STUDIA PHONOLOGICA II (1962)

An Automatic Recognition System of Speech Sounds*

By

Toshiyuki SAKAI and Shuji DOSHITA

Kyoto University, Kyoto, Japan

1. Introduction

The phonetic typewriter is a device for converting the sounds of the human voice into typed letters. The development of such equipment has long been a human ambition, because it is not only an instantanuous and untiring converter of speech to writing, but also an efficient tool for transmitting and processing information between human beings and machines.

Speech sounds are very complicated, having multiple dimensions (time, frequency, intensity, envelope, etc.), and the acoustic waveforms corresponding to a given word or speech sound are far from identical from person to person and dialect to dialect.

For the development of such equipment it is necessary to consider two processes: the automatic segmentation of continuous speech sounds and the pattern recognition of segmented sound.

This paper discusses the principles involved and some experiments related to the automatic segmentation of continuous speech sounds into discrete sound segments corresponding to phonemes.

For this processing we use primarily two criteria called "stability" and "distance" applied to the results of a zero-crossing-analysis expressed in binary form.

Pattern recognition of speech sounds is also discussed from distinctive features' point of view.[1)]

Logical combination of the binary representation of the signals derived from many parallel filter circuits is used to determine phoneme classification to the unknown input speech.

This phonetic typewriter is flexible and is useful for processing speech sounds of other languages as well as of the Japanese language.

Throughout the system 3,000 transistors and 5,000 diodes are used to the

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.

* The project has been supported by the Grant from the Japanese Ministry of Education.

1) Pattern recognition is performed on the same principle as described in the reference.

speech input, phoneme classification, analysis, segmentation, recognition and symbol output.

At the end of the paper the result of the trigram of the Japanese phonemes is presented. This was intended to obtain the basic data for the contextual recognition and segmentation in the phonetic level which we are trying.

2. Fundamental Structure of Phonetic Typewriter

Fig. 1 shows the block diagram of our phonetic typewriter named Sonotype, which is divided into two essential parts: segmentation part and recognition part.

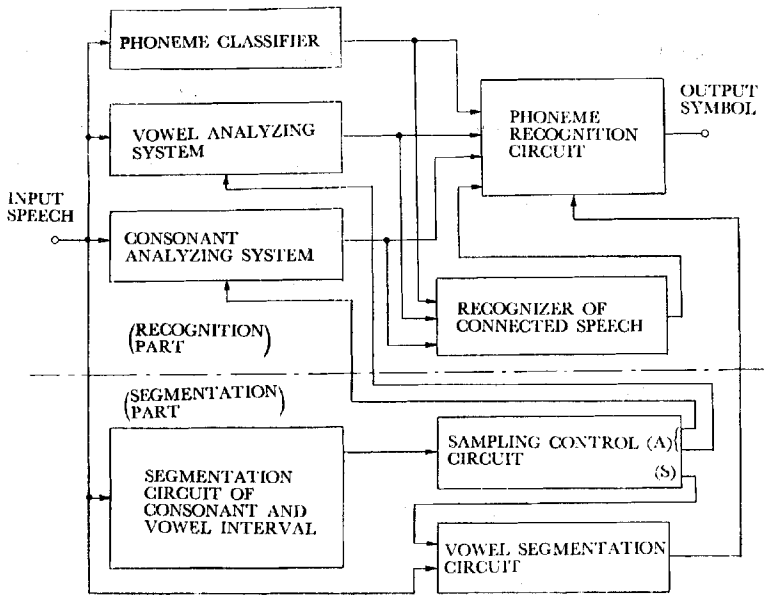


Fig. 1 Block diagram of phonetic typewriter.

Segmentation circuit of consonant and vowel interval in Fig. 1 divides the continuous analog speech wave into the consonant interval and the vowel interval, and vowel segmentation circuit further into phoneme units in the segment of vowel. Recognition part is to recognize the phoneme with parameters from the phoneme classifier and the analyzer for vowel and consonant operating in the respective segments divided by the sampling control circuit of the segmentation part.

Phoneme classifier is inserted with the aim of making groups in the segmented units corresponding to the phoneme classification which is used in the theory of speech as the manner of articulation.

Vowel analyzing system and consonant analyzing system are the series of analyzing and digitalizing circuits respectively for the corresponding parameters to the place of articulation.

Recognizer of connected speech is able to process the patterns of special phoneme sequences which may appear in conversational speech and are different from the ordinary combination of phoneme units.

Phoneme recognition circuit is the final recognition part for the unknown input, in which as the results one output in phonemic symbol or code is selected. In the case of Japanese, we are at present using as output the Kana letter system that is used in commercial telegraph system.

3. Segmentation

The information bearing parameters of speech sounds are continuous, analog quantity with respect to time.

Considered from the sampling theory, amount of sampled information of the speech sound is too large for the machines to process in real time. But the speed of phoneme production originated from articulatory organ is low enough for real time processing by machine owing to the speaking mechanism of human being.

The discrete phonemic series in the brain of human being is modulated by articulatory organs into the continuous speech wave that includes not only the manner and place of articulations but also non essential transitions.

As the philosophy of finding out the corresponding period to the phonemes, decision of vowel interval is considered first because vowel is the most dominant and important information obtainable from the continuous speech wave.

It is necessary to define the two terms, distance and stability used for the automatic segmentation of speech sounds. Let us suppose the time sequence of information in digital form. For example, it may be a spectrum analysis of speech sounds represented in binary form (it may be expanded into the case of multiple bits for one parameter).

At a sampling point of j, the i-th parameter (or distribution) is denoted by P_{ij}, then the parameter set may be expressed as

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

The whole pattern P is

$$P = \{P_j\} = \{P_1, P_2, \cdots\cdots\cdots\cdots, \}. \tag{2}$$

To distance d is defined by

$$d_j = \sum_i (P_{ij} \oplus P_{ij-1}) \tag{3}$$

Where $\oplus$ is the sign of exclusive OR (mod. 2), $\sum$ is the ordinary summation.

The distance of d_j is similar to the one used in the theory of coding, and in the case of spectrum or zero-crossing analysis it takes a fairly large value for non-stationary phonemes. For the stationary speech sound like vowels, the pattern in the acoustical analysis has usually small distance, but very large distance for plosive sounds.

The index of stability is defined as follows:

$$X_{ij}(l)=\frac{1}{l}\sum_{k=0}^{l-1}P_{ij-k} \tag{4}$$

where l is the number of quantized time points to be considered for the processing of pattern. X_{ij} (l) is the number of existence of "1" concerning to the i-th parameter, normalized by the maximum interval l before the sampling point of j. The value of X_{ij} (l) gives an important information in finding out the ending of stationary or transient intervals for suitable value of l.

In Fig. 2, the block diagram of segmentation part is shown in the case of a zero-crossing wave analysis of speech sounds.

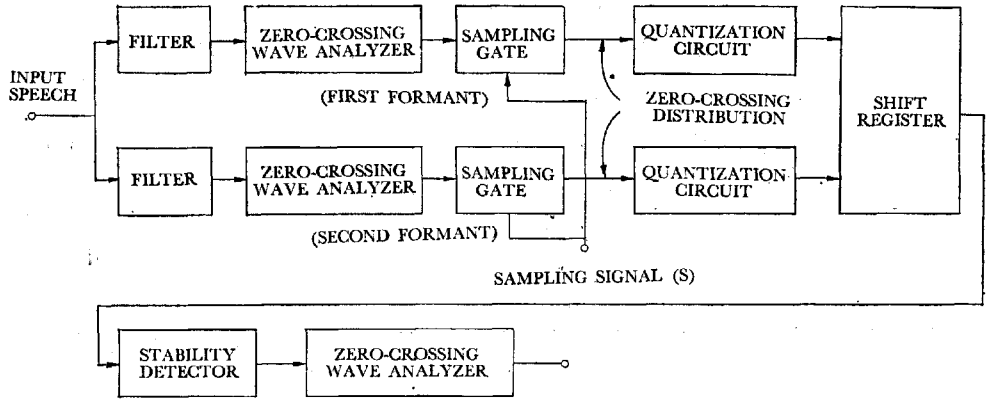


Fig. 2 Block diagram of vowel segmentation part.

The analysis of zero-crossing wave method will be discussed latter. The analysis during about 10 ms gives channel classified distribution W_{ij} corresponding to the number of counts existed in this period j.

The distribution of the zero-crossing analysis W_{ij} is digitized to P_{ij} of equation (1) in quantization circuit by setting up a threshold relative to the maximum value at this sampling period j.* The processing for distance and stability is made in a shift register having a length of l.

The segmentation signals to control phoneme recognition circuit are derived from the logical combination of stability which is the digitalized form of index of stability and distance.

The zero-crossing wave is a rectangular waveform as shown in Fig. 3, in which only the time points crossing the zero level of the original speech wave are maintained as the information bearing parameter, because the wave has constant levels of positive or negative according to the polarity of the original wave.

The zero-crossing wave analysis is the measurement of the width of rectangular waves, and may be integrated as a statistical distribution for a certain time interval T.

* These procedures are explained in reference.

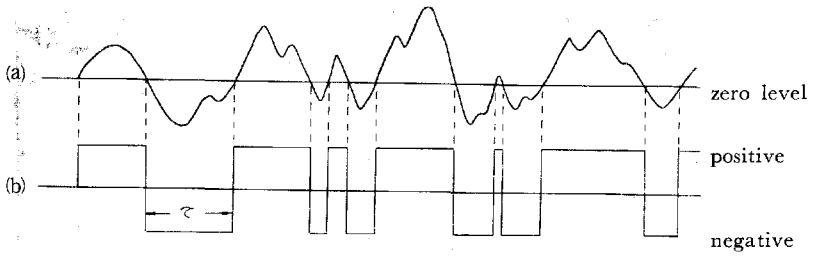


Fig. 3 Examples of (a) original speech wave.
(b) zero-crossing wave.

A pattern sequence of those statistics will be obtained by successive sampling.

We denote N_{ij} as the number of occurrance of zero crossing width τ, which appears in the i-th channel, whose channel width is $\Delta\tau_i$ and center value τ_i, and in the j-th sampling, whose interval is T. Then a statistical expression is

$$W_{ij}=W_j(\tau_i)=\frac{1}{T}\frac{N_{ij}\tau_i}{\Delta\tau_i}. \quad (i=1, 2, \cdots\cdots, n),\ (j=0, 1\cdots\cdots)$$

and the pattern is

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

The quantization circuit converts W_{ij} to the M bits signal. This W_j corresponds approximately to the spectrum of the original wave. Therefore we can obtain simply and effectively the formants by the following method.

Let us suppose that $M=1$ and the peak value in the n channel distribution W_j is W_{jmax}, then the threshold value of quantization circuit is decided as W_{jmax}/α ($\alpha\geq 1$), and W_{ij} is converted to P_{ij} of equation (1).

Fig. 4 shows some examples thus obtained.[1] F_1 and F_2 regions are divided into 5 and 8 channels respectively and for each of these F_1 and F_2 regions the digitized pattern is obtained automatically.

The sampling is repeated every 10 ms ($T=10$ ms). The shift register of Fig. 2 memorizes the pattern of the digitized sequence of the input speech sounds in the form of space instead of time. This register should have the memory space long enough to detect the stabitity. The stability detector computes the stability S_{ij} (h/l) from X_{ij} (l) with the following rules by simple combinational circuits.

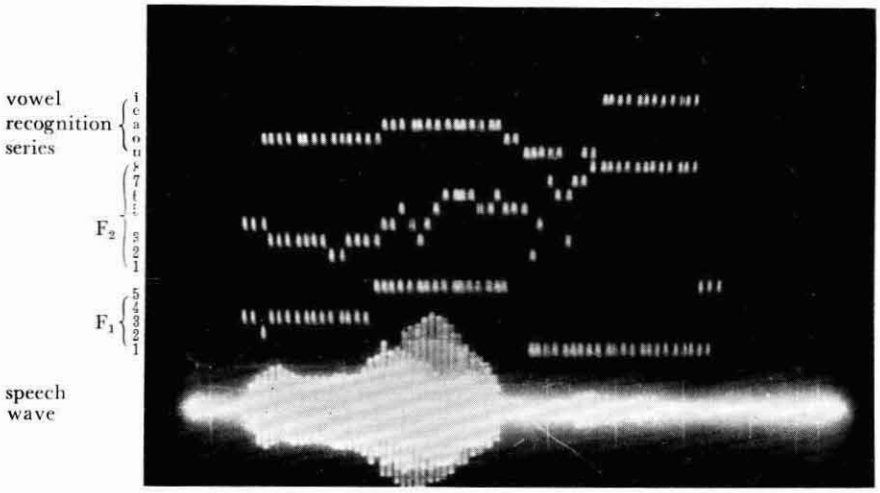
For example,

S_{ij} (6/6)=1 When X_{ij} (6)=1
S_{ij} (6/6)=0 When X_{ij} (6)<1
S_{ij} (4/5)=1 When X_{ij} (5)≥4/5
S_{ij} (4/5)=0 When X_{ij} (5)<4/5

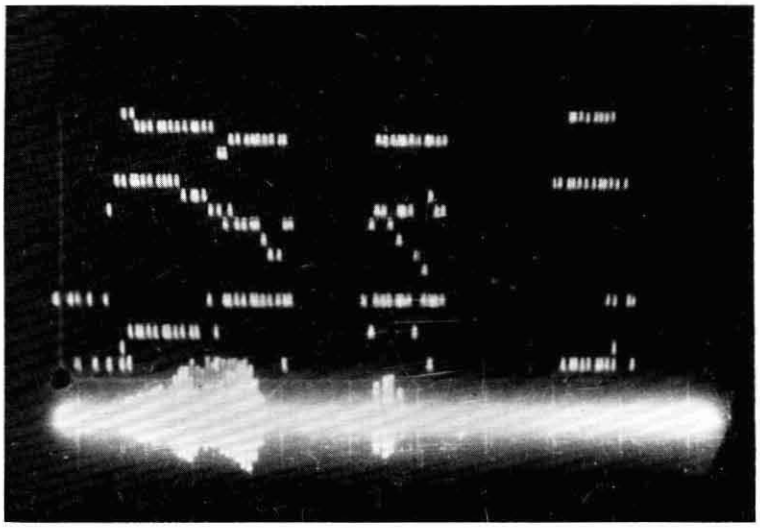
Fig. 5 shows examples of S_{ij}. The stability appears

channel No.		frequency characteristics (c/s)
F_2	8	2500–2050
	7	2050–1750
	6	1750–1450
	5	1450–1280
	4	1280–1140
	3	1140– 950
	2	950– 760
	1	760– 720
F_1	5	1360– 890
	4	890– 660
	3	660– 440
	2	440– 315
	1	315– 155

1) These patterns are also shown in reference.



(a) T O A M I



(b) M E A K A SH I

Fig. 4 Zero-crossing time pattern of F_1 and F_2 region sampled each 10 ms, (T = 10ms), and instantaneous vowel recognition series. (Speech samples are taken from Japanese words)

in certain channels characterized by the input sounds and the degree of stability has a different value according to the input sounds, so that, for example, suitable combination of either S_{ij} (6/6) or S_{ij} (4/5) is used for each channel.

SAMPLING POINT (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	0	0	0	0	0	0	0	0
S_{jj} ($^4/_5$)	0	0	0	0	1	1	1	1	1	1	0	0	0	1	0	0

Fig. 5 Examples of stabilities.

As the existence of the stability implies the existence of the formant, we recognize the interval, during which the stabiliey has been detected in both F_1 and F_2 regions, as a certain segment of a phoneme.

The segmentation signal generator of Fig. 2 is activated every time when a new combination of the stabilities S_{ij} is detected, and this segmentation signal controls the recognition part of Fig. 1 with the rule shown in Fig. 8.

The degree of the stability detector is influenced with the characteristics of the channel classification of the zero-crossing analysis, the setting of the threshold value α of quantization circuit and the value l for the decision of the pattern length. These values are determined by the experimental data statisfying both the detection of the stationary part and the suppression of the transition part.

Fig. 6 is a photograph of detected S_{ij} pattern from stability detector, in which α is set to 1 and k/l to 6/6. Speech sample used is [aoi-ie]. Channel arrangement and sampling time interval are same with those in Fig. 4. The pattern shows that noisy components are smoothed out and dominant channels are extracted.

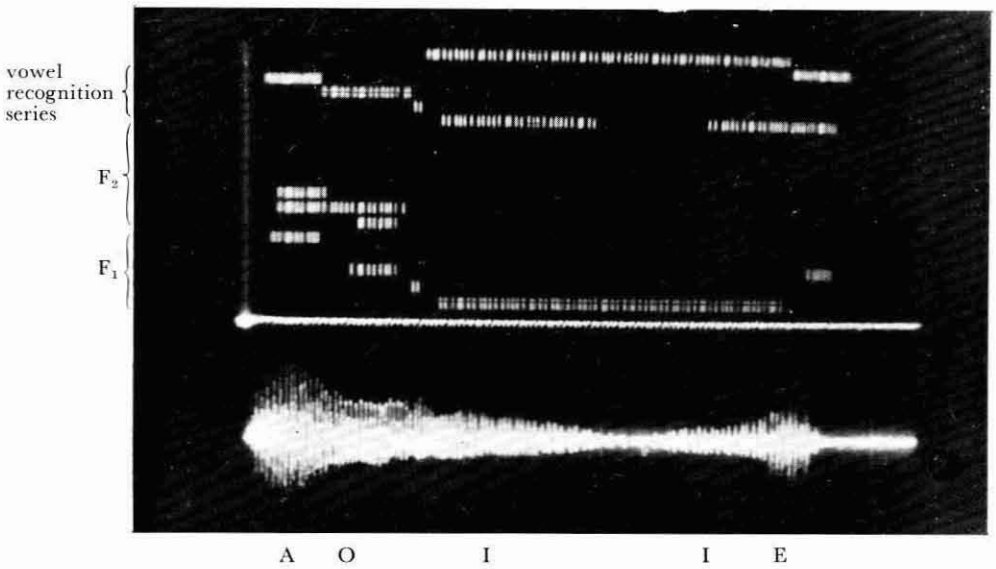


Fig. 6 Stability pattern detected in stability detector. Sample sound is an atterance of the word [aoi-ie]. Channel arrangement is the same to that of Fig. 4.

4. Recognition

Recognition is the other important aspect for the automatic recognition of speech sounds together with the segmentation described above.

The recognition part in Fig. 1 is controlled by the segmentation part, and afterwards recombined with the results of the segmentation part. The recognition of speech sounds is performed by selecting phoneme as the recognition unit.

The speech sounds are characterized in its pronouncing process, by the manner of articulation and by the place of articulation, and therefore we treat the speech sound from these two aspects. That is, for the former one we divide the phoneme into several groups, passing through distinctive feature extractor and phoneme classifier, and for the latter we separate the phoneme in the same group of manner of articulation against each other by the analysis.

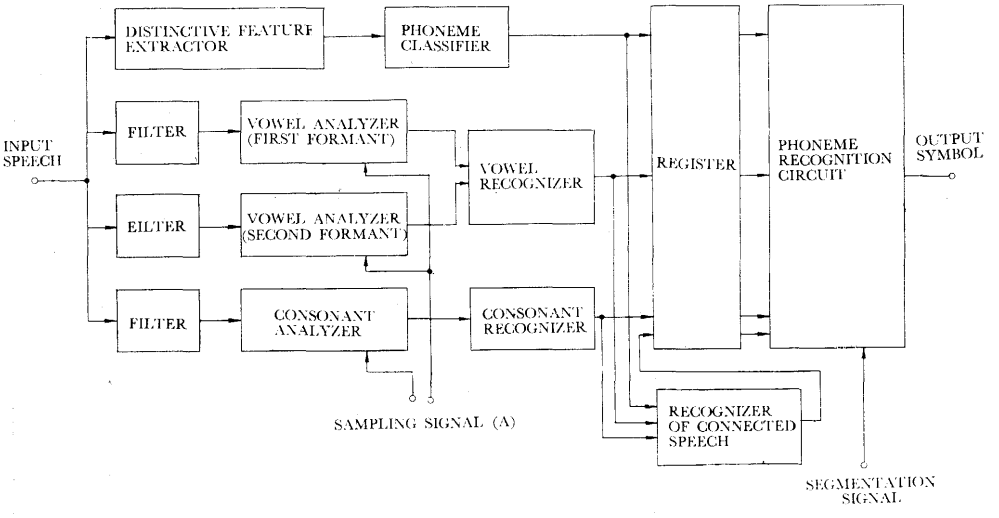


Fig. 7 Block diagram of recognition part.

Fig. 7 shows the block diagram of this recognition part. Distinctive feature extractor picks up the characteristic features from the output of low pass, high pass and band pass filters, considering not only the relative energy distribution of the input but also their time variations.

The phoneme classifier detects from these features, the vowel interval, unvoiced consonant interval, voiced consonant interval, nasal interval and plosiveness, from which grouping is accomplished as shown in the right column in Table 1.

Before being converted into the zero crossing wave, the sound wave is pre-

Table 1 Classification of the Japanese phonemes. (the right column shows the classification in phoneme classifier of Fig. 6)

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)	
			Nasal ; /m/, /n/	Nasal
			Semi-vowel ; /w/, /j/	Transient
	Unvoiced		Fricative ; /s/, /ʃ/, /h/ (including aspirate)	Unvoiced consonant
			Affricate ; /ts/, /tʃ/	
			Plosive ; /p/, /t/, /k/	

processed by the filters having the required characteristics.

The analyzing systems are different for the consonants and the vowels. For the vowels, after filtering into the first formant (F_1) and the second formant (F_2) regions, the zero-crossing wave interval is measured in real time by the vowel analyzer for the period decided by the sampling signal and gets zero-crossing distributions W_{ij} in both F_1 and F_2 regions. These distributions are sent to the vowel recognizer.

Sampling signals (A) in Fig. 7 are generated by the sampling control circuit in Fig. 1, in which the sound is segmented into the consonant interval and the vowel interval.

For the consonant, sampling interval is a definite period and for vowel successive sampling of constant period (e.g. 20 ms) is performed. This sampling signal A is synchronized with the sampling signal S of the vowel segmentation circuit in Fig. 1.

Vowel recognition is done in the vowel recognizer of Fig. 6 by finding the channels which have the peak in each of the F_1 and F_2 distributions. For example in case of the Jaganese vowels, the decision logic is shown in Table 2.

Table 2 Frequency characteristics of the channels of the vowel analyzer and its decision matrix in F_1-F_2 region.

		Second formant		
First formant	c/s	2500 1700	1700 1130	1130 720
	1440–790	no out put	a	a
	790–680	e	a	o
	680–470	e	o	o
	470–400	e	u	o
	400–170	i	u	u

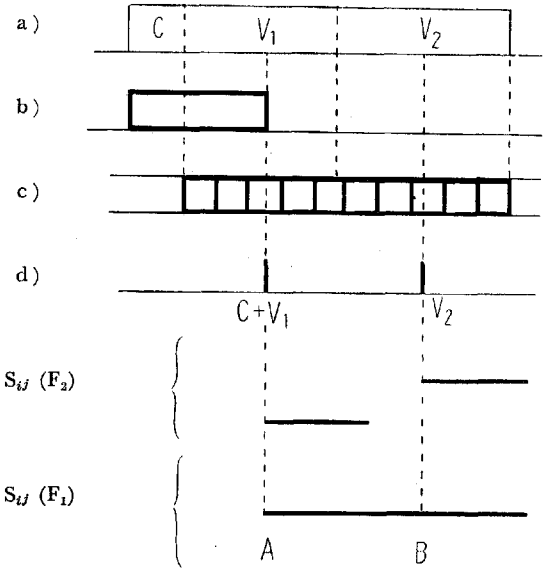
For the consonant recognizer only simple characteristics are necessary to discriminate the phonemes which belong to the same phoneme classification group.

If a sound is found to be unvoiced consonant and plosive in phoneme classifier, then there remains only the necessity of the discrimination among |k|, |t|, and |p|, which is done by the zero-crossing wave analysis summarized in consonant recognizer. Therefore different channel characteristics of W_{ij} is set up in the most suitable form for the discrimination belonging to the individual phoneme group and the output of these channels are digitalized into "1" or "0" with the threshold values which are determined statiscally.

5. Conbination of Segmentation and Recognition

As is stated above we have decomposed the sounds into several parts and have registered necessary parameters for the recognition in the register of Fig. 7 as the digitized form. The function of phoneme recognition circuit in Fig. 7 is to recognize, from these parameters stored in register, the final output by combining them by the segmentation signal.

Fig. 8 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 regisier, and the register for vowel part is refreshed successively each time the vowel part is sampled. The segmentation signal which controls the recognition output, is obtained from the segmentation signal generator in Fig. 2. 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.



a) Input speech b) registered results of consonant recognizer and phoneme classifier c) Vowel recognition series d) Segmentation signal

Fig. 8 Time chart of the operation in the phoneme recognition circuit.

In Fig. 8 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 already 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 results of the consonant recognizer already held in the register and that of the vowel recognizer received at that time, and prints out one "Kana" letter for the combination of $C+V_1$, and next one "Kana" (vowel) for V_2.

6. Statistics of Phoneme Sequence in Japanese

Speech recognition system will be considered at several levels of processing; 1) acoustical analysis, 2) phonetic processing, 3) linguistic processing. In level

1), parameter extraction is performed by the method which will not depend on the particular language, though 2) and 3) need different treatment for each language. In level 2), recognition is made based on the analyzed data and considering phonetic context. In level 3) phonetically recognized results are processed as a message or sentence, where linguistic information is utilized for the improvement of score and for the final decision. In our daily conversation perfect understanding of speech owes the lingustical redunduncy which complement the uncertainty in acoustical and phonetical recognition.

In the course of speech processing from the view point of phonetic context, the relationship between phonemes are rather complicated, but it may be simplified by taking the facts into consideration that the influence of primary importance to one phoneme may be limitted to the last preceeding and the next following phoneme. That is, in case of processing of a speech patten segment, a pattern corresponding to 3 phoneme sequence must be jointly processed. As the speech samples for these processing, many possibility may exist. One will be to select as samples the words or idioms which are frequently used in everyday conversation. But more systematic way is to select a group of words that contain some phoneme sequence to be examined. In practice, we classify the phonetic context, required for the recognition of conversational speech, into several essential items. These items are, for instance; relation of pattern between /ja/ and /ija/, relation between long vowel and short vowel, influence of adjacent consonants to vowel, and vice versa etc.. Next a group of words for each item is selected from the Japanese vocabulary.

These samples were selected for the purpose of examining the effect of phonetic context, but in construction of practical recognizing machine the words which frequently appear in conversation might have great importance. As we now treat speech sounds not as the word but as the three phoneme sequence, the phoneme sequences which frequently appear in conversation might have importance, too.

With those reasons trigram of Japanese phoneme sequence was examined. The statistics of Japanese is usually obtained by choosing the KANA letter as the symbols. As you know, KANA letter is an orthography and it is usually composed of consonant and vowel as in Roman letter expression, therefore it is not suitable for the expression of speech. So, having chosen as element of trigram the phoneme and some other simbols, we calculated the frequency of occurence of the combination of three phoneme sequence (i, j, k) by digital computer named KDC-1 (Kyoto University Digital Computer). The 23 elements selected are listed in Table 3. The selection is not ideal by the limtita-

Table 3 23 phonemes and elements used for the trigram of the Japanese phonemes.

vowel	a, i, u, e, o,
semi vowel	j, w,
unvoiced consonant	k, t, p, s, h,
voiced consonant	g, d, b, z, r, m, n,
other elements	long cowel symbol syllabic nasal assimilated consonant space

tion of the memory capacity of the computer, for instance; long vowel was expressed as "vowel phoneme" plus "long vowel symbol" and phenetic description /s/ and /ʃ/ were merged to "s", which may arise no confusion in Japanese. 50,000 phonemic data were sampled from sentences picked up from magazines, recent books, journals and novels.

As a measure of the data, entropy was calculated.

$$F_0 = -\log_2 \frac{1}{23} = 4.524 \qquad F_1 = -\sum_i P(i)\log_2 P(i) = 4.072$$

$$F_2 = -\sum_{ij} P(ij)\log_2 P_i(j) = 3.063 \qquad F_3 = -\sum_{ijk} P(ijk)\log_2 P_{ij}(k) = 2.620$$

Table 4 Distribution of elements of Japanese.

elements	probability	elements	probability
a	0.139	n	0.052
o	0.116	r	0.041
i	0.098	m	0.031
u	0.068	d	0.027
e	0.063	g	0.020
j	0.025	z	0.010
w	0.016	b	0.007
t	0.073	long vowel symbol	0.026
k	0.062	syllabic naals	0.025
s	0.055	space	0.013
h	0.013	assimilated consonant	0.012
p	0.003		

Above values indicate the same tendency with the the entropy of English. Table 4 shows the distribution of the elements. From the table it is observed that the total occurrence of 5 vowels reaches to extent of 0.484, which means that the Japanese language has the principal construction of "consonant+vowel", that "p" and "b" appeares seldom in Japanese speech and that about 5% of vowel is articulated as long vowel.

Table 5 Probability of occurence of phoneme sequences.

	group of phoneme sequence	probability (%)
1	V+VC+V	15.6
2	V+UVC+V	14.8
3	V+V	4.9
4	V+F	0.8
5	V+N	1.7
6	V+w+V	1.4
7	V+y+V	1.0
8	V+V+V	0.5

V ; vowel, VC ; voiced consonant
UVC ; unvoiced consonant
F ; assimilated consonant
N ; syllabic nasal

The numbe of possible combinations (i, j, k) of these 23 elements is $23^3 = 12{,}167$, but all of these cases do not occur. There are some combinations which do not appear at all. The trigram shows that 90% of data is concentrated to about 1,000 kinds of phoneme combinations. This may be thought that for recognition of phonemes by contextual method of three phoneme sequence, about 1,000 kinds of standard patterns must be

considered at least. This number is not so large compared with the possible number 23^3.

The trigram is scored for each group of sequences as shown in Table 5. The total percentage is 40.7% and the rest of these data has the combinations such as "consonant (c)+vowel (v)+consonant (c)" etc.. In Table 5 most sequences have the combination of V+C+V (or C+V+C), which are a basic structure of Japanese phonem sequence. Other types of sequence in Table 5 are deviated from this basic principle, and though the percentage of these types of sequence is low, they have importance and difficulty from the point of automatic recognition.

Based on these data, we have tried the contextual approach to vowel part and are now designing the equipment in which 100 standard patterns corresponding to three phoneme sequences are prepared and input pattern is refered to them. These data of trigram also give useful information to improve the recognized score.

7. Conclusoin

The system described here is a recognition model for Japanese speech sounds, which has the function of segmentation and of recognition of these segmented intervals. By taking the method by which the sound pattern is decomposed into elementary units (phonemes), input speech categorie is not limitted. The score for vowel is more than 90% and for consonant about 70%, though there left some phonemes that are not allowed as input at present.

This method of recognition of conversational speech sound is the first step trial and we are continuing to develop more efficient system. The statistics of phoneme sequence was proved to be very useful for the recognition by phonetic context and the new development is going on.

Acknowlegment

The authors wish to express their sincere thanks to Prof. Ken-ichi Maeda of Department of Electronic Engineering, Kyoto University, for his encouragement and support of this project. The authors owe very much to Mr. Kiyoshi Hashimoto of Electrical Communication Research Laboratory for his useful help in experiment, to Dr. Takeo Kurokawa, Mr. Haruo Tomonari, Mr. Toshiro Tsuzaki, Dr. Kunikazu Nagata, Mr. Hiroshi Sekimoto, Mr. Hisashi Kaneko, Mr. Yasuo Kato and Hiroshi Kondo of Nippon Electric Company Ltd. for their engineering design and construction of the whole equipment.

Reference

T. Sakai and S. Doshita "A Research Model of Japanese Monosyllable Recognition and Some Applications to Conversational Speech Analysis" STUDIA PHONOLOGICA I (1961)