The Automatic Speech Recognition System for Conversational Sound

Sakai, Toshiyuki; Doshita, Shuji

STUDIA PHONOLOGICA III (1963)

The Automatic Speech Recognition System for Conversational Sound*

Toshiyuki SAKAI and Shuji DOSHITA

Kyoto University, Kyoto, Japan

Summary

This paper describes the method and the system investigated to solve the problem encountered in the automatic recognition of speech sound. Starting from research of the automatic analyzer of speech sound, a monosyllable recognition system was constructed in which the phoneme is used as the basic recognition unit. Recently this system has been developed to accept the conversational speech sound with unlimited vocabulary.

The mechanical recognition of conversational speech sound requires two basic operations. One is the segmentation of the continuous speech sound into several discrete intervals (or segments), each of which may be thought to correspond to a phoneme, and the other is the pattern recognition of such segments. For segmentation, by defining two criteria, "stability" and "distance", the properties of the time pattern obtained by the analysis of input speech sound may be examined. The principle of the recognition is based on the mechanism of the articulation in our speech organ. Correspoding to this, the machine has the functions called phoneme classification, vowel analysis and consonant analysis.

Conversational speech recognition system with the phonetic contextual approach is also applied to the vowel recognition where the time pattern of input speech is matched with the stored standard patterns in which the phonetic contextual effects are taken into consideration.

The time pattern which has the great variety may be effetively expressed by the new representation of "sequential pattern" and "weighting pattern".

1. Introduction

Realization of an automatic recognizer of the spoken sound has long been a desire of human beings. Some experimental models have been tried which accept

* The project was supported by the Japanese Ministry of Education, and the machine was constructed in Nippon Electric Co. Ltd. (NEC) in Tokyo, Japan.

Toshiyuki SAKAI (坂井利之) : P.H.D. Professor of Department of Electrical Engineering, Kyoto University.

Shuji DOSHITA (堂下修司) : Assistant of Department of Electrical Engineering, Kyoto University.

some limited sounds as the input. Speech is one of the important media of communication, but in the present speech communication system, it is transmitted keeping its original form without knowing its essential properties. However, for the mechanical recognition of speech sound it is necessary not only to make clear the properties of speech sound but also to find a realizable system or method for the extraction of these properties. The authors first tried to examine spoech sound properties by the electronis methods which are further developed for the recognition of speech sound.

In considering the recognition system, we intended to accept, all the vocabulary commonly used, not limited to a certain category of words. To satisfy this requirement the phoneme is selected as the basic recognition unit and the operation of the machine is controlled by the input speech sound itself.

Speech sound or the parameters that prescribe its properties are continuous quantities varying continuously with time whereas the letter or code is discrete in space and time. Therefore speech recognition is the partition of the continuous quantities to the discrete domains in time axis and space axis. In the partition, to avoid the difficulties by the effect of large variety of speech sound property from speaker to speaker, relative time scaling was adopted; for the partition of parameters, several methods were used in parallel, corresponding to our articulation mechanism.

The automatic speech recognition system based on this principle was first desgned to work with a monosyllable input and it is extended to accept the general conversational speech sonnds as described below.

2. Principle of Conversation Speech Recognition System

Fig. 1 shows the whole block diagram of the conversational speech recognition system. It is divided into two parts according to function: the segmentation part (I of Fig. 1) which separates the input speech sound into discrete sections and the recognition part (II of Fig. 1) which performs the discrimination of the separated sections.

2.1 Segmentation Part

The principal operations of the segmentation part are to distinguish the time points, which separate the speech sound into the sections corresponding to the recognition unit (phoneme), from the time pattern of parameters extracted from the input speech sound and thereby to control the operation of the recognition part, such as sampling, discrimination and output timing. In this part the "segmentation circuit of consonant and vowel" divides the speech sound into consonant sections and vowel sections. Further, when a vowel section contains moret han one recognition unit (phoneme), the "vowel segmentation circuit" divides it into segments each of which is thought to correspond to the vowel phoneme. For this

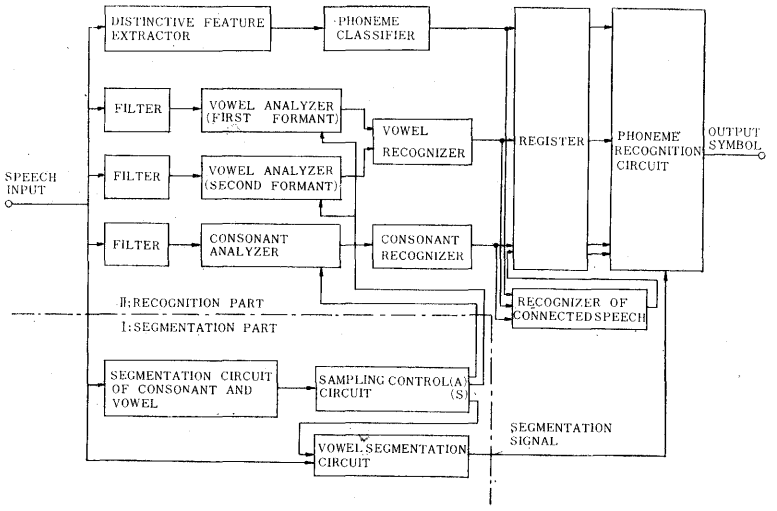


Fig. 1 Block diagram of speech recognition system.

operation of segmentation the pattern of time variation of speech parameters is exmined. This is based on the fact that speech sound is composed of such sections as; quasi-stable sections in which the parameters remain almost constant state, transient sections in which the parameters move gradually and sections or time points at which the parameters make abrupt change. The principle of the segmentation of consonant section and vowel section is like that of the phoneme classification of the recognition part.

Segmentation of vowel section is performed by extracting the parameters, that describe the aforementioned intervals, from the digital pattern of seech sound anayzed by the zero-crossing wave analysis. We defined two quantities called "distance" and "stability". Stability expresses the stationary property of pattern, that is, the property that the parameter remains almost constant state over a certain interval. On the contrary, distance is the quantity that expresses the change of the pattern. Stability is useful for discovering the vowel sound and the fricative sound and distance is for the sound with burst such as stop consonant.

In the operation of segmentation, the selection of the recognition unit is the basic problem. To accept conversational speech sound we selected the phoneme rather than the morpheme or the word, and the phonetic context was treated in another part of the machine.

2.2 Recognition Part

The recognition part recognizes the speech section, segmented by the signal of segmentation part, as the phoneme. In the phoneme classifier of Fig. 1 speech wave is classified to the several groups, each corresponding to the manner of articulation in our speech organ, by the distinctive features. In parallel with this operation vowel recognition system and consonant recognition system, each

of which is composed of filter, analyzer and recognizer, find the parameters corresponding to the place of articulation and recognize phomes classified as belonging to the same group by the phoneme classifier. The analyzing method is the zero-crossing wave analysis which is applied to the consonant section and vowel section separately, whose sampling is controlled by the sampling signal form the sampling control circuit. Sampling is performed once for one consonant section; for a vowel section it is periodically repeated.

As mentioned above we selected as recognition unit the phoneme and the phonetic interaction between phonemes which is essential to the conversational speech sound is treated in the recognizer of connected speech of Fig. 1. All the results from the previous stage of the machine are once stored in the register and are combined in the phoneme recognition circuit where final recognition of phoneme is made, controlled by the segmentation signal from the segmentation part. The output symbol is the Kana letter, which is the Japanese phonetic alphabet and also the orthography.

3. Segmentation to the Recognition Unit

3.1 Distance and Stability

As speech sound is an anlog and continuous signal, its recognition is after all the coding of the speech signal to the letter symbol. Further, it will be desirable to process the speech sound in digital from. We convvert the input speech wave into the digital pattern which is the time series of parameters by digitizing the analog quantities with appropriate unit and by sampling it with the time unit sufficient to maintain the characteristics of time variation of speech sound.

The information of speech sound wave is, according to the sampling theory, too large to treat directly. On the other hand it contains much redunduncy when the linguistic information of speech sound is considered. From the view point of time domain, typical sections of the speech sound are as follows; a) quasi-stable state in which parameters at each time point are closely related to each other and are repeated only with slight change, b) transitional sections in which parameters change is gradual except for some time point at which parameters change abruptly. Segmentation may be performed by paying attention to such time change characteristic of the parameters. For this purpose we defined two criteria "stability" and "distance" as follows:

On the time axis the time points 1, 2,$\cdots j$, $j+1,\cdots$ are selected and the interval between j-th and $j+1$-th time points is celled j-th sampling interval. We denote by P_{ij} the i-th element of parameters or distributions ($i=$, 2,$\cdots$, n) in the j-th sampling time interval normalized in each time interval. Then

$$P_j=\{P_{1j},\ P_{2j},\ \cdots\cdots,\ P_{nj}\}$$

is the parameter set or distribution in the j-th time intterval. The whole pattern of speech sound is expressed as follows;

$$P=\{P_j\}=P_1,\ P_2,\ \cdots\cdots,\ P_j,\ \cdots\cdots$$

We define the index of stability $X_{ij}(l)$ as

$$X_{ij}(l)=\frac{1}{l}\sum_{k=0}^{l-1}P_{ij-k}$$

where l is the number of sampling intervals. $X_{ij}(l)$ is defined in each time interval and in each channel and has the value 0∼1. This index represents the rate in the i-th channel during the l intervals before the j-th time interval. When this takes a large value it may be considered that the neighbouring part of the pattern belongs to one segment corresponding to a phoneme. Thus by selecting a proper value of l, $X_{ij}(l)$ gives important information for distinguishing the stationary part of pattern from the transient part which inevitably appears bewteen the stationary parts.

The distance d_j is defined by

$$d_j=\sum_i|P_{ij}-P_{ij-1}|$$

where the sign — may be replaced by ⊕ (exclusive OR) where P_{ij} is a variable of Boolean algebra and d_j is the Hamming distance. The distance is the quantity used to measure the magnitude of change of pattern and takes large values for the time intervals where the pattern makes an abrupt change and small value for stationary parts where the pattern keeps a nearly constant state.

3.2 *Segmentation of Successive Vowels by Zero-crossing Wave Analysis*

Experiment of segmentation using the zero-crossing pattern was applied to the vowel segmentation circit of Fig. 1. Fig. 2 shows the block diagram of this circuit. The zero-crossing distribution

$$\{W_{ij}\}\qquad i=1,\ 2,\ \cdots\cdots,\ n$$

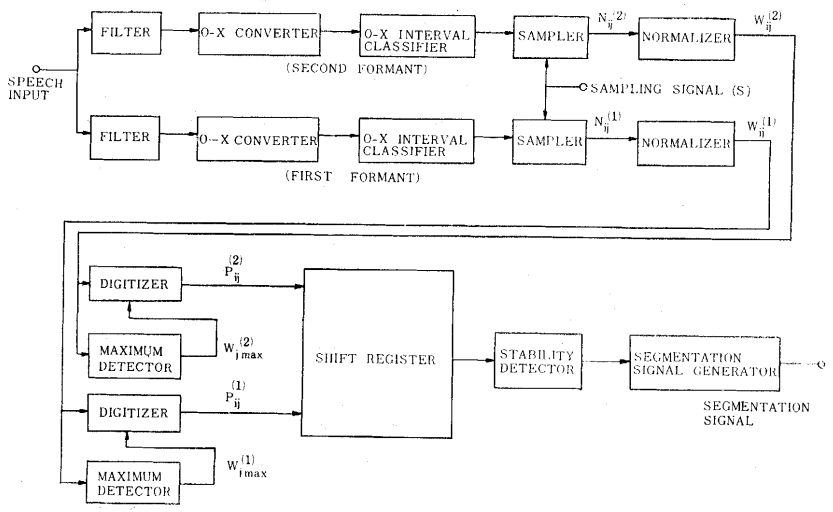


Fig. 2 Block diagram of vowel segmentation circuit.

is obtained in each of successive sampling intervals. Therefore the pattern is formed as the series of the zero-crossing distributions,

The method of zero-crossing wave analysis will be presented later. The anaysis of speech sound during, for example, 10 ms gives channel classified distribution $\{W_{ij}\}$. Then W_{ij} is digitized to the aforementioned P_{ij} by setting up a threshold level relative to the maximum value at that sampling interval. The processing of pattern for distance and stability is made in a shift register having the length of l. Though this method may be applied to the general speech sound, we applied this to the vowel section for the reasons that, because of the "consonant+vowel" construction of the Japanese syllable, the vowel has an important funtion and that the scale of hardware is too large to apply the method to the whole speech sound.

As the parameters of vowel, the first formant and the second formant are selected. Before the zero-crossing wave analysis, frequency regions of the first formant (F_1) and the second formant (F_2) are picked up by passing the input speech sound through the filters; a low pass filter of 1,500 c/s for the F_1 region and a band pass filter of 800~2,500 c/s for the F_2 region. (The first formant and the second formant are indicated by the suffix (1), (2), respectively, but since the operation of both regions are the same as is seen in Fig. 2, its suffix is often omited.) In Fig. 2 the zero-crossing wave converter (O-X converter) converts input speech wave into zero-crosssing wave, a series of rectangular waveforms in which only the time points at which the original speech sound crosses zero level are left as the information bearing parameter and the wave has constant levels of positive and negative according to the polarity of the original wave.

The zero-crossing wave analysis[1][2] is a measurment of the width of each rectanglar waveform, which is executed in the zero-crossing interval classifier (O-X classifier) of Fig. 2, and the classified number is integrated as a zero-crossing distribution for a certain sampling interval T. By repeating the sampling in a constant period, a zero-crossing pattern is obtained.

We denote by N_{ij} the number of occurences of zero-crossing width τ, which appear in the i-th channel of width $\Delta\tau_i$ and central value τ_i and in the j-th sampling interval of T. N_{ij} is the number of rectangular waves in the interval T which have the width between $\tau_i-\frac{\Delta\tau_i}{2}$ and $\tau_i+\frac{\Delta\tau_i}{2}$. Then a statistical expression of zero-crossing interval measurment is

$$W_{ij}=W_j(\tau_i)=\frac{1}{T}\frac{N_{ij}\tau_i}{\Delta\tau_i} \qquad \begin{matrix}(i=1,\ 2,\ \cdots\cdots,\ n)\\ (j=1,\ 2,\ \cdots\cdots)\end{matrix}$$

and the zero-crossing distribution in the j-th interval is

$$W_j=\{W_{1j},\ W_{2j},\ \cdots\cdots,\ W_{nj}\}$$

(suffix n may be n_1 for F_1 and n_2 for F_2.) By the multiplication of τ_i as is seen in the above expression, W_{ij} is the ratio of time interval, which is the summa-

tion of the width of the rectangular waves classified to the i-th channel in the j-th interval, to the total time T' and $\sum_{i=1}^{n} W_{ij}$ has almost constant value.

The circuit of zero-crossing wave analysis has the same construction as the vowel analyzer. Therefore as is shown in Fig. 6 a pair of the zero-crossing distributions in F_1 region and in F_2 region, each of which having n_1 and n_2 channels respectively, is obtained and they are treated separately.

The normalizer of Fig. 2 converts the N_{ij}, which is given as a pulse number, to the proportional analog voltage W_{ij} of the above equation. The zero-crossing pattern is closely related to the spectra of original speech sound and its peak corresponds to the formant. Therefore we can obtain simply and effectively the following method. The maximum detector picks the peak value $W_{j\max}$ of the n channel distribution in each time interval and the digitizer converts W_{ij} to the aforementioned P_{ij} with the threshold level $W_{j\max}/\alpha(\alpha \geq 1)$. This digitized pattern P_{ij} is sent to the shift register of Fig. 2.

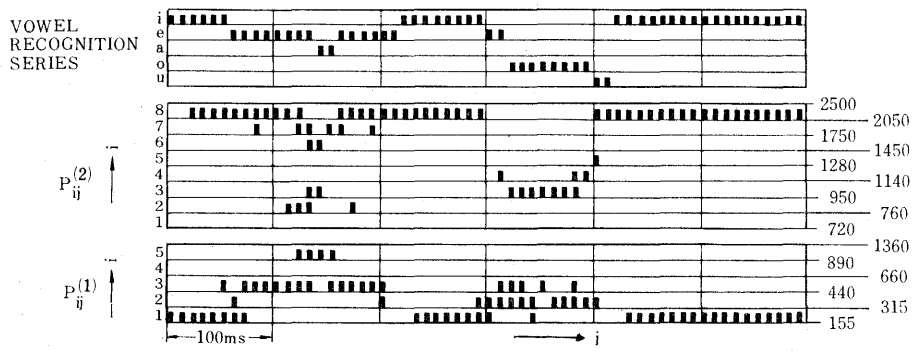


Fig. 3 Digitized pattern of zero-crossing distribution and a series of vowel recognition for the input sound "YAYOI", with the channel classification characteristics (c/s).

Fig. 3 shows an example of the digitized pattern. The distribution $P_{ij}^{(1)}$ for F_1 region and $P_{ij}^{(2)}$ for F_2 region which have 5 and 8 channels, respectively, are separatly digitized with the threshold level $\alpha \fallingdotseq 1$. In the figure the sampling is repeated every 10 ms (T=10 ms) and one black point represents 10 ms. In the upper part of the pattern a series of vowel recognition repeated every 20 ms is added. (As the pattern display cycle is 10 ms, one recognition of vowel series is shown by two black points.) The shift register of Fig. 2 memorizes the digitized pattern for the required time for processing, by which the time pattern is expressed in the form of space instead of time. The register has 13 channels and l bit length, enough to detect the stability.

The stability detector computes the stability $S_{ij}(h/l)$ by digitizing the index of stability taken from the pattern P_{ij} in the shift register.

$$S_{ij}(h/l)=1 \quad \text{when} \quad X_{ij}(l) \geq h/l$$
$$S_{ij}(h/l)=0 \quad \text{when} \quad X_{ij}(l) < h/l$$

For the vowel section, S_{ij} (6/6) and S_{ij} (4/5) were used and appropriate one of them was selected for each channel. An illustration of stability is shown below.

Sampling Interval (j)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input Pattern (i)	0	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0
S_{ij} (6/6)	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
S_{ij} (4/5)	·	·	·	·	1	1	1	1	0	1	0	0	0	1	0	0

As is seen in example, S_{ij} (4/5)=1 means that in i-th channel there appeared more than four 1's during the five intervals just before the j-th sampling inteval.

An example of stability pattern is shown in Fig. 4 where α is set close to 1 and h/l to 6/6. The pattern shows that noisy components are smoothed out and dominant channels are extracted.

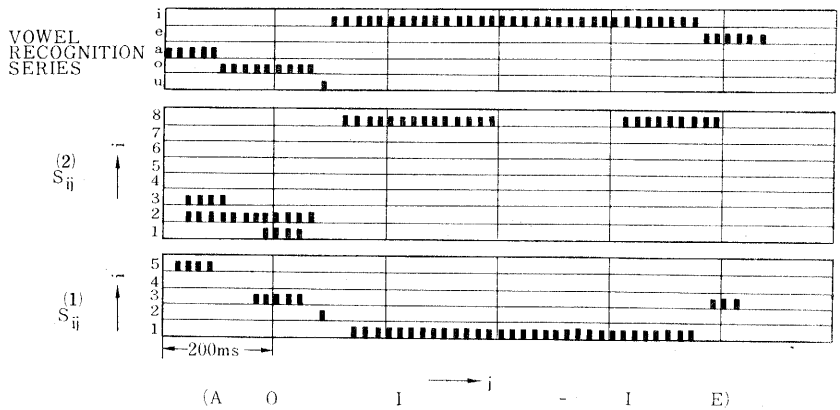


Fig. 4 Stability pattern and vowel recognition series of a connected vowel sound (Japanese) "AOI-IE"). The channel arrangement is the same as that of Fig. 3.

As the existence of the stability implies the existence of the formant, we recognize the section, during which the stability has been detected in both F_1 and F_2 regions as one segment corresponding to a phoneme.

The segmentation signal generator of Fig. 2 generates the segmentation signal every time a new combination of the stability is detected; this signal controls the recognition part of Fig. 1 with the rule shown in Fig. 7.

The function of stability detection is influenced with the characteristics of the channel chassification of the zero-crossing wave analysis, with the setting of the threshold value α of the digitizer and with the value l. These values are determined by the experimental data so as to satisfy both the detection of the stationary part and the suppression of the transition part.

In Fig. 5 schematic diagram of the stability detector and the segmentation signal generator is shown.

The length of shift register along the time axis is l=6, therefore in the j-th time interval the stored phttern is as follows;

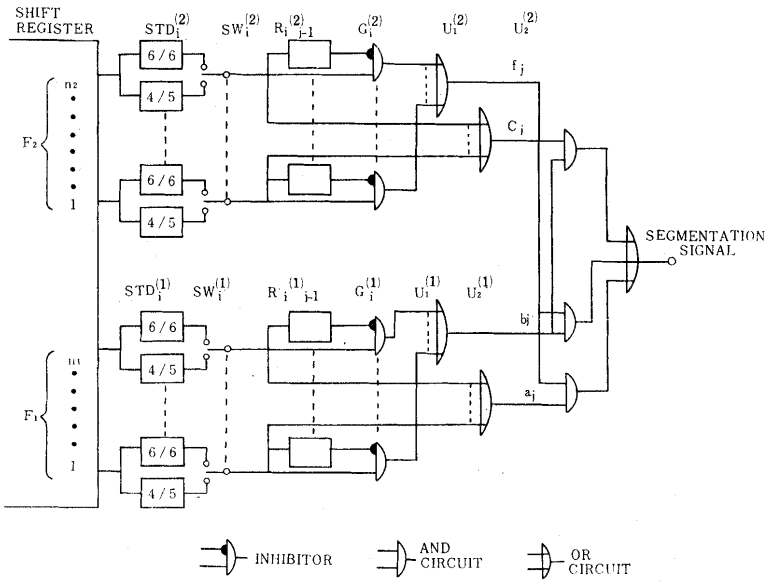


Fig. 5 Schematic diagram of stability detector and segmentation signal generator.

First formant pattern: $P_{ij}^{(1)}$ $i=1, 2, \cdots\cdots, n_1$; $j=j, j-1, \cdots\cdots, j-(l-1)$

Second formant pattern: $P_{ij}^{(2)}$ $i=1, 2, \cdots\cdots, n_2$; $j=j, j-1, \cdots\cdots, j-(l-1)$

The contents of the shift register in the j-th time interval are shifted by one bit along the time axis in the next interval. The logical circuits are connected to each memory cell of the shift register. As all the circuits are static logic using transistors and diodes, the contents of shift register and the state of the logical circuits are held constant in one interval. The stability detectors $\mathrm{STD}_i^{(1)}$ and $\mathrm{STD}_i^{(2)}$ connected to each channel give S_{ij} (h/l). 6/6 means the circuit for S_{ij} (6/6) and 4/5 for S_{ij} (4/5) and the suitable one is selected in each channel by switch $SW_i^{(1)}$ and $SW_i^{(2)}$. Merory circuit $R_i{}^{(1)}{}_{j-1}$ and $R_i{}^{(2)}{}_{j-1}$ which memorizes the detected S_{ij} for one time interval and AND circuit $G_i^{(1)}$ and $G_i^{(2)}$ are used for the detection of the beginning point of the stability.

The outputs of the OR circuits $U_1^{(1)}$, $U_2^{(1)}$, $U_1^{(2)}$, and $U_2^{(2)}$ are as follows.

a_j; shows the existence of the stability in $\mathrm{F_1}$ region.

b_i; shows the beginning point of the stability in $\mathrm{F_1}$ region.

c_j; shows the existence of the stability in $\mathrm{F_2}$ region.

f_j; shows the beginning point of the stability in $\mathrm{F_2}$ region.

The segmentation signal which tells the detection of the new vowel segment is obtained by the logical combination of the above four singnals.

3.3 Segmetation of Consonant and Vowel Section and Samping Control

The method applied to the vowel segmentation may be applied to the segmentation between consonant section and vowel section. As both sections have

distinctive characteristics, their segmentation is performed in another way in the segmentation circuit of consonant and vowel section. The lower frequency components by vocal cords excitation and the higher frequency components by formant, hiss etc. are detected, after the input speech sound is passed through low pass filter and high pass filer, respectively. By logical combination of these signals consonant section is separated from vowel section.[3] The sampling control circuit sends the sampling signal (A) to control the vowel analyzer and consonat analyzer in vowel section and in consonant section, respectively, and also sends the sampling signal (S) to the vowel segmentation circuit in vowel section.

4. Recognition of Phoneme[3]

Recognition is the other aspect of the automatic recognition of speech sound together with the segmentation described above.

The recognition part in Fig. 1 is controlled by the segmentation part and afterwards combined with the results of segmentation part. The phoneme was selected as the recognition unit.

The speech sound is characterized in its pronouncing process by the manner and place of articulation and, therefore, we designed the machine to treat the speech sound from these two aspects. That is, for the former we divide the segment of speech sound into several phoneme groups by the distinctive feature extractor and the phoneme classifier, and for the latter we discriminate the phonemes that belong to one phoneme group against each other by the analysis. The block diagram of this part is shown in part II of Fig. 1.

4.1 Classification of Phoneme

The distinctive feature extractor picks up the characteristic features from the outputs of low pass and high pass filters, considering not only the relative energy distribution, but also the time variation. The phoneme classifier detects from these features the vowel section, unvoiced consonant section, voiced consonant section, nasal section and plosiveness, from which grouping is accomplished as shown in the right column of Table 1.

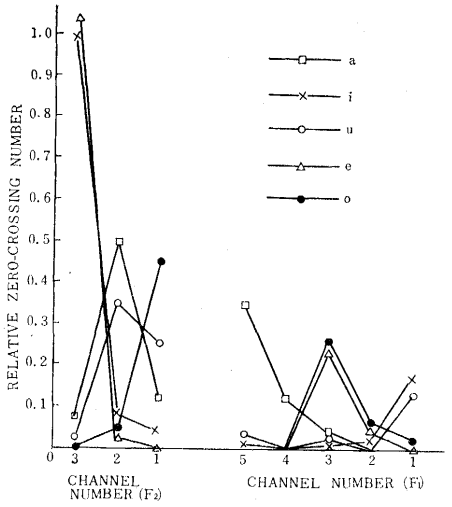
Table 1 Classification of the Japanese phonemes. (The right column shows the classification in the phoneme classifier of Fig. 1)

Phoneme	Voiced	Vowel	Vowel ; /a/, /i/, /u/, /e/, /o/	Vowel
		Consonant	Fricative ; /z/, /ʒ/, (/dz/, /dʒ/) (including affricate)	Voiced Consonant
			Plosive ; /b/, /d/, /g/, /r/ (including flapped)	Voiced Consonant
			Nasal ; /m/, /n/	Nasal
			Semi-vowel ; /w/, /j/	Transient
	Unvoiced		Fricative ; /s/, /ʃ/, /h/ (including aspirate)	Unvoiced Consonant
			Affricate ; /ts/, /tʃ/	Unvoiced Consonant
			Plosive ; /p/, /t/, /k/	Unvoiced Consonant

4.2 Vowel Recognition

In parallel with the phoneme classification, speech sound is analyzed by the zero-crossing wave analysis method in the same way as used in vowel segmentation.

Before being converted into the zero-crossing wave, speech sound is filtered into a first formant (F_1) region and a second formant (F_2) region. Zero-crossing wave intervals are measured in the vowel analyzer controlled by the sampling signal of 20 ms which is successively applied during the vowel section and is synchronized with the sampling signal (S) of the vowel segmentation circuit. For each sampling the zero-crossing distribution is obtained in both F_1 and F_2 region.



a)

First formant: Channel number	Channel number	Second formant: Channel number 3	2	1
	Frequency characteristics (c/s)	2500 ~ 1700	1700 ~ 1130	1130 ~ 680
5	790–1440	*	a	a
4	680– 790	e	a	o
3	470– 680	e	o	o
2	400– 470	e	u	o
1	170– 400	i	u	u

* No output

b)

Fig. 6 a) Zero-crossing distribution of male vowels (Japanese) obtained by the vowel analyzer, and b) vowel recognition logic in F_1–F_2 domain.

Vowel recognition is made in the vowel recognizer of Fig. 1 by finding the channels which have the peak value in each of the F_1 and F_2 distributions. Fig. 6 a) is the zero-crossing distributions in F_1 and F_2 regions of the Japanese vowels pronounced as monosyllable, whose samping time interval is 20 ms. (Ordinate is the reletive zero-crossing number. To get the distribution this must be normalized.) Fig. 6 b) is the decision logic of the Japanese vowel. Among the fifteen cells a cell situated at the cross point of the peak channels is selected and the assigned vowel is recognized. As is seen in Fig. 1, before reaching the vowel analyzer the input speech sound is passed through filter to pick up the formant region. The filters are at present adjusted for the male voice and for the female voice their characteristics must be modified.

4.3 Recognition of Consonant

The consonant analyzer has the same function as the vowel analyzer. The

sampling, however, is performed only once for one consonant section and is continued during the interval within the limits from 10 ms to 40 ms. As the function necessary to the consonant analyzer is to discriminate the phonemes which belong to each phoneme group such as unvoiced consonant, voiced consonant and nasal, an individual analyzer matched to each phoneme group is prepared.

In the consonant recognizer the zero-crossing distribution, which is classified into 14 channels, is reorganized to another channel characteristics suitable for the phoneme group to be recognized. Then the reorganized distribution is digitized to 1 bit with the threshold level previously decided from the data. Each digitized output is used as the Boolean variable for decision logic in the phoneme recognition circuit.

In the recognizer of connecctd speech of Fig. 1, some cases of phonetic contexts are treated which should be especially considered in conversational speech sound. Though it is desirable to process the whole speech sound taking the phonetic context into account, it would increase the complexity and scale of the machine. The cases considered here are the vowel insertion when a vowel sound is not pronounced in some contextual condition, the assimilated consonant (In procouncing the series of phonetic symbols "i-ts-te", /ts/ is replaced by the implosion without explosion and /i/ is pronounced briefly.) and the syllabic nesal. For the vowel insertion, an appropriate vowel phoneme estimated from the preceding and thc following consonant is added to the output of the machine.

An assimilated consonant is detectdc by the implosion for a fairly long interval which exists between the end of the preceding sound and the following unvoiced consonant with burst.

5. Combination of Segmentation and Recognion

As is stated above we have decomposed the sound into several segments and have registered necessary parameters for the recognition of the segments in the register of Fig. 1 as the digitized form. The function of the phoneme recognition circuit in Fig. 1 is to recognize, from the parameters stored in the register, final output by combining them by means of the segmentation signal.

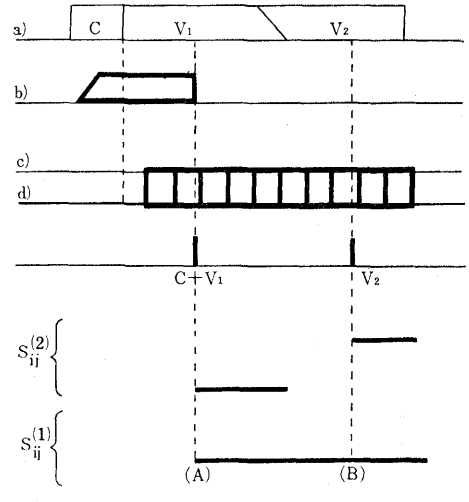


Fig. 7 Time chart of the phoneme recognition circuit; a) Input speech (consonant: C, vowel : V). b) Stored results of consonant recognition and phoneme classification. c) Vowel recognition series. d) Output control signal from segmentation part of Fig. 1.

Fig. 7 is the time chart of the phoneme recognition circuit. For example, for the input sound of $C+V_1+V_2$ (C; consonant, V; vowel) the results of the consonant analyzing system and the phoneme classifier are stored in the register; the register for vowel part is refreshed successively during vowel section. The segmentation signal which controls the recognition output, is led from the segmentation signal generator in Fig. 1. That is, the signal comes out each time a new combination of the stabilities S_{ij} in both F_1 and F_2 regions are detected.

In Fig. 7 the point (A) is the case when the stabilities are detected at the same time and the point (B) is the case when a stability has been alrrady present in one of the F_1 and F_2 regions and a new stability is found in the other region.

The phoneme recognition circuit combines, at the time when this segmentation signal is generated, the reusults of the consonant recognizer and phoneme classifier already held in the register and that of the vowel recognizer received at that time, and, in this example, prints out one "Kana" letter for the combination of $C+V_1$, and next one "Kana" (vowel) for V_2.

The phoneme recognition circuit is the decoder of the recognized code stored in the register to the "Kana" letter system, and is composed of a diode AND matrix circuit. In this decoding logic the influence of vowel on the preceding consanat is considered. After the output is sent out, all the registers are reset in preparation for the next operation. When there are no significant codes in the register of the consonant recognizer and there is vowel code in the register of the phoneme classifier, a vowel "Kana" letter is sent out.

6. Conversational Speech Recognition by Phonetic Contextual Approach

In the speech recognition system described above the speech sound was divided into several segments, that seem to correspond to the phoneme, from the time variation of the analyzed pattern and then discrimination was performed for each segment. But the parameters or distributions of a segment are affected by the phonetic contextual effect from the neighboring phonemes. The speech sound, therefore, must be recognized as a whole pattern considering phonetic context. This principle, however, is impossible to apply to the recognition system of general conversational speech.

It will be a valid simplification to consider that, though reciprocal interaction will occur between neighboring but non-adjacent phonemes, the dominant influence upon one phoneme will be the influence from a just preceding phoneme and a just following phoneme. In the general principle of phonetic contextual approach we selected, as the baisic recognition unit (or segment), the sgment of pattern corresponding to the three phoneme sequence.

As a preparatory step, a trigram of the Japanese phoneme sequence was examined, which will give us the data to design the recognition system. That is, though the possible number of combinations of the three phoneme sequence is too large, it was proved that the number of combinations actually used in daily conversation is reduced to a realizable scale. On this basis we adopted a conversational speech recognition system in which the phoneme is chosen as the recognition unit and the phonetic contextual effect from adjacent phonemes is considered. The principle was actually applied to the vowel recognition.

6.1 Principle of Recognition

The basic principle of recognition is the matching of the analyzed pattern of input unknown speech sound with the stored standard patterns corresponding to three phoneme sequences. The speed of the time variation of the analyzed pattern shows different tendencies according to the phoneme sequence that the speech sound contains. Further, it is affected largely by the articulating condition and individuality. Because of this variety the necessary number of standard patterns prepared in the system would reach an unrealizable size. As an effective method to solve this we tried to express the pattern of speech sound by a set of patterns; the sequential pattern and the weighting pattern.

The sequential pattern is the pattern that shows the sequence of states, that is, how the parameters or the distributions of speech sound at a sampling point change with time.

The weighting pattern is a series of numbers that show the number of the successive occurences of each state composing the sequential pattern. These expressions of pattern may be useful for any pattern, but it works more effectively for the pattern of speech sound where the parameters change is negligible or gradual in the most section except for some time points. The effect of the speed of articulation on the pattern in this method is expressed mainly in the size of the weighting pattern, and the sequential pattern and the shape of the weighting pattern are not significantly intfluenced by the articulation speed.

In the system the matching operation is first executed upon the sequential pattern and the weigihting pattern in parallel and then both results are combined to produce final matching of the whole input pottern. By this two stage logic of matching, the complexity of logic and circuit is reduced.

6.2 System of the Recognition Circuit

Fig. 8 is the block diagram of the system. The timing of the whole system and that of the input pattern are synchronized by the shift pulse timing. A new input coming to the input terminal is compiled either to the weighting pattern or to the sequential pattern in the shift register. During the time the information flows from the input stage R_n to the output staga R_o of the shift register memory, logical operation is successively carried on its contents. Thus, the memory capacity

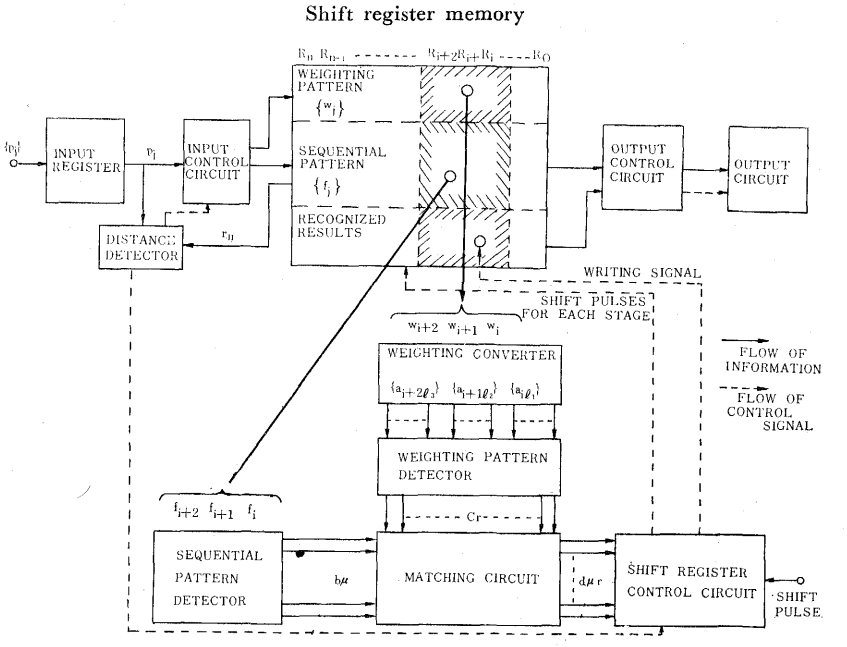


Fig. 8 Block diagram of speech recognition system, considering the phonetic contextual effect.

of the shift register need not be too large and in principle there is no limitation no the length of the input speech sound.

The logical circuits, whose input is fed from the shift register memory, are designed to work in parallel. After the shift register changes the state by means of the shift pulse, in the time interval to the next shift pulse the shift register, the logical circuit and the shift register control circuit are held in a constant state. By the next shift pulse the new input is fed to the shift register and at the same time the shift register control circuit puts its output logic into use to control the shift register and the output circuit. The processing of the input pattern is, therefore, finished in real time.

Let us think of the speech pattern P as the time series of the parameters or distributions P_j obtained in the time interval j. Then P is written as

$$P=\{P_j\}=P_1,\ P_2,\ \cdots\cdots,\ P_j,\ \cdots\cdots.$$

We define the word "run" as follows: a longest series of the parameters taken from the pattern $\{P_j\}$, where all the adjacent parameters are recognized as the same based on some criterion. The sequence of runs obtained from the parameters $\{P_1\}$ is denoted by $f_1, f_2, \cdots, f_k, \cdots$, where it is supposed that $P_{j-1}=f_k$, and the length of f_k denoted by w_k. Then a set of $\{f_k\}$ and $\{w_k\}$ have the one to one correspondence with the original pattern $\{P_j\}$; they also correspond to the sequential pattern and weithting pattern, respectively, as we have defined them above.

In a real system, P_j is given as a time series. In the j-th interval t_j between the time points j and $j-1$, the latest part of the input pattern is stored as $\{f_k\}$, $\{w_k\}$ in the shift register and the next input P_j exists in the input register. The distance detector of Fig. 8 computes the distance between the contents of the input register R_{n+1} and that of the R_n stage of the shift register. The input control circuit compiles, under the control of the distance detector, a new input in R_{n+1} to the pattern which is expressed as $\{f_k\}$, $\{w_k\}$ in the shift register.

Now let r_{nj} be the contents of the register R_n in the time interval t_j. Then $r_{n+1\,j}=P_j$. On the other hand $P_{j-1}=f_k$, then $r_{nj}=(w_k, f_k)$ and $r_{n-1\,j}=(w_{k-1}, f_{k-1})$.

The logic in the distance detector is as follows:

(a) For a given threshold ε, if the distance $g_j=|P_j-f_k|\leq\varepsilon$, it is recognized as $P_j=f_k$, and in the next time interval t_{j+1}, R_n is changed to $r_{nj+1}=(w_k+1, f_k)$, without any change in the other part of the shift register.

(b) If $g_j>\varepsilon$, then it is recognized as $P_j\neq f_k$, and in the next time interval t_{j+1} all the contents of shift register are shifted by one to the output side and the R_n is set to $r_{nj+1}=(1, f_{k+1})$, where $P_j=f_{k+1}$. (There are some subsidiary functions omited here.)

The pattern stored in the shift register is deteceed by the diode logics of the detection circuits which are connected to stages R_i, R_{i+1} and R_{i+2}. That is the length of pattern to be looked up at one time is the length of three runs. The sequential pattern f_i, f_{i+1}, f_{i+2} is led to the sequential pattern detector, where one out of the standard patterns $\{b_\mu\}$ which are set as diode logic in the circuit is selected. A sequential pattern generally corresponds to more than one phoneme sequence. Which phoneme sequence is the desired output is then decided by the detection of the weighting pattern.

The weight which is stored in the shift register as count number is converted to an inequality relation by the weight converter. This circuit may be thought as a decoder that generates, for each of the weights w_i, w_{i+1} and w_{i+2}, a number of outputs $\{a_1\ l_1\}$, $\{a_{i+1}\ l_2\}$ and $\{a_{i+2}\ l_3\}$ by the following logics.

When some values l are given for w_i, the logic of a_{il} is

$$a_{il}=1 \quad \text{when} \quad w_i<l$$
$$a_{il}=0 \quad \text{when} \quad w_i\geq l$$

and there are also the outputs of its negation $\{\overline{a_{il}}\}$

The weighting pattern detector decodes the weighting pattern as a combination of the inequality relations, and gives signals to some of the prepared channels $\{c_\nu\}$. For example if it is required to generate the signal for the weithting pattern

$$l_1'\leq w_i<l_1,\quad l_2',\leq w_{i+1}<l_2,\quad l'_3\leq w_{i+2}<l_3$$

then the deconding logic is

$$c_\nu=a_{il_1}\cdot\overline{a_{1l_1'}}\cdot a_{i+1l_2}\cdot\overline{a_{i+1l_2'}}\cdot a_{i+2l_3}\cdot\overline{a_{i+2l_3'}}$$

The matching circuit combines the detected sequential pattern b_μ with some weighting pattern c_ν's that are necessary to distinguish the detcted pattern b_μ. Generally there are several weighting patterns to be combined with a certain sequential pattern. The logics among such c_ν's are exclusive, but not always exclusive between any c_ν's which are not combined to the same sequental pattern. The result of the matching is the decision for the run of length three, that is, a part of the whole input pattern $\{P_j\}$. By repeating this procedure successively in each time interval, the final recognition result is obtained.

After a shift pulse, selection of one of the output lines $d_{\mu\nu}$ is being performed. The function of shift register control circuit is the control of the shift pulse at each stage (a shift pulse from R_n to R_{n-1} is written as s_{n-1}) and the writing of the recognized results. The shift pulse at each stage is independently controlled by the logic of this circuit which is determined by the selected input line $d_{\mu\nu}$. That is, if it is required to eliminate the contents of the R_{i+1}, then

$$s_0=s_1=\cdots=s_i=0 \text{ and } s_{i+1}=s_{i+2}=\cdots=s_n=1$$

($s_i=0$ means that no shift pulse is added to R_{i-1}.)
The elimination peocedure occures when the matched part of the pattern is the transient interval from one phoneme to the next. On the contray, when the part was recognized as a recognition unit, the writing signal puts the result into the shift register memory corresponding to the pattern.

After the processing of the pattern is finished it flows out of the output side R_0 of the register, where the parts of the pattern recognized as the recognition unit are picked up by the output control circuit to be sent to output circuit.

6.3 Experiment in Vowel Recogntion

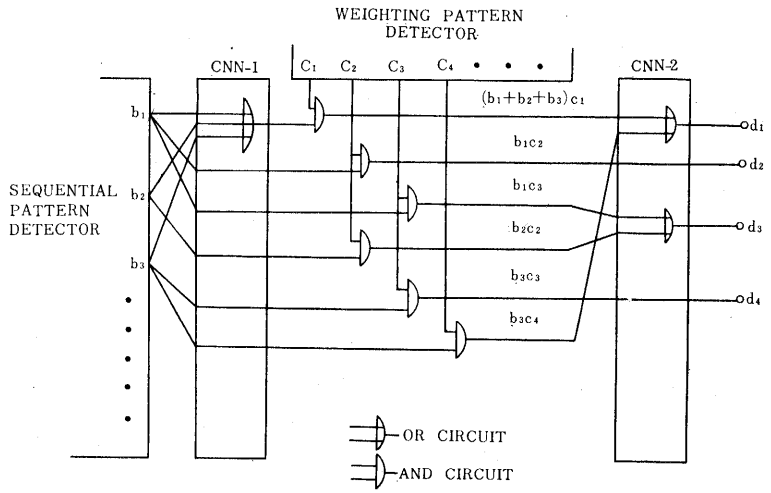
The application of this principle to the conversational speech recognizing machine would need considerably large scale devices.

The system was put into practice for the vowel part and the aforementioned functions of vowel segmantation and vowel recognition was replaced by this system. For simplicity the input series is the vowel recognition series sampled each 20 ms as shown in Fig. 3. Though the possible number of the standard sequential patterns is $5^3=125$, the prepared number of the patterns $\{b_\mu\}$ is reduced to about 80 by merging them. The memory capacity of the shift register is 6 stages and the maximum weight is 16 (320 ms).

An example of operation of the distance detector (the same as the coincidence circuit in this experiment) is shown in Fig. 9 a). In the case when the contents of both R_4 and R_5 are equal, R_5 is added to R_4 and, when not equal, by the next shift pulse R_5 is shifted to R_4, R_4, to R_3, etc.. As is seen in the pattern at t_4 of Fig. 9 a), in case R_4 and R_{in} have the same contents but differ from R_5, the parameter in R_5 is regarded as a noisy component and is added to the weight of R_4.

	R_6	R_5	R_4	R_3
t_1	a_1	•	•	•
t_2	a_2	a_1	•	•
t_3	e	a_2	a_1	•
t_4	a_3	e	$2a_1$	•
t_5	i_1	a_3	$3a_1$	•
t_6	i_2	i_1	$4a_1$	•
t_7	•	i_2	i_1	$4a_1$
t_8	•	•	$2i_1$	$4a_1$
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•

a) An example of states of shift register for input series [a_1 a_1 e a_3 i_1 i_1···]. $2a_1$ means that weighting of a_1 is 2. R_6 is the input register.



CNN : connection board

b) An example of connection of matching circuit.

Fig. 9 Example of the vowel recognition system.

The sequential pattern detector and the weighting pattern detector are connected to the stages 2, 3 and 4 of the shift register. The sequential patterns are grouped into the following categories: (1) patterns that may belong to transient parts, (2) patterns to be recognized as one vowel and (3) patterns that may represent semi-vowls. Whether a pattern of group (1) is a transient part from one phoneme to the next or is to be recognized as one phoneme is decided by the state of the corresponding weighting pattern; for example, the decision whether "o" in the vowel series "eou" is transient or not. The pattern of group (2) is recognized as a phoneme, because the movement of formants show the distinctive

change. For example, the series "ioi" will show the presence of the phoneme /o/. For the pattern of group (3), whether it is a semi-vowel or not must be decided. (We have two semi-vowels /j/ and /w/ in Japanese.)

The threshold values l of the weight converter used here as follows; 2, 3, 4, 5, 8, 9, 10, 13. The sampling time interval is 20 ms and 13 means 260 ms. 45 output lines $\{c_\nu\}$ of the weighting pattern detector are selected to combine with the sequential patterns $\{b_\mu\}$. A connection example of the matching circuit is shown in Fig. 9 b) based on the logics shown in the figure. Connection circuits (CNN) are used to merge the logics. The number of final matched outputs $\{d_{\mu\nu}\}$ is about 25. This vowel recgnizer recognizes the five Japanese vowels, long vowels and semi-vowels. When this recognizer is combined with the speech recognition system described above, the system can process the spoken words, though there are a few limitations in category and in the speed of articulation at the present stage. The score using this recognizer is better than method of stability. The faults in judgement are mainly caused by the errors in the input series itself and insufficient information in the series because it simplified to the five vowels.

7. Conclusion

It is not yet clear whether the mechanical recognition of the conversational speech is possible or not. The methods described in this paper are some attempts at the mechanical recognition. We described several problems to be solved in the course of the speech recognition and attempts at solving them, some of which were put into the actual machine. In principle we did not limit the input vocabulary of the conversational speech. With the problems of what parameters ought to be employed, it is also necessary to find a general principle to process the conversational speech from the phonetic contextual point of view. The combined expression of the pattern by the sequential pattern and weighting pattern is, we think, one of the more effective ways of the pattern processing.

Acknowledgement

We would like to express sincere thanks to all persons who concerned themselves with this project, especially to Dr. K. Maeda, professor at Kyoto University for his encouragement, to Dr. S. Inoue, Mr. K. Kagiyama, Mr. K. Shirai, Mr. T. Kurashita and Mr. K. Hashimoto for support in the experiments, and to Dr. T. Kurokawa, Mr. H. Tomonari, Mr. T. Tsuzaki, Dr. T. Sekimoto, Dr. K. Nagata, Mr. Y. Kato, Mr. H. Kaneko and Mr. H. Kondo of the Nippon Electric Co. Ltd. (NEC), Tokyo, Japan, for the construction of the systems.

References

1) T. Sakai and S. Inoue. "An Analyzing Equipment for Zero-Crossng Interval and Its Application to Speech Analysis" The Journal of the Institute of Electrical Communication Engineers of Japan. April, 1956 (Japanese)

2) T. Sakai and S. Inoue. "New Instruments and Methods for Speech Analysis" JASA Vol. 32, No. 4, April, 1960.

3) T. Sakai, S. Doshita and K. Hashimoto. "The Automatic Recognition System of the Japanese Monosyllable", Convension of the professional group on Automaton and Automatic Control of the Institute of the Electrical Communication Engineers of Japan, Jan. 1961 (Japanese)