); Current Problems in Speech Recognition, J. Acoust. Soc. Japan, 28, 33-41.
2. Preprints of U.S.-Japan Joint Seminar on Dynamic Aspects of Speech Production, December 7-10, 1976.
3. F.S. Cooper, P.C. Delattre, A. M. Liberman, J. M. Borst and L. J. Gerstman (1952); Some Experiments on the Perception of Synthetic Speech Sounds, J. Acoust. Soc. Am., 24, 597-606.
4. A. Cohen and S. Nooteboom, eds., (1975); Structure and Process in Speech Perception, Springer-Verlag, Berlin
5. H. Fujisaki, 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.
6. H. Fujisaki, Y. Sato, Y. Noguchi and T. Yamkura (1975); Automatic Recognition of Semivowel in Spoken Words, J. Acoust. Soc. Japan, 31, 696-697.
7. Y. Noguchi and H. Fujisaki (1976); Classification of Voiced Stop Consonants Using Features Extracted on the Basis of a Model of Coarticulation, Records of Spring Meeting, Acoust. Soc. Japan, 339-340.
8. H. Fujisaki and H. Sudo (1971); A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent, J. Acoust. Soc. Japan, 27, 445-453.
9. H. Fujisaki and H. Sudo (1972); A Generative Model for the Prosody of Connected Speech in Japanese, Conference Record, 1972 Conference on Speech Communication and Processing, IEEE-AFCRL, 140-143.
10. H. Fujisaki and M. Sugito (1976); Acoustic and Perceptual Analysis of Two-Mora Word Accent Types in the Osaka Dialect, Ann. Bull. RILP, No. 10, 157-177.
11. H. Fujisaki and S. Sekimoto (1975); Perception of Time-varying Resonance Frequencies in Speech and Non-speech Stimuli, Structure and Process in Speech Perception, A. Cohen and S. Nooteboom, eds., Springer-Verlag, Berlin, 269-282.
12. H. Fujisaki, K. Nakamura and T. Imoto (1975); Auditory Perception of Duration of Speech and Non-speech Stimuli, Auditory Analysis and Perception of Speech, G. Fant and M. Tathâm, eds., Academic Press, London, 197-219.