SPEECH SYNTHESIS AND RECOGNITION BY COMPUTER

KENJI OHTANI

Department of Information Science
Kyoto University

July, 1972

SPEECH SYNTHESIS AND RECOGNITION BY COMPUTER

KENJI OHTANI

Department of Information Science
Kyoto University

July, 1972

PREFACE

With increasing improvements in information science, the pattern processing research has become highly popular which facilitates man-machine interaction. There is a growing demand for a new powerful pattern transmission and processing technique by which man can communicate directly in his natural tongue or in handwriting.

Of all information transmission mediums between man and machine, speech transmission is the best means that can presently be thought of, when viewed from the following considerations.

(1) Man is capable of making any abrupt interruptions to a machine by his own speech at any time under any circumstances.
(2) For the transmission and reception of speech, man needs no kinds of tools at all.
(3) From the viewpoint of practical applications, an economical and rapid transmission of information between machines and a large number of people can very effectively be achieved by voice transmission over a popular telephone network.

How, then, should speech be considered in the first place? Its information content is, as a rule, classified into two groups:

(1) linguistic information
(2) prosodic information

These can't necessarily be separated from each other. However, in most cases, the latter information is relevant to naturalness and is varied according to individual speakers, conversational environments and so-on. The former corresponds to the written equivalent of information contained in speech. This is particularly the main concern in the matter of speech transmission between man and machine. The prosodic information is also necessary for the natural speech transmission particularly from machine to man.

From this standpoint, such speech researches should be considered as follows:

(1) The vocoder aims at the efficient encoding, transmitting, and decoding of speech waves.

(2) Speech synthesis should approximately reproduce a speech sound wave from a compactly compressed linguistic and prosodic information.

(3) The main objective of speech recognition is to find out some efficient scheme which can get rid of redundancy in original speech and extract its essential linguistic information.

In this thesis, we are discussing new speech processing systems, such as a speech synthesis system and a speech recognition system, both of which can facilitate man-machine communication conveniently. The so-called zero-crossing analysis is commonly used in these systems.

In Chapter 2, we are discussing the fundamental properties of a zero-crossing wave and the potentiality of applying the zero-crossing analysis method to speech processings.

In Chapter 3 is presented a new speech synthesis method, which can synthesize any speech sentence in real-time in on-line mode. This synthesis scheme can be utilized for multi-speech output.

In Chapter 4 is discussed a new speech analysis-synthesis scheme, which can analyze original speech, transform it to a compressed information and lastly reproduce intelligible speech from the information. This can be used as a speech output system for a limited vocabulary of words.

In Chapter 5 is discussed a speech recognition system, which is connected to the speech analysis-synthesis system as will be described in Chapter 4. Phonemes in spoken words are recognized by this system which exhibits a rather good recognition performance.

July 1972
Kenji Ohtani

CONTENTS

4項欠

LIST OF SYMBOLS

Chapter 2

$P(s)$	:	transfer function of observed speech
$S(s)$	:	transfer function of voicing source
$T(s)$	:	transfer function of vocal tract
$R(s)$	:	radiation impedance
s	:	complex frequency
s_k	:	complex pole
F_k	:	k-th formant frequency
σ_k	:	damping constant of s_k
B_k	:	band width of k-th formant
T_0	:	pitch period
T_1	:	rising time of triangular volume flow
T_2	:	falling time of triangular volume flow
Ts	:	sampling interval
τ	:	zero-crossing interval
OX	:	abbreviation of "zero-crossing"
OXI	:	abbreviation of "zero-crossing interval"
mpa	:	abbreviation of "maximum peak amplitude" in a pitch
CH1	:	low frequency part of input speech
CH2	:	high frequency part of input speech
T	:	time window
N_i	:	number of zero-crossing points of i-th channel
$\overline{\tau_i}$	:	average zero-crossing interval of i-th channel
f_i	:	i-th extracted formant frequency
f_Z	:	extracted frequency parameter from the speech without any pre-filtering
A_0	:	peak amplitude of original speech in a frame
A_1	:	peak amplitude of CH1 in a frame
A_2	:	peak amplitude of CH2 in a frame
R_1	:	ratio of amplitude A_1 to A_0 in 100Xdb
R_2	:	ratio of amplitude A_2 to A_0 in 100Xdb
AMP	:	root mean square of A_1 and A_2

$\boldsymbol{x}$: vector representation of parameters set

Π_i : category i

C_i : event that vector $\boldsymbol{x}$ belongs to category Π_i

$P(C_i/\boldsymbol{x})$: probability that the vector $\boldsymbol{x}$ belongs to Π_i

μ_i : mean vector of category i consisting of five parameters

Σ_i : covariance matrix of parameters of category i

D_i : measure of distance of $\boldsymbol{x}$ from category i

Chapter 3

τ : zero-crossing interval

τ' : distributed zero-crossing interval

F_i : i-th formant frequency set in terminal analog model

P : wave-element

a : amplitude parameter of wave-element

r : repetition parameter of wave-element

a^{r} : segment (wave-element expression)

F^i_{ON} : i-th on-set formant frequency

F^i_{OFF} : i-th off-set formant frequency

V^i_{ON} : i-th formant frequency of preceding vowel

V^i_{OFF} : i-th formant frequency of succeeding vowel

$A^i_{ON}, B^i_{ON}, C^i_{ON}$: constants used in empirical formula which calculate F^i_{ON}

$A^i_{OFF}, B^i_{OFF}, C^i_{OFF}$: constants used in empirical formula which calculate F^i_{OFF}

d : information bit indicating whether the wave-element is voiced or not

b : information bit indicating whether the wave-element is plosive or not

OXI : abbreviation of "zero-crossing interval"

τ_i : i-th numerical zero-crossing interval in a wave-element

w : register setting τ_i

(V–V) : vowel-vowel context

(V–C) : vowel-consonant context

(C–V) : consonant-vowel context

CPU : abbreviation of "central processing unit"

Chapter 4 and 5

V : voiced frame or part

C : voiceless frame or part

F : voiceless fricative frame or part

X : silence frame or part

S : voiced steady frame or part

T : voiced transitional frame or part

CH1 : low frequency part of input speech

CH2 : high frequency part of input speech

A_1 : average amplitude of rectified speech of CH1

A_2 : average amplitude of rectified speech of CH2

R : ratio of amplitude A_1 to A_2

$\overline{\tau_1}$: average zero-crossing interval of CH1

$\overline{\tau_2}$: average zero-crossing interval of CH2

f_1 : extracted frequency parameter of CH1

f_2 : extracted frequency parameter of CH2

${}_if_1$: f_1 parameter in i-th frame

${}_if_2$: f_2 parameter in i-th frame

σ_1 : variance of logarism of f_1 in voiced part

σ_2 : variance of logarism of f_2 in voiced part

${}_ig_1$: logarism of ${}_if_1$ normalized by σ_1

${}_ig_2$: logarism of ${}_if_2$ normalized by σ_2

d_i : Euclidean distance between vectors $({}_ig_1, {}_ig_2)$ and $({}_{i-1}g_1, {}_{i-1}g_2)$

Z : voiced steady frame (temporary judgement)

D : voiced transient frame (temporary judgement)

P : wave-element

$\boldsymbol{\alpha}$: vector expression of amplitude parameters of wave-element

r : repetition parameter of wave-element

f_S : parameter calculated in the same manner as in f_1 and f_2 from the wave in CH2 passed through a low-pass operation

$\boldsymbol{x}$: vector expression of parameters in log scale

$\boldsymbol{\mu}_i$: average vector of $\boldsymbol{x}$ of category i

Σ_i : covariance matrix of parameters vector $\boldsymbol{x}$ of category i

$h_i(\boldsymbol{x})$: measure of similarity of $\boldsymbol{x}$ to category i

α : time change characteristic of A_1 in a slight longer time scale

β : time change characteristic of A_1 in a short time scale

8 項欠

Chapter 1 Introduction

It has for many years been a cherished dream of mankind to be able to communicate with machines by speech. Speech synthesis and recognition performed by machine, therefore, have been the main theme of speech engineering research from its beginning.

One of the objectives of such speech synthesis by machine is to investigate the phenomena of speech perception. The famous pattern play back project at Haskin's laboratory in the first half of 1950's[6] demonstrated numerous excellent results in this field. In particular, the experiments which showed the relation between voiced consonants and their second formant loci are very instructive. In the meanwhile, the MIT research group[56] developed another speech synthesis scheme based on the acoustic tube model in 1950's. In 1960, Fant[13] published the famous static vocal tract model, which has since contributed toward helping many speech synthesis researchers.

Since the beginning of 1960's, with increasing improvements in digital computer techniques, several kinds of speech synthesis schemes have been developed with the aid of the simulation programs and/or by the computer controlled machines. Thanks to these techniques, speech synthesis researches have made good progress and new practical speech synthesis researches including multi-speech output system have been demonstrated.

Speech synthesis schemes are estimated on the basis of the following aspects;

(1) quality of the synthetic speech.
(2) available multiplicity of the output terminals.
(3) scope of output vocabulary.

The speech output schemes reported hitherto have competed against each other concerning these specifications. Their methods exhibit a wide range of variation. However, speech output methods developed up to now may be classified into the following four groups:

(1) voice recording-reproducing method.
(2) compilation method.
(3) syntnesis method.

(4) analysis-synthesis method.

The voice recording-reproducing system reproduces continuous speech by connecting words or syllables recorded in analog or digital forms. In 1963, Dixon[10] developed a speech output scheme by using dyphone as a fundamental speech reproducing unit. There remains in this method a problem of introducing effects of co-articulation and prosodic information into the synthetic speech. When the problem is to be taken into account, the total memory size for storing the speech reproducing units increases rapidly.

The compilation method is based on the synthesis technique supplemented with the voice recording-reproducing technique. Matsui[30] proposed a method in which impulse responses of vocal tract systems in various shapes are used as the acoustical elements of compilation. The elements have been produced beforehand by means of the electrical vocal analog tube. On the other hand, Nakata et al[34] used damped sinusoids and band-pass noise as compilation units. Speech synthesis method as described in Chapter 3 of this thesis is one of the synthesis schemes, in which a sequence of zero-crossing intervals is used as compilation unit. Those methods for speech synthesis can usually synthesize any multiple speech sentences on line and in real time.

The speech synthesis methods are most popular, and in it the vocal-tract analog method and the terminal analog method are included. In the vocal-tract analog method[5][31], the tract behavior is simulated by the non-uniform acoustic tubes; ordinarily abutting right-circular cylinders represent the vocal and nasal tracts. On the other hand, the terminal analog speech synthesizer[23][40] simulates the tract behavior in terms of over-all transmission characteristics. The method attempts to duplicate the transmission properties of the tract as viewed from its input and output terminals. These methods have their strong points, not only in its application to practical use, but also in its employment as the materials of speech perception test. However, there still exists a difficulty as to how the synthesis circuits of consonants can be simplified. Even so these methods are said to be easily applicable to the terminals of the multiple-speech output system, with the increasing development of the integrated circuit technique.

The speech analysis-synthesis method has lately attracted considerable

attention. Input original speech is analyzed, transformed to a compressed information and stored in a computer. The speech sound is reproduced from the compressed information by reproducer. IBM[2] developed the vocoder type speech analysis-synthesis system. The PARCOR system by Itakura[22], and the speech analysis-synthesis by linear prediction by Atal[1] are used as a convenient tool of efficient coding of speech and of practical speech output. Matsui[32] has presented a new speech analysis method, which is based on Kalman filter theory. The method can be applied both to speech output and to speech recognition. In Chapter 4, we will describe a new speech analysis-synthesis scheme, which can easily be connected to a speech recognition system.

Traditional studies on speech recognition owe much to Sonagraph developed by Koenig[27] in 1946. In 1952, Davis et al[7], BTL, experimentally recognized spoken digits with 97∿99% accuracy of recognition with regard to one speaker, by measurement of zero-crossing points. Denes[8] and Fry[17], in 1959, presented recognition systems in which the spectral pattern matching method and the linguistic information were utilized together.

Since 1960's, the electronic computer has been used as a powerful tool. Denes et al[9] have obtained good results in the recognition experiments of a restricted vocabulary of words. Forgies[16], Lincoln Lab., developed a system which recognized vowels and voiceless consonants in some restricted contexts by the first and second formants extracted from the spectral patterns. Reddy et al[41] published a speech recognition system by computer without the aid of any special machine or devices.

Reddy[42] and Fant[14] has recently proposed new powerful speech recognition systems which contain the acoustical, phonological, syntactic and also semantic speech processing stages. In these systems, the main path toward higher levels of recognition is paralleled by a feedback path for enabling reexamination at lower stage. This approach would be promising, and will provide us with much progress in speech recognition by computer.

In our country, in 1963, Sakai and Doshita[43] developed a conversational speech recognition machine, in which zero-crossing measurement and some phonological rule were utilized. Kato et al[24] presented a spoken digits recognizer in 1964. In recent years, Chiba and Sakoe[53] proposed a new time-

normalization technique by dynamic programming in limited word recognition. Itahashi et al[21] tried to utilize a word dictionary positively in speech recognition.

The speech recognition system is generally estimated on the basis of recognition accuracy. However, the following specifications must be put together in its estimation:

(1) number of speakers (strictly limited or otherwise).
(2) scope of vocabulary.

Reddy[42] has pointed out the additional aspects such as the response time, memory size of the recognition program and cost problem. However, these aspects may be considered as the future problem.

The limited word recognizer can be used in, for example, a directive to the computer by speech, key punching by speech, automatic division of mail letters. When the recognizer is put to practical use, a rather high recognition accuracy is required, such as can bear comparison with punching accuracy by key punchers. The final goal of accuracy of such a recognizer will be three nine or so with regard to an unlimited number of speakers. On the other hand, the speech recognition system dealing with an unlimited vocabulary of words sets the goal of recognition to be 1,000 or 10,000 words in respect to some restricted number of speakers.

In Chapter 5 of this thesis, we will describe a new speech recognition system which recognizes the phonemes in spoken words, where the zero-crossing analysis is used. The system is connected to the speech analysis-synthesis system as described in Chapter 4. The composite system will propose a new speech research method.

The zero-crossing analysis method has been used in speech processing by many researchers, since 1948 when Licklider et al[29] demonstrated the intelligibility of the zero-crossing wave. In our studies, the zero-crossing analysis methods are used for the purpose of speech synthesis, analysis-synthesis and recognition system. In Chapter 2, we will show the fundamental experiments on the zero-crossing wave.

Chapter 2 Zero-crossing Analysis of Speech Sounds

2.1 Introduction

Speech synthesis and recognition are, in principle, to reduce the redundancy in speech and to express the acoustic properties of phonemes, effects of their coarticulation and traits of speakers, if necessary, in the form of a set of parameters. The study of speech synthesis and recognition is begun with the study of speech analysis, in which the properties of speech are investigated.

One of the most popular method taken as the pre-processing of speech is the spectral analysis, in which several kinds of filter banks are used. This analysis method is a simple and direct means to detect the envelope of the spectrum, and is utilized for the traditional sound spectrograph, too.

Another speech analysis method is the so-called zero-crossing wave analysis. Ever since Licklider[29] and others showed that up to 95% could be achieved in the clarity of syllables, zero-crossing sound waves have been used as a powerful tool for the study of speech. A zero-crossing wave is the wave that is obtained from an original waveform by amplifying it infinitely and limiting it to two valued amplitudes. In this waveform, the only sure information is the time when the waveform crosses the zero point and the only information obtainable from measurement is the interval between two successive zero-crossing time points.

One of the most favorite characteristics of this zero-crossing analysis method is that the measurements in this case are best suited to digital processing, And by the use of a relatively small number of wide band filters together with zero-crossing measures associated with each of these band, dynamic response penalties of narrow band filters would be reduced. By the infinite peak clipping of speech sounds, vowels are more distorted than consonants. From the fact that personal traits are included more in vowels than in consonants, clipped speech is less indicative of an individual speaker's voice traits.

Now, the handling methods of zero-crossing wave are classified into two

groups. One of them is the method by which the sequence of zero-crossing intervals (fine structure of zero-crossing waves)[54] is taken into consideration. Another one is the method by which the distribution of zero-crossing intervals or average zero-crossing intervals in a constant time interval (coarse structure of zero-crossing waves)[54] is considered to be parameters.

In general, it is very difficult to handle such zero-crossing waves mathematically and in particular, their fine structure. Scarr[54] and Teature[58] have tried to extract the formants by measuring the zero-crossing width under the high amplitude parts of the waveform within one cycle at larynx frequency. These trials are the analytical examples of the fine structure. However, because these methods need exact pitch detection, they are not reliable feature extraction methods. And yet it is doubtless that information is included in the sequence order of such zero-crossing intervals. In spite of the difficulty how to clarify the relation between the sequence order and the other parameters, the sequence order is sure to have the essential information in speech.

The coarse structure of zero-crossing waves is used together with a few broad band filters in extracting the formants. Sakai[43], Ewing[12] and Reddy[41] have used the analysis method for automatic speech recognition. It is possible to extract parameters from the coarse structure of zero-crossing waves by a simple device or a computer program. However, no sufficient results can be obtained by the sole use of these parameters. Also, it is difficult to construct any satisfactory recognition scheme from the hardware whose processing is limited to one direction of time. For this kind of automatic speech recognition, it is necessary to construct a recognizer by some computer program in which the other kinds of parameters, such as the amplitude information are used together.

In this chapter, we will discuss several kinds of results obtained from the zero-crossing analysis experiments. The fine structure of zero-crossing waves is described in section 2.3. In section 2.4, the coarse processing method and some recognition experiments by the method are presented.

2.2 Zero-Crossing Analysis of Synthetic Speech

2.2.1 Transfer Function of the Vocal System

The speech sound is the sound pressure output from "transmission networks" excited by several types of sources. Speech sounds are classified into two groups; voiced sounds and voiceless sounds according to their respective voicing sources. The voiced sounds of speech are produced by vibratory action of the vocal cords. The variable area orifice produced by the vibrating cords permits quasi-periodic triangular pulses of air to excite the acoustic system above the vocal cords. On the other hand, manners of speech output of voiceless sounds are classified into two groups. One of them is produced by a turbulent flow of air created at some point of stricture in the tract (for example, the source of the fricative sounds /s/ and /ʃ/). Another one is created by a pressure buildup at some point of closure (for example, the source of the plosive sounds /p/, /t/ and /k/). An abrupt release of pressure provides a transitory excitation. The air flow excited by such source is articulated in a particular manner according to the articulatory positions, for instance, the tongue, lips and velum. Thus, several kinds of speech waves are obtained.

In general, the observed speech P(s) is expressed as follows;

$$P(s)=S(s)\cdot T(s)\cdot R(s). \qquad (2.1)$$

Here, S(s), T(s) and R(s) mean the transfer functions of the speech source, the vocal tract and the radiation impedance respectively. Vowels, in particular, do not include zeroes in T(s). Therefore, vowels may be expressed as follows:

$$T(s)=\prod_{k=1}^{N}\frac{s_k\cdot\bar{s}_k}{(s-s_k)\cdot(s-\bar{s}_k)}, \qquad \left|\begin{array}{l} s_k=-\sigma_k+jw_k \\ \sigma_k=\pi B_k \\ w_k=2\pi F_k \end{array}\right. \qquad (2.2)$$

Here, s_k means the k-th complex pole. The imaginary part F_k of s_k means the resonant frequency which is called the formant frequency. The band width

B_k of the k-th formant is prescribed by the damping constant σ_k. On the other hand, the radiation impedance R(s) is approximated by the differential characteristic. The speech output P(s) may be obtained by the so-called terminal analog speech synthesis model, in which the triangular wave is input as the voicing source S(s) and the transfer functions T(s) and R(s) are simulated by filter bank. The block diagram of the model is shown in Fig. 2.1. Here, T_0 is a pitch period of the voicing source, and T_1 and T_2 are the rising and falling time respectively of the triangular glottal volume flow.

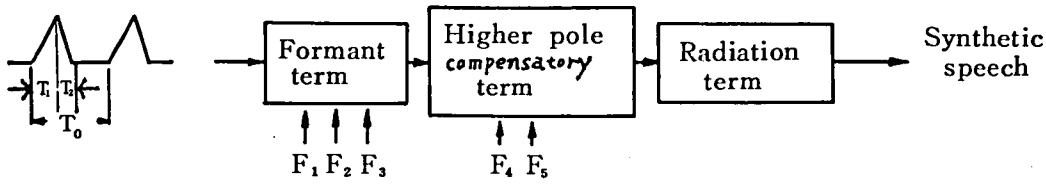


Fig. 2.1 Terminal-analog speech synthesis model

The complex poles are accounted for up to third formants (F_1, F_2, and F_3), and two fixed higher formants (F_4 and F_5) are added to compensate the effect of eliminating the higher poles. The transfer functions of the synthetic speech output P(s) is expressed as follows;

$$P(s) = S(s) \cdot \prod_{k=1}^{5} \frac{s_k \cdot \bar{s}_k}{(s - s_k)(s - \bar{s}_k)} \cdot s \qquad (2.3)$$

By setting the samplers between S(s), Π(···) and s, its z transform equals to,

$$P^*(z) = S^*(z) \cdot \prod_{k=1}^{5} \frac{(1 - 2e^{-\sigma_k T_s} \cos\omega_k T_s + e^{-2\sigma_k T_s})(1 - z^{-1})}{1 - 2e^{-\sigma_k T_s} \cos\omega_k T_s z^{-1} + e^{-2\sigma_k T_s} z^{-2}} \qquad (2.4)$$

Here, T_S is the sampling interval and is set to 10^{-4} sec.

In the following sections, we will investigate the relation between the transfer functions of the voicing system and its zero-crossing wave, by the simulation program of the synthesis model.

2.2.2 Effect of Voicing Source upon Zero-Crossing Intervals

The spectrum of the glottal volume flow is generally irregular and is characterized by numerous spectral zeroes. The taper of the spectrum is approximately 12db per octave. To a first order approximation, an epoch of peak amplitude of speech wave in a pitch period coincides with the greatest change in the derivative of the glottal waveform. For a triangular wave, for example, it would be at the apex.

The zero-crossing intervals seriously depend on the relative intensity of the lower two formants. However, when the relation is critical, the frequency and the shape of pitch play important roles for the patterns of zero-crossing intervals.

In this section, we will discuss the sequence of zero-crossing intervals of the synthetic vowels synthesized by the above-mentioned terminal analog model shown in Fig. 2.1. In Fig. 2.2, the output speech wave of vowel [e] and its sequence of zero-crossing intervals are presented for various duty factor $(T_1+T_2)/T_0$. The zero-crossing intervals are measured by counting the number of sampling pulses (T_S=100μsec) between the successive zero-crossing points. They are presented on the time axis in the Figure. Here, the ratio T_1/T_2 of the voicing source is fixed at 2. Formant frequencies F_1, F_2, and F_3 of [e] are set to 500 Hz, 2000Hz and 2600Hz, and also as the compensatory terms, formants, F_4 and F_5, whose frequencies are 3500 Hz and 4500 Hz respectively are added. Bandwidths of these formants $B_1 \sim B_5$ are 50Hz, 60Hz, 100Hz, 155Hz and 280Hz respectively. The pitch interval T_0 is set to 6.4 msec.

In all cases, the number of zero-crossing points in a pitch interval is even. In general, according as the duty factor of the voicing source decreases, the number of zero-crossing points increases, because the taper of the spectrum of the voicing source becomes more flat.

An epock of peak amplitude of speech wave shifts according the duty

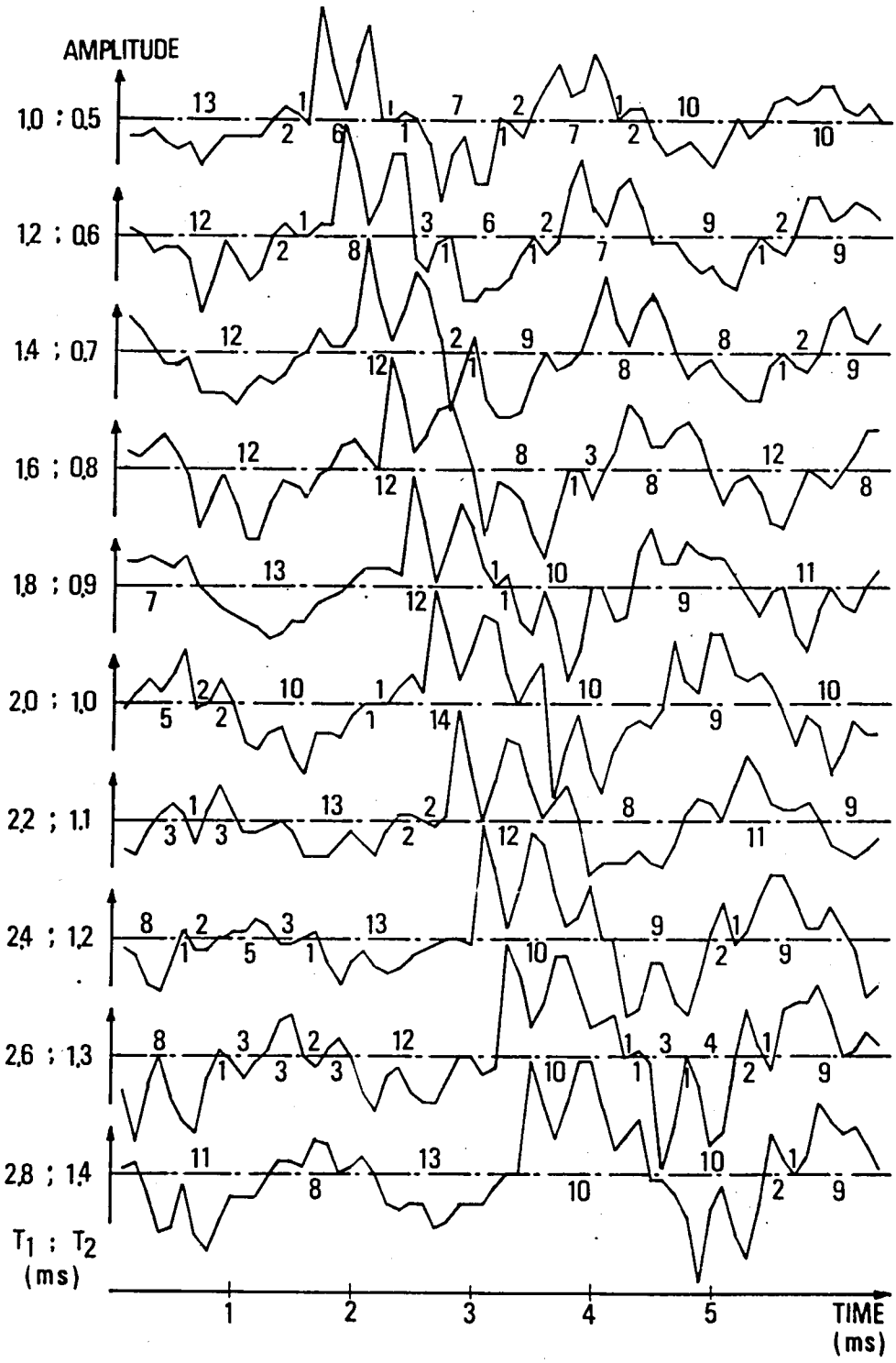


Fig. 2.2 Synthetic speech waveform of [e]

factor of the voicing source increases. The zero-crossing interval with maximum amplitude is about 10 samples interval, that is, 1 msec as shown in Fig. 2.2. In other words, half the inverse of the zero-crossing interval corresponds to the first formant frequency 500Hz. In the zero-crossing interval with high amplitude, the frequency components generally coincide in their phases, so that the most prominent spectral component, probably the first formant, has the strongest influence upon the zero-crossing interval. On the other hand, the zero-crossing intervals, in low amplitude part in particular, seriously depend upon the higher formants. In this part, the number of zero-crossing points depends on the relative **values of products of formant amplitudes and their respective frequencies. This fact agrees with what Teature[58] pointed out.**

2.2.3 Effect of Characteristics of Vocal Tract Resonances upon Zero-Crossing Intervals

Now, we are investigating the effect of the formants upon zero-crossing intervals. The simulation is carried out at 100Hz steps for the combination of the first and second formant frequencies (F1, F2), whereas the third formant frequency is fixed at 2700 Hz. Here, the pitch interval is 6.4 msec, and T_1 and T_2 are equal to 2.2 msec and 1.2 msec respectively. Fig. 2.3 (a), (b) and (c) show the zero-crossing pattern as a function of the second formant frequency for fixed first formant frequency; 300Hz, 600Hz and 900Hz. The second formant frequency is changed from 800Hz to 2500Hz by the step of 100Hz. As drawn, the distance τ between the contour lines represents a zero-crossing interval. One might expect half the inverse of zero-crossing interval $1/(2x\tau)$ to bear some relationship to the formant frequencies.

The figures show that the zero-crossing intervals gradually approach to the values estimated by the second formant frequency, according as the given F2 value increases. When the F2 value is small, the zero-crossing intervals are nearly equal to the value estimated by the first formant. However, these aspects are very much complicated. There are many mutually complicatedly related factors such as the phase relationship between the first formant and the second formant, and the effect of the higher order formants.

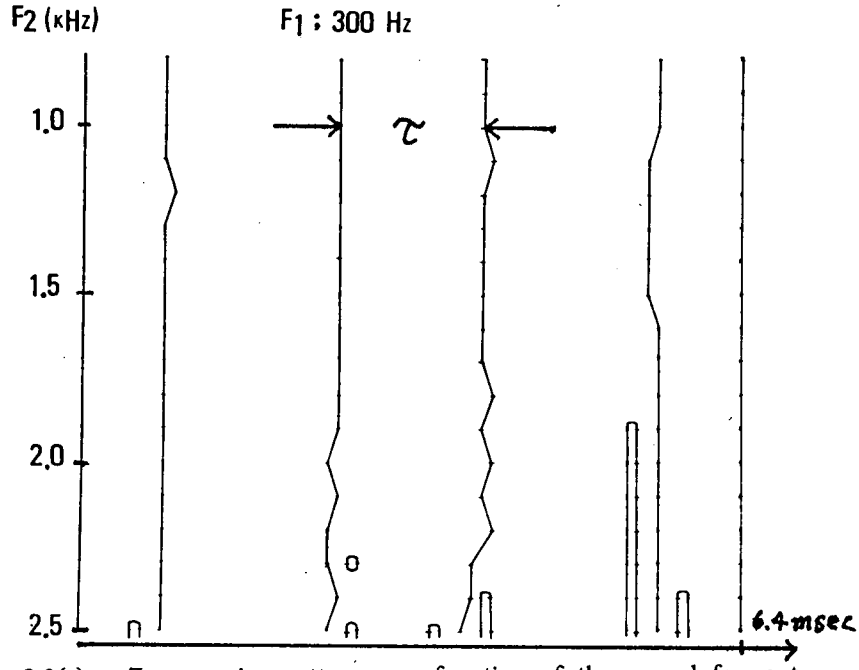


Fig. 2.3(a) Zero-crossing pattern as a function of the second formant frequency for fixed first formant frequency (F1=300Hz)

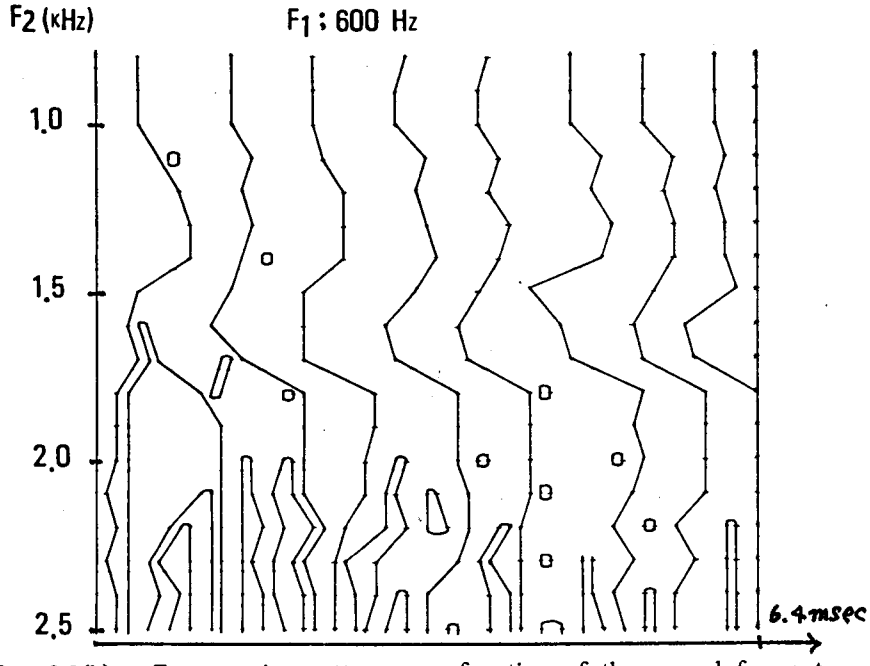


Fig. 2.3(b) Zero-crossing pattern as a function of the second formant frequency for fixed first formant frequency (F1=600Hz)

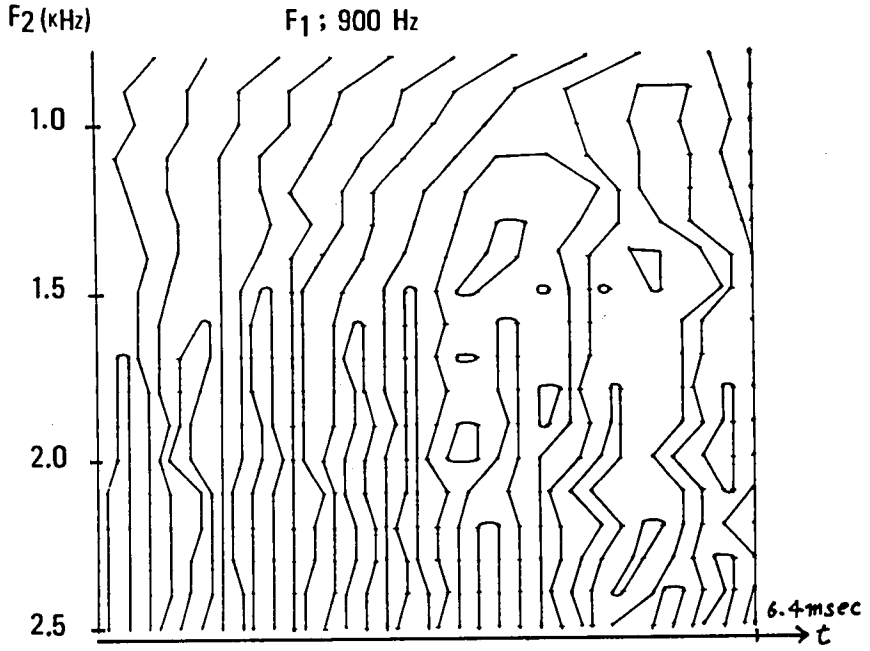


Fig. 2.3(c) Zero-crossing pattern as a function of the second formant frequency for fixed first formant frequency (F1=900Hz)

2.3 Sequence of Zero-Crossing Intervals

2.3.1 Pattern of Zero-Crossing Intervals of Natural Speech

It is most desirable for a speech researcher to be able to *show* the results of speech analysis in a visible form. The sound spectrogram is a well-known example which has been widely used for the display of a short-time spectrum. Thus, the speech research of the acoustic and perceptual features such as the formants, pitch and stress owes much to spectrograph. As the visible output method of the zero-crossing analysis, however, the intervalgram has also been developed by Chang et al[3], which gives the transitory distribution of the zero-crossing intervals in a short time window. In the intervalgram, the ordinate is the zero-crossing interval, the abcissa is time, and darkness of the pattern represents the number of frequencies of the zero-crossing interval in a

short time window.

It is true that the intervalgram is an effective means to represent the distribution of zero-crossing intervals, but for the output of the sequence of zero-crossing (OX) intervals in a visible form, still another method is necessary, which is shown in the upper half of Fig. 2.4, where the abcissa represents the order of occurence of zero-crossing points, and the ordinate is the zero-crossing

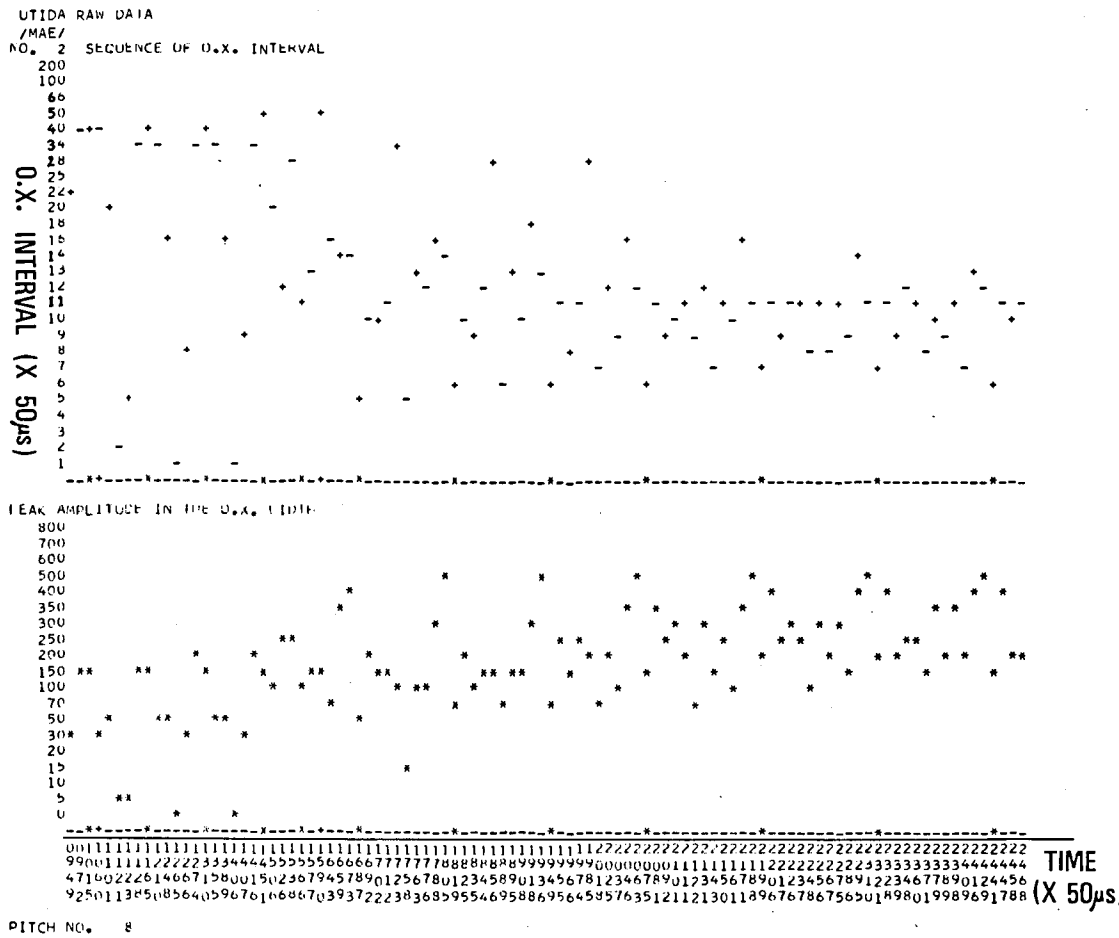


Fig. 2.4 Sequence of zero-crossing intervals of natural speech

intervals whose value is scattered into 26 steps. These scattered 26 values are decided so that half the inverse of the values equally separated. Symbols (+) and (-) represent the zero-crossing intervals having positive and negative amplitudes respectively. In the lower half of the figure are shown sequence of peak

amplitudes of the corresponding zero-crossing intervals. By such a figure, we can obtain a visible output of detailed zero-crossing pattern. Here, the star symbols (*) on the time axis show the result of automatic pitch extraction which will be described in the following section. It is the defect of this display that the abcissa is not the linear time scale. Then the time scale must be attached to each point of the horizontal axis, as is shown in the figure. The time scale is represented by the number of sampling pulses counted by 20 kHz sampler.

The example shown in Fig. 2.4 is a part of the transitional part between [m] and [a] in a word [mae](前). At the right half of the figure, a steady part of the vowel [a] is found, and in it a rather stable sequence of zero-crossing intervals is repeated at the larynx frequency (about 166 Hz).

2.3.2 Pitch Extraction from a Pattern of Zero-Crossing Intervals

The extraction of the period of the glottal source, that is, the pitch period is an inevitable and fundamental problem concerning the development of the general speech analysis-synthesis system, in addition to the vocoder which aimes at the compression of band width for speech transmission. Nevertheless, the problem does not come to a satisfactory solution. The main difficulty is that not only does the exciting glottal waveform vary in period and amplitude, but it also varies in shape, and the over-all spectrum is seriously affected by such variation. Besides, the wide range of pitch frequency variation attributable to the subjects, contexts and conversational environments, besides, is a reason for the difficulties of the pitch extraction.

Methods of pitch extraction are classified into two groups. One of them is a method in which the periodicity of the wave form is utilized. Another one uses the so-called cepstrum technique, which detects a frequency interval between abutting harmonics of frequency spectrum.

In this section, we are discussing the pitch extraction technique which utilizes the periodicity of sequence of zero-crossing intervals. The sequence of zero-crossing intervals in a sustained part of voiced sounds has a rather clear periodicity repeated at pitch intervals. Hence, by measuring the auto-correlation

of the sequence, we can detect the pitch intervals in order. The flowchart of the pitch extraction program is shown in Fig. 2.5. The explicative figure is given in Fig. 2.6. The horizontal line is a time axis, under which a sequence of OXI is shown by number of sampling pulses. The input speech materials are sampled at 20kHz and fed into a computer, without any pre-emphasis and filtering process. The extraction method is as follows:

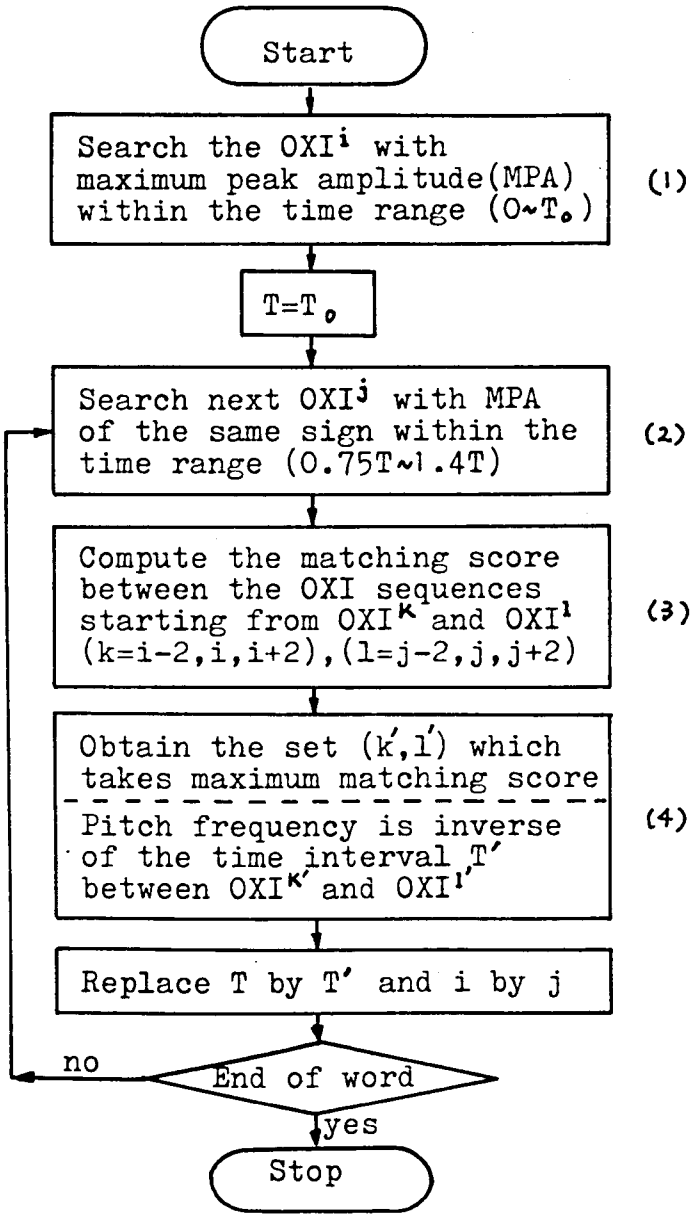


Fig. 2.5 Pitch extraction procedure

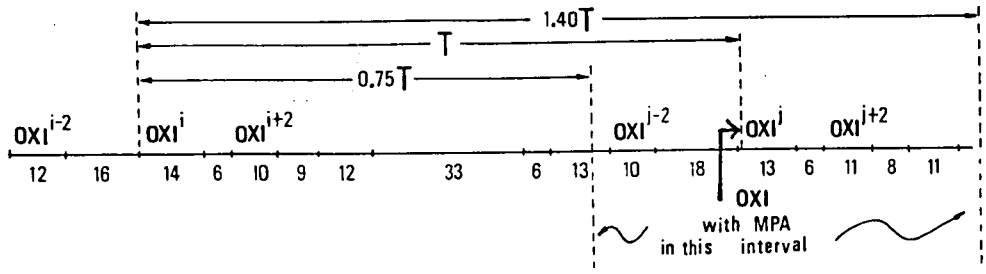


Fig. 2.6 Explicative figure of the pitch extraction

(1) In the initial T_o time interval, the zero-crossing interval with maximum peak amplitude is searched (hereafter, the amplitude is abbrebiated as mpa). Now, let it be the i-th zero-crossing interval (expressed as OXI^i), as is shown in Fig. 2.6.

(2) The zero-crossing interval (OXI^j) with "mpa" is detected from the range $0.75T \sim 1.4T$ (T is the preceding pitch interval) measured from the i-th zero-crossing interval. Here, (j-i) is an even number.

(3) The matching score (correlation score) is computed between the sequence OXI^k, OXI^{k+1} ..... $OXI^{\ell-1}$ and the zero-crossing waves with OXI^{ℓ}, $OXI^{\ell+1}$, ... Here, k is either "i-2", "i" or "i+2", and ℓ is either "j-2", "j" or "j+2". In such a way, nine matching scores are obtained.

(4) Let (ℓ', k') be the combination which has the greatest matching score. The pitch is the time interval T' between the $OXI^{k'}$ and $OXI^{\ell'}$. T is substituted by this time interval T' and the above-mentioned processes (2), (3) and (4) are continued until a given word is completed.

In Fig. 2.7(a), (b), (c) and (d), the extracted pitch frequencies and the corresponding matching scores (-100%∼100%) are presented. The loci of pitch frequencies obtained in such a way are rather smooth. Pitch frequencies gradually fall down at the end of the utterances. Comparatively low pitch frequencies are shown in the voiced consonants such as [d_3] in Fig. 2.7(c) and [m] in Fig. 2.7(b). There is very low matching score in the voiceless parts such as the voiceless burst in Fig. 2.7(d), the fricative part in Fig. 2.7(c)

which have no periodicity of the sequence; while, almost all parts of voiced sounds have more than 50% matching scores. Therefore, it is possible to discriminate the **voiced sounds from the voiceless sounds by the matching score.**

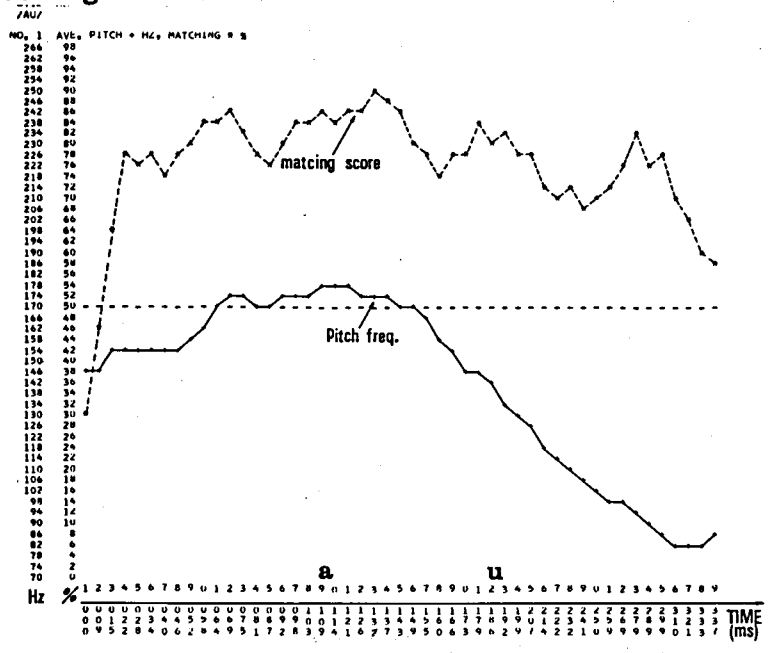


Fig. 2.7(a) Example of pitch extraction [au] (会う)

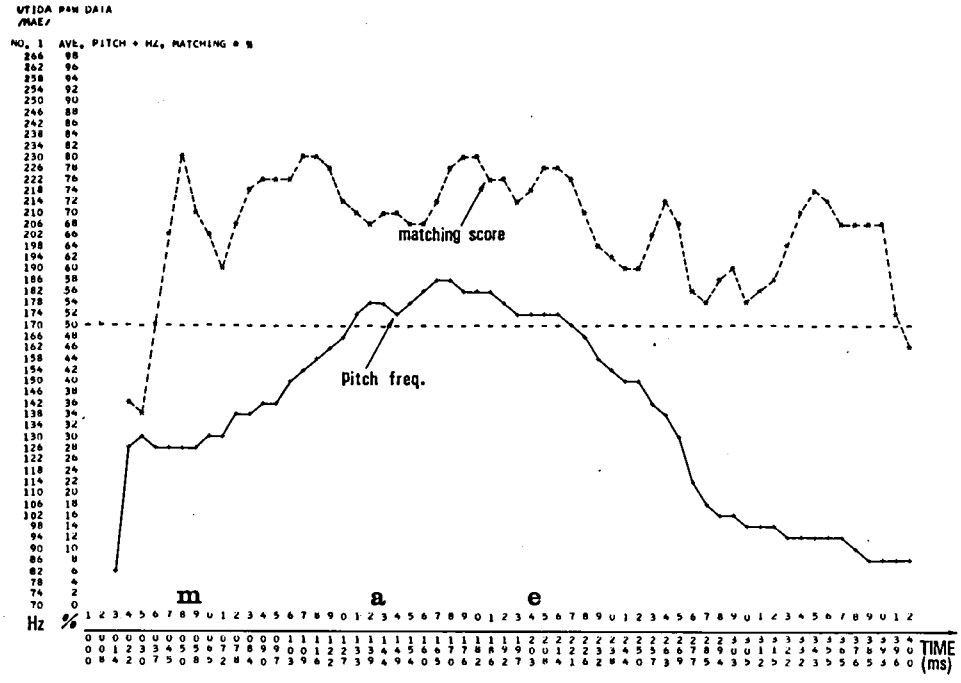


Fig. 2.7(b) Example of pitch extraction [mae] (前)

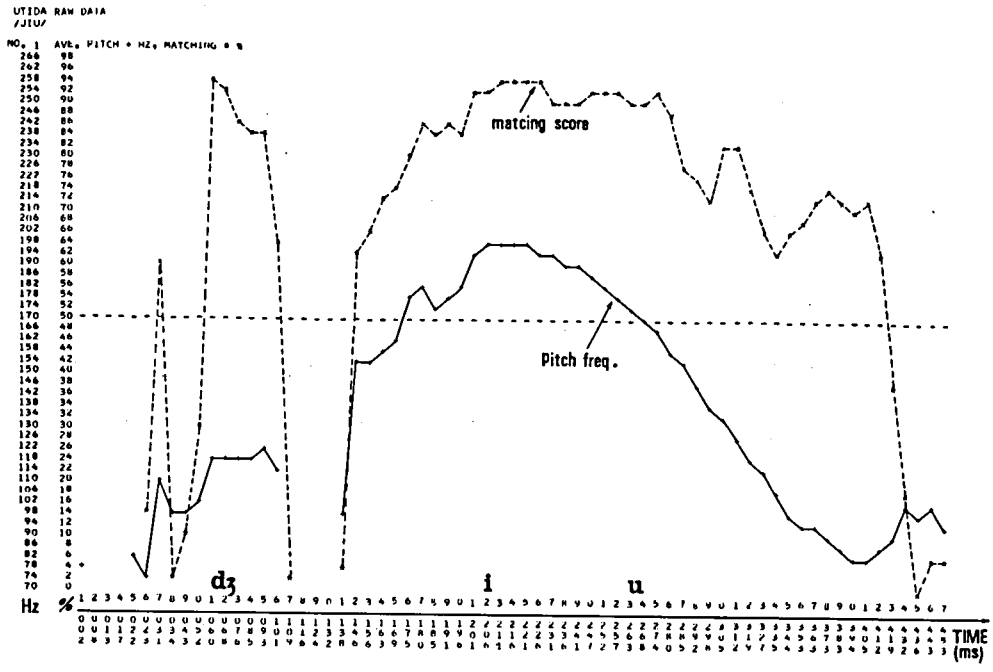


Fig. 2.7(c) Example of pitch extraction [jiu] (滋雨)

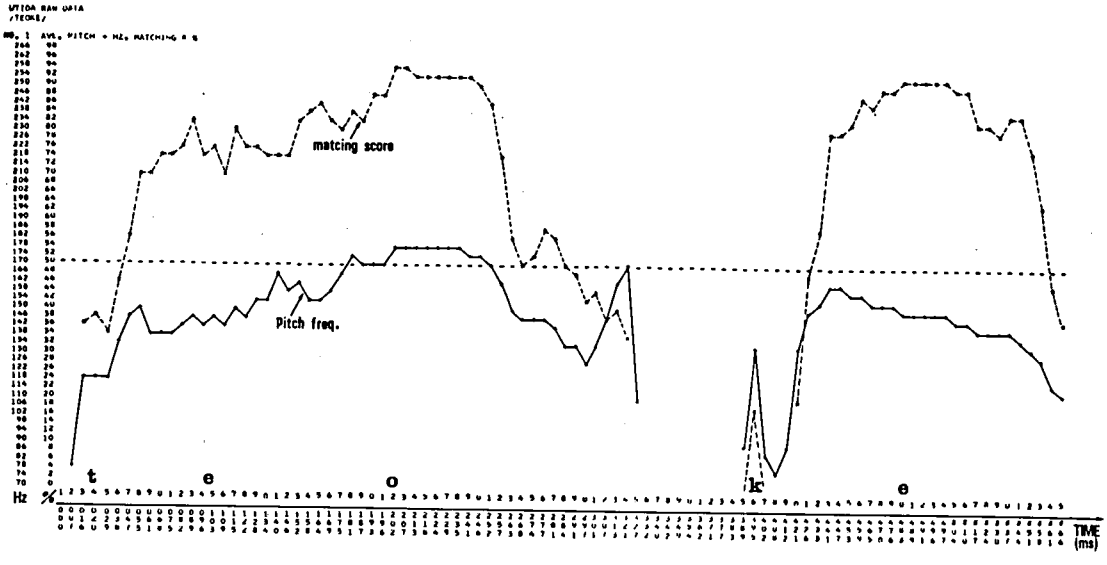


Fig. 2.7(d) Example of pitch extraction [teoke] (手桶)

2.3.3 Expression of Speech by Wave-Elements[45][49][50]

Speech waves in the steady parts of vowels and voiced consonants have a rather precise periodicity. The same is also true of the sequence of zero-crossing intervals in such the parts. From the consideration of the sequence in a pitch period as the fundamental information bearing units in these stable voiced parts, the high grade of information compression can be obtained.

In fricative sounds, zero-crossing points occur at random and the distribution of zero-crossing intervals is stable. Besides, the duration of such part is very long. It has been experimentally assured that the original zero-crossing wave is approximately reproduced by arranging at random the set of zero-crossing intervals in a short time period.

In other words, it has been ascertained that the speech can be expressed with the compressed information under rules, such as the repetition and re-arrangement of the sequence of zero-crossing intervals. We, henceforth, will call the fundamental sequence of zero-crossing intervals used in these rules, as the wave-element. The notion of this wave-element is frequently used in the speech synthesis by rule in Chapter 3 and in the speech analysis-synthesis system in Chapter 4.

2.4 Distribution of Zero-Crossing Intervals

2.4.1 Formant Extraction by Average Zero-Crossing Intervals

Voiced sounds are produced exclusively by vocal cord excitation of the vocal tract. The speech output is damped sinusoid repeated at the pulse rate. The envelope of the amplitude spectrum has its maximum at a frequency essentially equal to the resonance frequency of the tract.

The resonances of the vocal tract are, of cource, multiple. The output time waveform is therefore a superposition of damped sinusoids and the amplitude spectrum generally exhibits multiple peaks. If the individual resonances can be suitably isolated, say by appropriate pre-emphasis circuits or by filtering,

measurement of zero-crossing intervals might be a useful indication of formant frequency.

Chang et al[4] have demonstrated mathematically and experimentally that the average rate of zero-crossing of the undifferentiated speech wave and of differentiated wave is a measure of the first formant frequency and the second formant frequency respectively.

Scarr,[54] Ewing et al[12] and other researchers have used the zero-crossing technique for formant extraction, in which the speech signal is pre-filtered into frequency ranges appropriate to individual formants.

Another zero-crossing technique for extracting the formant were proposed by Teature et al[58] and Scarr.[54] They have indicated that measurement of zero-crossing intervals under the high amplitude parts of the waveform within one cycle at larynx frequency might give a measure of the center of spectral energy. These pitch synchronous analysis, however, have their disadvantage to have to extract the accurate pitch periods.

Among various zero-crossing techniques, such as the above-mentioned attempts, the average zero-crossing measurement of the wave which is passed through several appropriate filters might be the most promising method for its stability and simplicity. However, the disadvantage remains that the method is still subjected to the overlapping of the formant frequency range. In other words, the setting of the pass-bands of filters remains to be an difficult problem.

The pattern of the time intervals between consecutive zero-crossings cannot easily be analyzed theoretically. Theoretical approach to only two component sinusoidal waves has been tried by Scarr[54] and Doshita[11]. But it is still difficult to extend this theoretical analysis to a signal consisting of three or more components, and no trial in this line has ever been made. The cut-off frequencies of filters must be experimentally decided by cut-and-try methods.

2.4.2 Device for Formant Extraction

Vowels are characterized by several formant frequencies which correspond to resonance frequencies of the vocal tract. In particular, it is well known that

the lower two or three formants are very important features for speech recognition. Hence, when the formant frequencies are to be extracted by the zero-crossing measurements, speech signals must be divided into some formant frequency bands by broad band-pass filters. The pre-emphasis circuit is ordinarily used together in order to average the slope of amplitude spectrum of the input speech.

Consequently, we have constructed a speech pre-filtering circuit for the formant extraction as shown in Fig. 2.8. CH1 and CH2 mean the output of

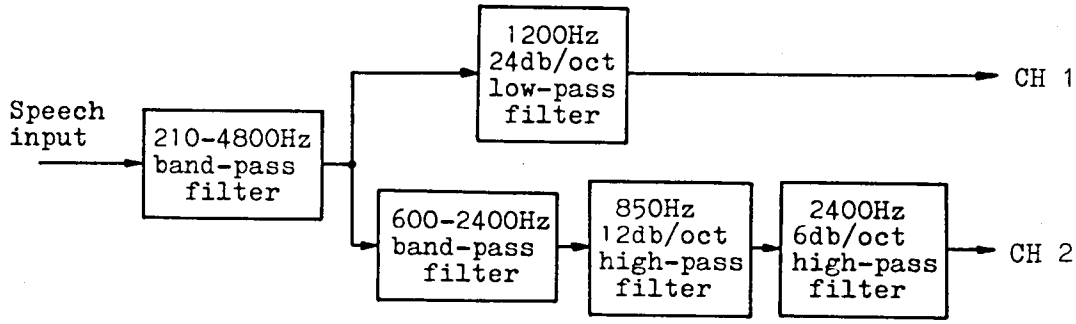


Fig. 2.8 Speech pre-filtering circuit

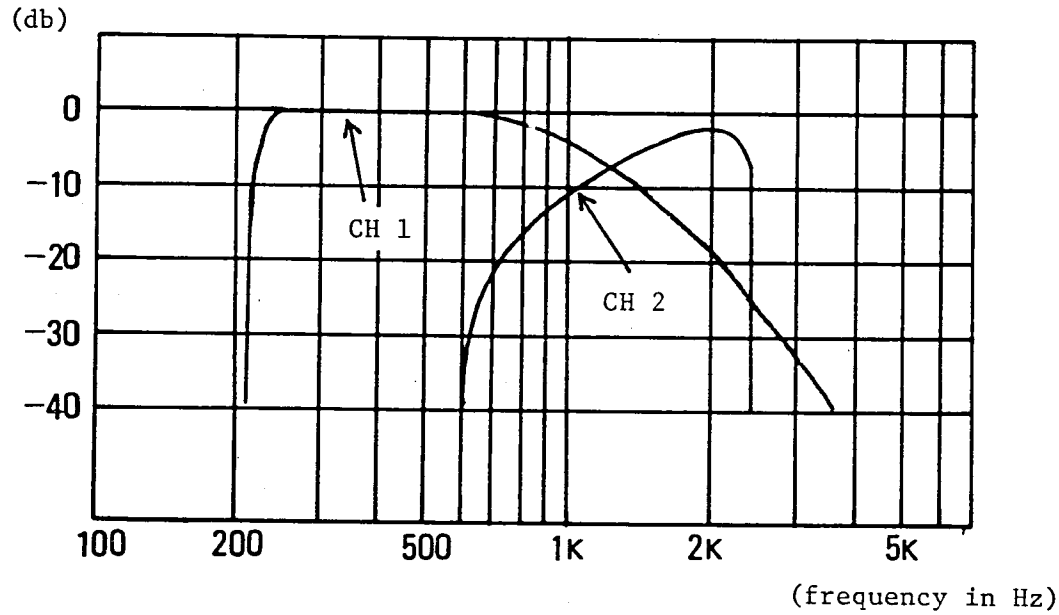


Fig. 2.9 Frequency characteristics of the speech pre-filtering circuit

the first and the second formant frequency bands respectively. The frequency characteristics of CH1 and CH2 are illustrated in Fig. 2.9. The abcissa represents the frequency in log scale and the ordinate is the amplitude in decibel scale. CH1 is the output of the serial concatenation of two second order low-pass filters (1.2 kHz cut-off, 12 db/oct.), and CH2 is the output of the serial concatenation of the sharp cut-off filter (2.4 kHz cut-off) and two second order high-pass filters (850 Hz cut-off, 12db/oct. and 2.4 kHz cut-off 12db/oct.). Other bands were also tentatively used, but the above-mentioned settings seemed to do as well or better than any others and are found to be more reasonable for the extraction of lower two formants.

Speech signals passed by these circuits are sampled at a rate of 10,000 samples per second. At each sample time, the amplitude of the signal was converted to a 11 binary digit number, which was then written on a auxiliary memory of a computer. As a first step in processing, the digitized speech was divided into the abutting constant time windows. The average zero-crossing time interval $\overline{\tau_i}$ is obtained by counting the number of the zero-crossing points N_i in the time window T, as follows.

$$\overline{\tau_i} = \frac{T}{N_i} \qquad (i=1, 2), \tag{2.5}$$

Here, the subscript i represents the channel number. Formant frequencies f_i are given by

$$f_i = \frac{1}{2\overline{\tau_i}} = \frac{N_i}{2T}. \tag{2.6}$$

In order to test the characteristics of the speech pre-filtering circuit, we have analyzed the synthetic speech by this device and extracted formant frequencies. Here, we have adopted as voicing source the impulse being passed through two integrators so that the spectral zero has not appeared. And pitch period T is 6.4 msec. In Table 2.1 and 2.2 are listed the formant frequencies, f_1 and f_2, which are obtained from the outputs of CH1 and CH2 respectively.

The extracted frequencies f_1 and f_2 are multiple of 1/T, that is, 156 Hz because N_1 and N_2 are even numbers. They are approximately consistent with F1 and F2 values given in the synthesis model. When F_2 closes to the boundary region of CH1 and CH2, the extracted frequencies f_2 show a tendency to fluctuate above and below the F_2 values; while rather stable f_1 values are extracted in almost all combination of F_1 and F_2 values.

Table 2.1 Extracted f_1 parameter

F_2 (Hz) ↑ / f_1 extracted (Hz)								
2500	312	625	625	625	1250	781	937	1093
2400	312	468	468	625	781	781	937	1093
2300	312	468	468	625	781	781	937	1093
2200	312	468	468	625	781	781	937	1250
2100	312	468	468	625	781	781	937	937
2000	312	468	468	625	781	781	937	1093
1900	312	468	468	625	781	781	937	1093
1800	312	468	468	625	781	781	937	937
1700	312	468	468	625	781	781	937	937
1600	312	468	468	625	781	781	937	1093
1500	312	468	468	625	781	781	937	937
1400	312	468	468	625	781	781	937	937
1300	312	468	468	625	781	781	937	1093
1200	312	468	468	625	781	781	937	937
1100	312	468	468	625	781	781	937	1093
1000	312	468	468	625	781	781	937	
900	312	468	468	625	781	781		
800	312	468	625	625	781			
	300	400	500	600	700	800	900	1000 → F_1 (Hz)

F_2 in synthesis model

F_1 in synthesis model

Table 2.2 Extracted f_2 parameter

F_2 (Hz) / f_2 extracted (Hz)								
2500	2500	2500	2500	2500	2500	2500	2500	2500
2400	2343	2343	2343	2343	2343	2343	2343	2343
2300	2343	2343	2343	2343	2343	2343	2343	2343
2200	2187	2187	2187	2187	2187	2187	2187	2187
2100	2031	2031	2031	2031	2031	2031	2031	2031
2000	2031	2031	2031	2031	2031	2031	2031	2031
1900	1875	1875	1875	1875	1875	1875	1875	1875
1800	1875	1875	1875	1718	1718	1562	1718	1718
1700	1718	1718	1718	1718	1718	1718	1718	1718
1600	1562	1562	1562	1562	1562	1562	1562	1562
1500	1562	1562	1562	1406	1406	1250	1250	1406
1400	1562	1406	1406	1406	1406	1406	1406	1406
1300	1562	1406	1250	1406	1250	937	1093	1250
1200	1250	1250	1250	1093	1250	937	1093	1093
1100	1250	1250	1093	1093	1093	1093	1093	1093
1000	1406	1093	1093	781	937	937	937	
900	1250	937	937	937	937	781		
800	1093	937	937	1093	781			
	300	400	500	600	700	800	900	1000

F_2 in synthesis model

F_1 in synthesis model

F_1 (Hz)

By the zero-crossing measurement of the original synthesized speech, we can define f_Z frequency which may represents the total effects of whole spectral components and in particular, the lower three or four formants. Extracted f_Z frequency is listed in Table 2.3. The f_Z values have a slight tendency to be higher than the average of F_1 and F_2 values set up in the model.

Table 2.3 Extracted f_z parameter

F_2 (Hz) f_z extracted (Hz)

F_2 in synthesis model	F_1 = 300	400	500	600	700	800	900	1000
2500	2734	2578	2578	2500	2500	2500	2578	2578
2400	2812	2890	2734	2734	2734	2890	2812	2734
2300	2500	2421	2656	2578	2656	2656	2656	2578
2200	2265	2578	2421	2734	2578	2734	2500	2734
2100	1640	2578	2265	2343	2812	2109	2421	2265
2000	1562	2265	2500	2343	2578	2656	2578	2812
1900	1796	2578	2109	2265	2421	2031	2421	2343
1800	1250	2109	1953	2421	2421	2265	2421	2421
1700	1328	2109	1875	1640	1953	2031	2265	2187
1600	781	1796	1953	2109	2109	2109	2109	2421
1500	859	1953	1875	1640	1953	1484	1562	1796
1400	1015	1718	1562	1484	1718	1718	1640	1562
1300	781	1328	1015	1171	1640	1562	1484	1640
1200	625	1640	781	859	1718	1093	1328	1328
1100	625	1093	1015	1015	1328	1093	1171	1093
1000	781	1171	781	468	937	937	1015	
900	625	1250	781	703	1015	781		
800	781	781	859	859	859			

F_1 (Hz)

F_1 in synthesis model

Generally, they fluctuate in a wide range of frequency. Effects of higher formants can be found in this parameter.

2.4.3 Speech Material

We have tried the formant extraction of natural speech. The samples of this experiments are phonemes in spoken words: five vowels (/a/,/i/,/u/,/e/ and

/o/), voiceless fricative sounds (/s/ and / ʃ /) and also nasals (/m/ and /n/).

Nine-hundred words uttered by 20 males including five radio announcers were used for this experiment. Forty-five words spoken by each person are shown in Table 2.4.

Table 2.4 List of words used in the experiment

V–V	naate	hai	au	mae	kao
	kiatsu	suii	jiu	chie	shio
	suashi	ruiji	suu	sue	uo
	teate	teire	neuchi	maee	teoke
	soaku	koi	ou	koe	anooka
V–s–V	asa	ise	musume	seseragi	osoi
V–ʃ–V	ashi	ishi	mushi	deshi	hoshi
V–m–V	ama	imi	umu	seme	tomo
V–n–V	ana	hinichi	kunugi	genetsu	ono

2.4.4 Extracted Parameters and their Statistical Properties

We have extracted the frequency parameters f_1, f_2, and f_Z and also R_1 and R_2 parameters which indicate the ratio of amplitudes. R_1 and R_2 are defined by

$$\begin{cases} R_1 = 2000 \times \log_{10} \dfrac{A_1}{A_0} \\ R_2 = 2000 \times \log_{10} \dfrac{A_2}{A_0} \end{cases} \qquad (2.7)$$

Here, A_1 and A_2 are the peak amplitudes of CH1 and CH2 respectively in the time window. A_0 is the peak amplitude of the original speech wave in the same time interval.

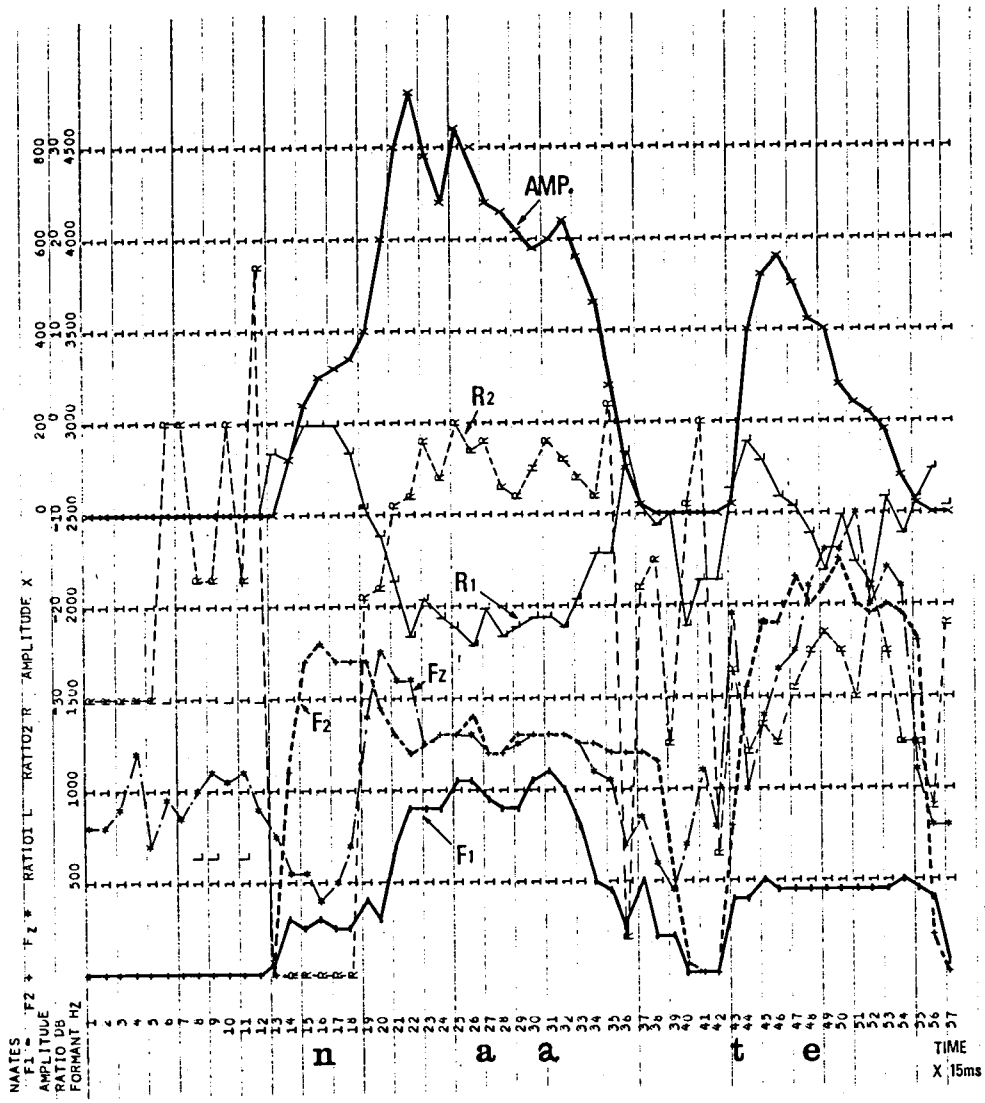


Fig. 2.10(a) Example of extracted parameters [naate](名宛)

Examples of the extracted parameters are shown in Fig. 2.10(a), (b), (c), and (d). The time axis is quantized by a time window (15 msec.), which we henceforth call a frame. The f_1 parameter of [a] in Fig. 2.10(a) has a slight tendency to be higher than the actual first formant frequency because the effect of the second formant frequency is found in the frequency region of CH1; while the f_2 parameter of the former [o] in Fig. 2.10(d) is higher than the actual second formant frequency because the second formant exists in the lower boundary region of CH2 and the third formant contributes to the extracted value. Fricative sound [s] has a

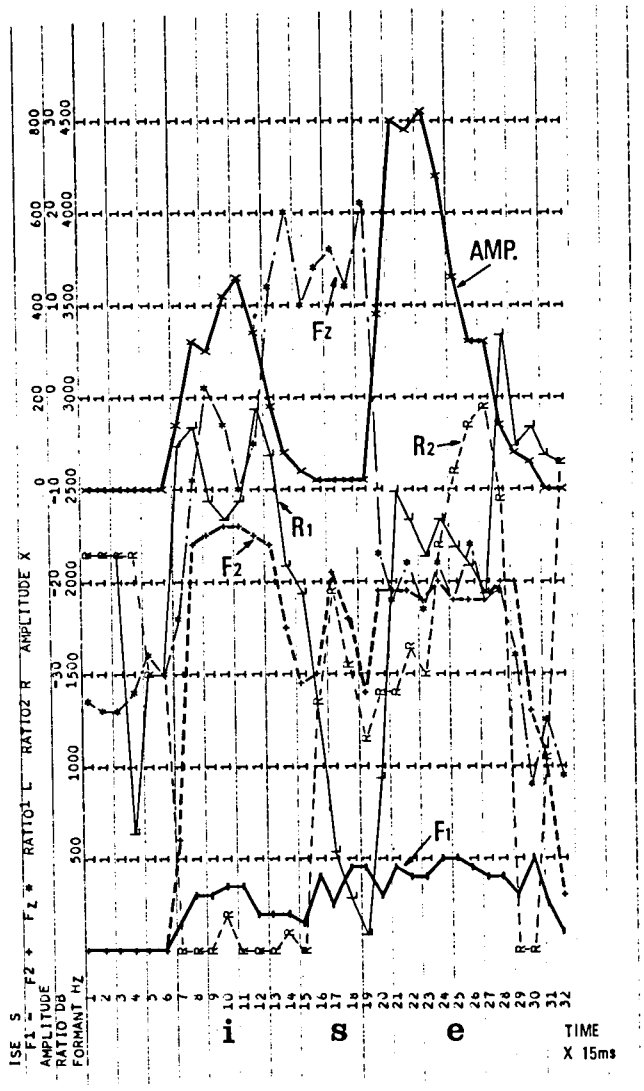


Fig. 2.10(b) Example of extracted parameters [ise] (伊勢)

large f_Z value as is shown in Fig. 2.10(b). Nasal [n] in Fig. 2.10(d) has a comparatively large f_2 value. In the transitional part between the consonant and neighboring vowels, smoothly changing f_2 loci are found. Parameter AMP is calculated by root mean square of A_1 and A_2, which is a powerful parameter to discriminate between vowels and consonants.

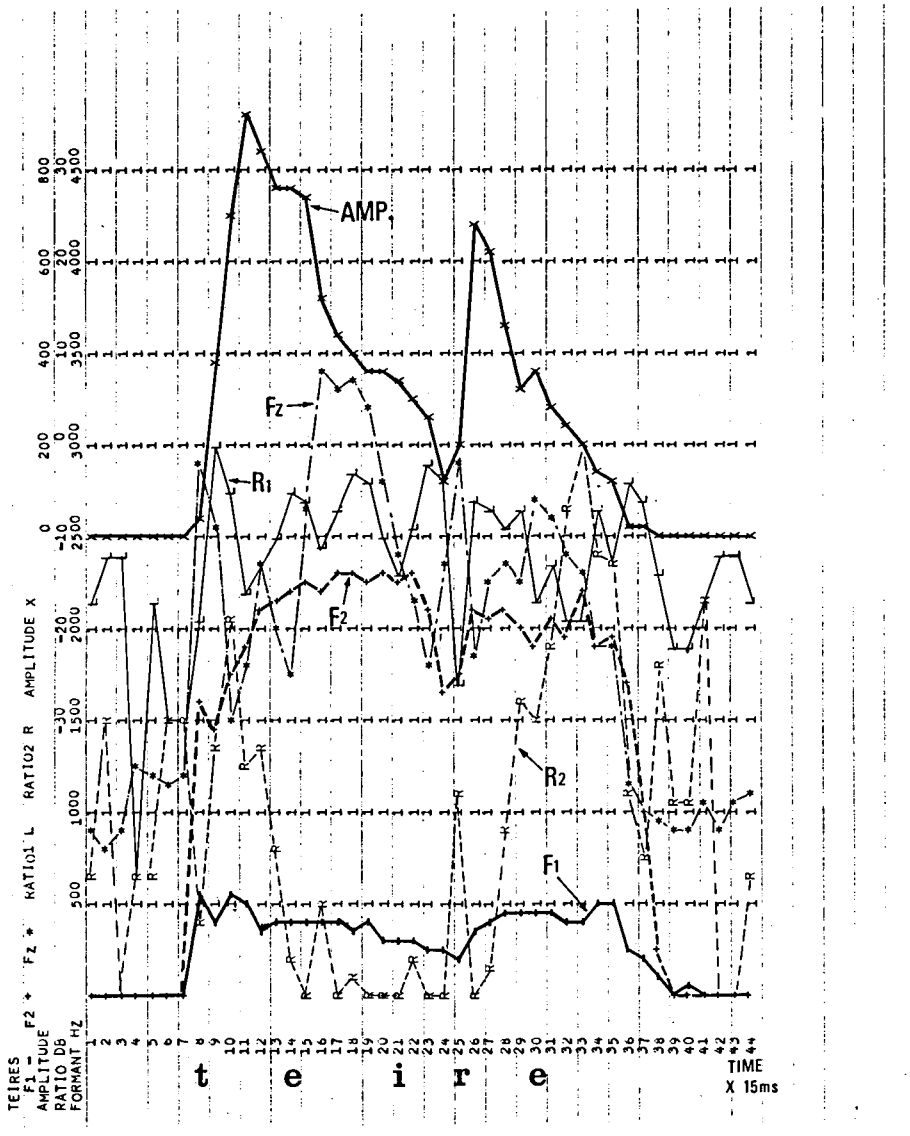


Fig. 2.10(c) Example of extracted parameters [teire] (手入れ)

As the materials for individual phonemes used in the following experiments, we have chosen three successive sets of parameters from the retentive parts of phonemes in spoken words. A set of five parameters (f_1, f_2, f_Z, R_1, and R_2) is considered as a point in five dimensional space.

Average parameter values of selected samples (a set of five parameters) for vowels and two consonant groups are shown in Table 2.5. Covariance matrices

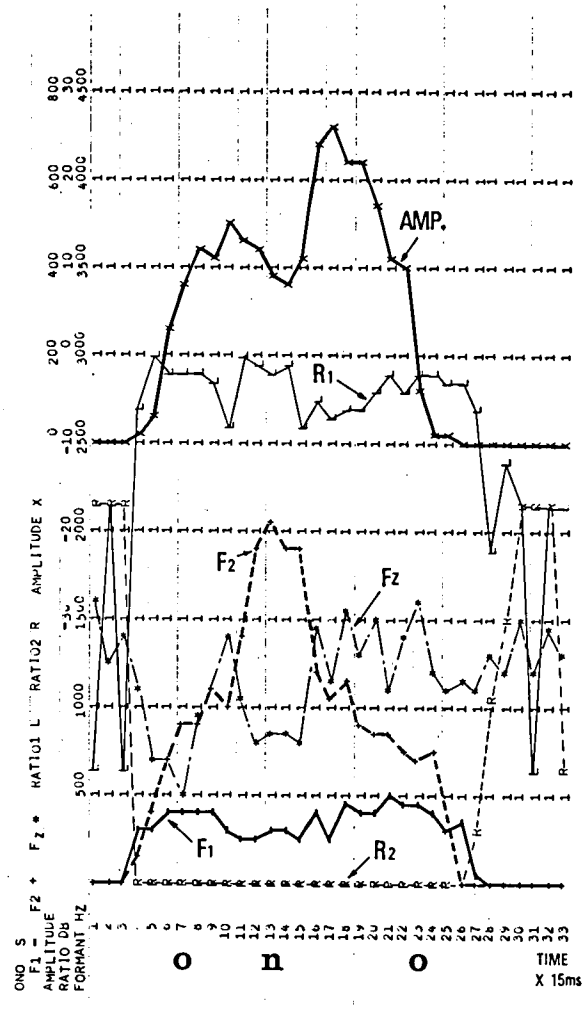


Fig. 2.10(d) Example of extracted parameters [ono] (斧)

Table 2.5 Average of extracted parameters

Category

Parameter	/a/	/i/	/u/	/e/	/o/	/s/	/n/	
f_1	805.1	297.7	352.6	498.3	524.2	321.1	285.8	(Hz)
f_2	1285.2	2176.8	1382.0	1933.5	964.8	1712.5	1630.9	(Hz)
fz	1449.3	2418.5	1283.8	1875.0	1126.0	3626.0	1359.4	(Hz)
R_1	-376.7	-301.0	-94.3	-360.5	-110.6	-1593.4	-141.5	(db×100)
R_2	-210.9	-1159.4	-1161.1	-273.6	-1236.5	-912.0	-1043.1	(db×100)

for these categories are shown in Table 2.6. Because the covariance matrices are symmetric, left half entries are omitted. Meaningless values are obtained for average f_1 and f_2 of voiceless consonants. Because the frequency components higher than 2.5 kHz is cut off in CH2, extracted f_2 value of vowel /i/ is slightly lower

Table 2.6 Covariance matrices of extracted parameters

Σ_a

2075.5	-2474.0	-5703.0	-2592.7	11667.8
	17956.0	11953.0	-6156.8	-11224.4
		83840.0	-10102.0	-2110.3
			13696.2	-5992.8
				45022.8

Σ_i

2587.7	-123.0	-1263.8	-1296.8	5404.4
	23535.0	22840.0	-3766.8	-6805.0
		293943.0	-69535.9	-8549.0
			49986.4	-23020.4
				229855.0

Σ_u

4902.3	-217.7	1655.6	-783.2	5529.3
	95162.0	49269.0	-9677.0	75798.0
		334473.0	-28152.2	130612.0
			7994.9	-24211.3
				226165.0

Σ_e

5549.3	-2179.9	2148.0	-5398.4	6019.4
	18044.0	14142.0	1361.8	-3047.7
		127525.0	-37168.9	28648.0
			56868.2	-28417.8
				74893.3

Σ_o

7013.2	114.3	6242.7	-1642.4	11486.9
	16562.0	11336.0	-3064.9	23354.0
		225367.0	-17712.3	56447.0
			5306.3	-13427.8
				144548.0

Σ_s

65113.8	43781.2	15384.0	-40778.8	-2231.0
	143303.0	45324.0	-109681.0	-48825.0
		194040.0	-176110.0	-162914.0
			412832.0	286891.0
				327625.0

Σ_n

3643.1	6217.3	3566.0	-1163.7	6867.3
	210310.0	53981.0	-9152.9	58510.0
		295725.0	-41898.7	149061.0
			35551.0	-28340.1
				228425.0

than the actual second formant frequency. It is comprehensible by the Table 2.6 that the variances of f_1 and f_2 of vowel /a/ are rather small and that these samples of /a/ are brought together into a small region of the dimensional space. On the other hand, variance of f_2 of nasals is very large.

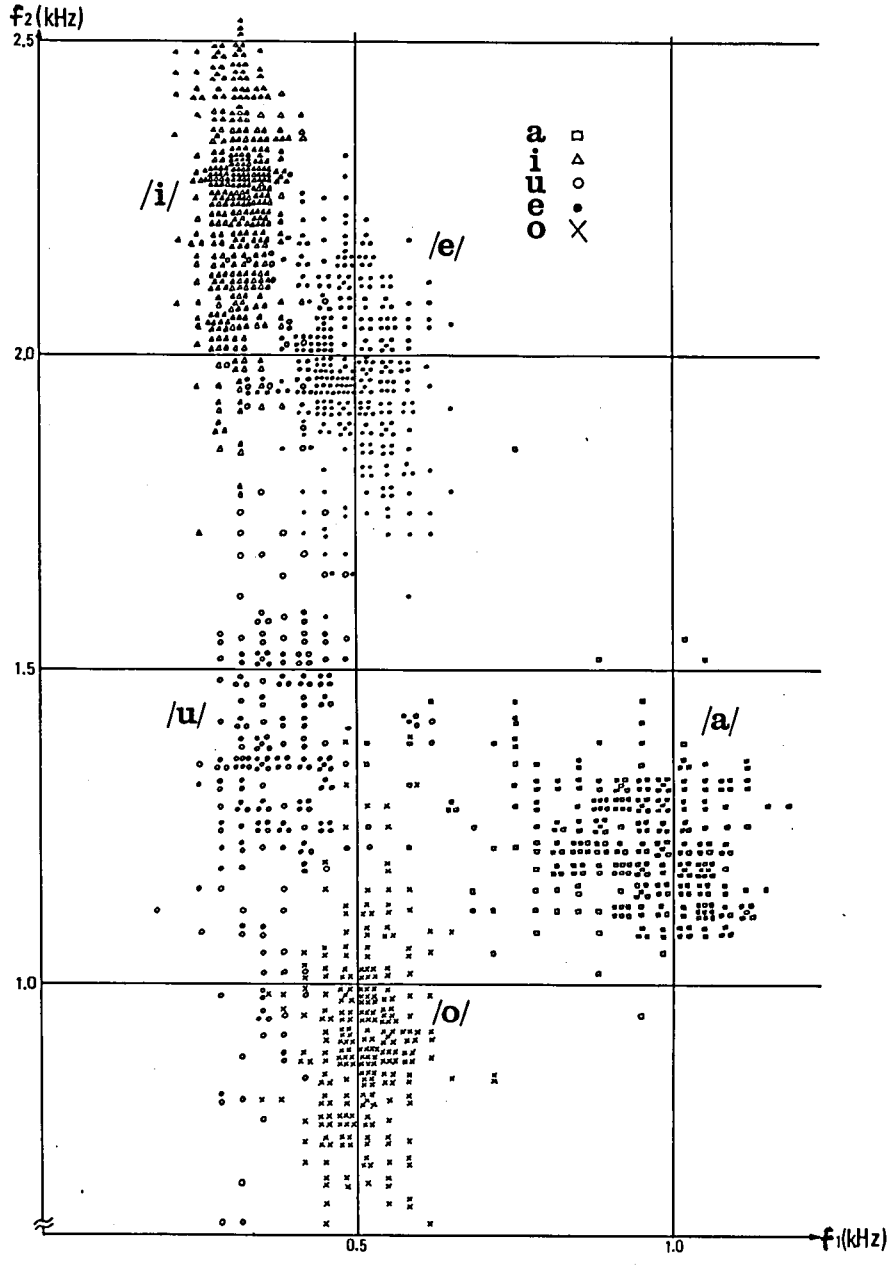


Fig. 2.11 Extracted samples plotted in f_1-f_2 plane

The parameters f_1 and f_2 sets of extracted samples of spoken words produced by the said five announcers are plotted in X–Y plane, and are shown in Fig. 2.11. Generally, f_2 value is distributed more widely than f_1 value. Particularly, f_2 values of the vowel [u] distributed in a wide range and some of them have f_2 values comparable with those of the vowel [i]. This is caused by the fact that the characteristics of the vowel [u] are seriously affected by the context.

2.4.5 Discrimination of Parameters Set

A set of parameters $(f_1, f_2, f_Z, R_1, R_2)$ can be considered as points in five dimensional vector space. Therefore, the recognition of the categories is to divide the Euclidean vector space into several subspaces by the linear or non-linear hyperplanes.

Classification by linear discriminants is based on whether the values of linear combination of weighted parameters are greater than a threshold or not. This classification is simple and is applicable to a wide range of applications. On the other hand, classification by non-linear discriminants is complicated and can not be constructed simply by the hardware. By these discriminants, however, the correlation among parameters can be taken into consideration, and generally better results may be obtained from them than the linear discriminants.

Now, let x be the vector representation of parameters set, and c_i be the event that the vector x belongs to category Π_i, then the discrimination with a minimum error rate may be obtained by the following decision scheme.

$$\begin{aligned} &\text{if} \quad p(c_i/x) \leqq p(c_j/x) \qquad (i=1,2,\ldots 7) \\ &\text{,then } x \text{ belongs to category } \Pi_j \end{aligned} \tag{2.8}$$

Here, $p(c_i/x)$ is probability that the vector x belongs to Π_i. Now, from Bayes' theorem, the following equation is derived, that is,

$$p(c_i/x) = \frac{p(c_i) \times p(x/c_i)}{p(x)} \tag{2.9}$$

Supposing that the distribution of parameter sets in a category is normal, $p(x/c_i)$ may be written, as follows.

$$p(x/c_i)=\frac{1}{2\pi\sqrt{|\Sigma_i|}}\exp\left[-\frac{1}{2}(x-\mu_i)\,{}^t\Sigma_i^{-1}(x-\mu_i)\right] \qquad (2.10)$$

Here, Σ_i is the covariance matrix and μ_i is the mean vector. If the *frequency of appearance* of every category is equal, by substituting the equation (2.8) with (2.9) and (2.10) and by engaging logarismic operation, equation (2.8) is re-written as follows:

$$(x-\mu_i)\,{}^t\Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i| \geqq (x-\mu_j)\,{}^t\Sigma_j^{-1}(x-\mu_j)+\log|\Sigma_j| \qquad (2.11)$$

Let D_i be defined by

$$D_i=(x-\mu_i)\,{}^t\Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i| \qquad (2.12)$$

Then, the equation (2.11) is re-written as in the following.

$$\begin{aligned}&\text{if } D_j \leqq D_i \ (i=1,2,\ldots 7)\\ &\text{then } x \text{ belongs to category } \Pi_j\end{aligned} \qquad (2.13)$$

The following recognition experiments are based on this Bayes' discriminant functions.

2.4.6 Results and their Discussion

According to the above-mentioned classification algorism, a total of 8,208 samples were classified into seven categories. Because the extracted parameters depends heavily on individual speakers, there may be sharp differences between the scores obtained by applying personally adjusted discriminant functions and by

applying the common discriminant functions to every person.

In Table 2.7, the results obtained by the common discriminant functions are shown. Generalization results of 84.6%-correct sample classification were

Table 2.7 Recognition results
(20 male, by common discriminant functions)

Phoneme in spoken words \ Recognized as	/a/	/i/	/u/	/e/	/o/	/s/	/n/
/a/	95.2	0.0	0.4	0.8	3.0	0.5	0.1
/i/	0.0	91.0	0.7	3.2	0.0	0.2	4.8
/u/	0.4	1.7	74.3	1.8	6.3	0.1	15.3
/e/	1.0	2.7	2.6	91.7	0.0	0.6	1.5
/o/	1.3	0.0	5.0	0.0	92.6	0.0	1.1
/s/	0.0	0.6	0.5	0.7	0.0	97.6	0.7
/n/	0.0	12.7	34.2	2.1	0.5	0.5	50.1

(%)

Table 2.8 Recognition results
(20 male, by personally adjusted discriminant functions)

Phoneme in spoken words \ Recognized as	/a/	/i/	/u/	/e/	/o/	/s/	/n/
/a/	99.1	0.0	0.1	0.1	0.4	0.1	0.1
/i/	0.0	94.7	1.1	1.7	0.0	0.0	2.6
/u/	0.0	1.0	83.9	0.9	1.9	0.0	12.2
/e/	0.1	1.4	1.1	96.5	0.0	0.1	0.7
/o/	0.3	0.0	1.9	0.0	97.3	0.0	0.5
/s/	0.0	0.0	0.1	0.1	0.0	99.6	0.2
/n/	0.1	4.2	9.8	0.8	0.9	0.1	84.1

(%)

Table 2.9 Recognition results
(20 male, by common discriminant functions, except /n/)

Recognized as

Phoneme in spoken words	/a/	/i/	/u/	/e/	/o/	/s/
/a/	95.3	0.0	0.4	0.8	3.0	0.5
/i/	0.0	94.4	2.1	3.2	0.0	0.2
/u/	0.7	4.3	86.3	2.4	6.3	0.1
/e/	1.0	3.2	3.5	91.8	0.0	0.6
/o/	1.6	0.0	5.8	0.0	92.6	0.0
/s/	0.0	0.7	0.7	0.7	0.0	98.0

(%)

achieved. Confusions between [n] and [u], [n] and [i], and [u] and [o] are conspicuous. In Table 2.8 are shown the results obtained by the personally adjusted discriminant functions. Recognition scores of [a] and [s] are rather good. About 93.6% of samples were classified correctly. It is clear that the latter scores are better than the former. The difference of the scores of [n] are notable in particular. The confusion between [u] and [n] occured very frequently, in any case. Recognition experiments with the exception of /n/ by the common discriminant function are shown in Table 2.9. Recognition scores of /u/ were improved by about 10%. Total recognition score of the experiments is 93.1%.

Vowels pairs which are apt to be confused are (/a/, /o/) (/i/,/e/), (/i/, /u/), (/u/, /o/) and (/u/, /e/). The confusion between /u/ and /o/ is conspicuous in particular.

The parameters f_1, f_2 and f_Z obtained by the zero-crossing measurements and the parameters R_1 and R_2 indicating the amplitude ratios were used for the classification of vowels and two consonant groups. f_1 and f_2 were independent of each other, and were very important parameters for the description of

phoneme characteristics. f_Z and R_1 played the important role for the discrimination of fricative sounds. Except for the discrimination between /u/ and nasals, sufficient scores could be obtained.

2.5 Conclusion

In this chapter, we discussed the fundamental experiments for speech processing by the zero-crossing analysis methods.

In section 2.2, effects of voicing source and characteristics of vocal tract resonances upon zero-crossing intervals were analyzed by using the synthetic speech. These results were too complicated for any precise mathematical expression. However, macroscopic understanding of these effects were clarified.

In section 2.3, the sequences of zero-crossing intervals of natural speech waves were analyzed. The repeated sequences of zero-crossing intervals were observed in voiced sounds. The pitch extraction has also been tried by using the auto-correlation of the sequences. Fundamental notions of "wave-element" were introduced and will be used in the following chapters.

In section 2.4, we described the method of formant extraction by zero-crossing measurements. Frequency parameters and other two parameters were used for phoneme classification. Rather good scores of correct classification were obtained and the dependency of applied discriminants upon the rate of accurate classification was also shown.

Chapter 3. Speech Synthesis by Rule

With increasing improvements in on-line computer applications it has become highly desirable to develop a practical system of speech synthesis for use as a computer output. Such a system would at least require a good degree of speech intelligibility and should ideally achieve synthesis in real time. Moreover, the program for synthesis and the synthesizer to be attached to a computer should be simple and applicable to any type of computer.

In this chapter is discussed a speech synthesis system capable of fulfilling these various requirements. The vocal fragments treated herein as digital data are equivalent to unit pitch durations of speech and are referred to as wave-element. In this system, the previously stored data (i.e., wave-element) are combined by means of a computer according to a certain set of rules and these vocal combinations are fed to a synthesizer to produce speech waveforms. The chapter describes segmental expressions of phonemes, the program for the synthesis, and the synthesizer, all of which are fundamental to the system.

3.1 Introduction

The electronic computer has become far superior to man in its ability to memorize numerical data or to process information via suitable algorithms. At the same time, development of the techniques involved in time-sharing systems has enabled a number of people to utilize the data-processing ability of the computer more economically and with greater speed.

Teletypewriters are commonly used as terminal equipment for data communication. Although a teletypewriter is capable of transmitting codes and can function as a computer input and output unit, it is relatively expensive, and cannot provide the general-user convenience of a telephone. It is therefore considered that economical and rapid transmission of output information from an information processor to a large number of people could very effectively be achieved by voice transmission through a telephone network.

Under these circumstances, interest in the study of speech synthesis has intensified of late. In the past, speech synthesis has been achieved by a few well-known systems such as terminal analog synthesis [23] [33] [40], wherein voice source waveforms and transfer functions of the vocal tract are simulated by means of electric networks, or by vocal tract analog synthesis [5] [19] [31], wherein the speech output mechanism is simulated more directly by constructing an electric network imitating a vocal tract.

In order for the terminal analog synthesizing system to attain the objective, various speech-defining parameters must be collected in advance of the speech production, through statistical analysis of the natural voice or trial-and-error adjustment of various parameters with respect to each tentatively synthesized speech. The number of speech-defining parameters should be as small as possible to simplify the data storage and the synthesizing process. This is important particularly when an electronic data processor such as an electronic computer is employed to store and process those parameters.

In contrast with these systems, which control synthesizers by making use of various kinds of parameters, there is a system which synthesizes speech by combining various speech segments that are stored for processing by an analog or digital machine [18] [20] [28] [57]. This system is used when the response to an inquiry mostly contains numerals with only a few accompanying words and has a fairly definite form of expression, as in stock-market quotations and volume of trading or telephone directory assistance. As a special example, we can point out a system which is capable of reading letters for the blind by editing and synthesizing speech corresponding to each word (about 20,000 words).

In systems based on the compilation of speech elements the vocal unit used for compilation determines the required capacity of the memory and the naturalness of the synthesized speech. Some of the speech elements most often employed are phonemes, syllables, and words. If words are used as speech elements, we obtain the best naturalness, but the sentences that can be synthesized are limited. On the other hand, if phonemes are used, any sentence can be synthesized in principle, but the naturalness is sacrificed to a considerable extent. Difficulties in speech synthesis lie in the reproduction of speech

elements (speech quality, cutting of fragments), naturalness in the synthesized words (continuity of formants, pitch and "transitional" accent). Some of the technical problems involve selection of appropriate methods for storage and readout, the number of speech segments to be used, and how to arrange segments to produce patterns closest to natural speech.

This chapter describes a speech synthesis system which was developed for the purpose of synthesizing any sentence by means of an electronic computer. The system is geared toward male Japanese speech synthesis. As speech elements used for compilation, units smaller than phonemes were selected (e,g., the pitch of particular voiced sound) and thus the consideration of coarticulation effects and continuous connections of formant loci were made possible. Storage of the compilation units is achieved by a system in which speech waveforms are transformed into zero-crossing waves to reduce the required amount of information and also to facilitate digital processing by electronic computer through quantization of the zero-crossing intervals. The synthesizer which was used to transform the compiled segments into speech was constructed with about eighty integrated circuts and connected to the computer for on-line operation. This system was also tried to English speech synthesis. The supplemental problem is to transform the English sentence into a phonemic string.(46)

We can get multi-channel outputs of synthetic speech sounds easily from this speech synthesis scheme because the ability of the transmission rate of the data channel of the computer by far exceeds the zero-crossing rate of the speech sounds. A multitude of speech synthesizers with the common random-access memories can be controlled at a time with no delay.

3.2 General Description of the System(44)∿(49)

The instantaneous speech synthesis system with the computer consists of a dictionary of speech elements memorized as digital data, a set of transformation rules for speech elements (transformation table and program), and a synthesizer which transforms, into speech waveforms, series of compiled and combined speech

elements based on these rules. The prepared speech elements (hereafter called "wave-element") consist of about 550 units of both vowel and consonant types. For synthesis the pre-recorded digital data are transferred from magnetic tape to magnetic drum*, where they are randomly accessible for magnetic core memory. The transformation table describes a set of rules by which (1) Romanized sentences are changed to phoneme symbols and (2) sequences of phoneme symbols are transformed to wave-elements, i.e., actual speech data, and also to sequences of information necessary along with the wave-elements. Figure 3.1 shows a block diagram of this system.

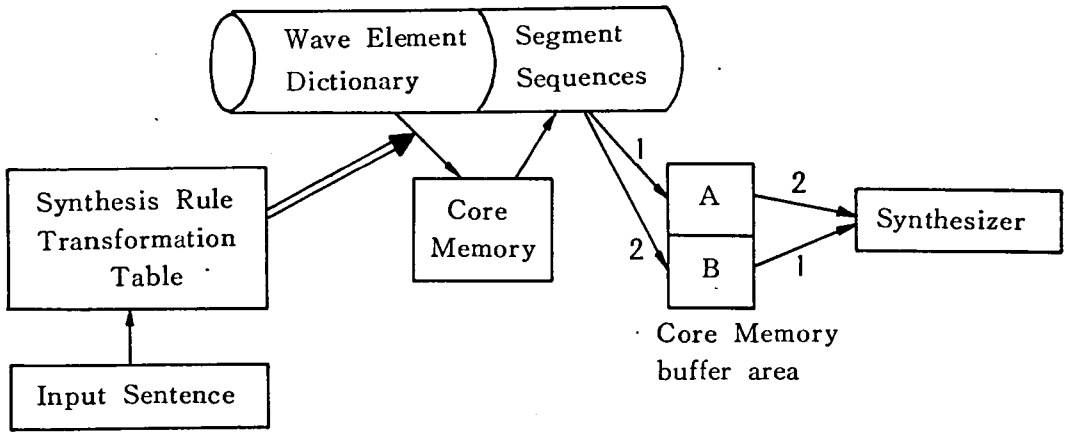


Fig. 3.1 Block diagram of the speech synthesis system

The input is a paper tape in which Romanized Japanese sentences (the essentially same result could be obtained, even when these sentences were written in Kana letter) and accent symbols are perforated. The input is first transformed into sequences of phoneme symbols according to the transformation table. This is done because there are pairs like /ta/ and /ti/ or /sa/ and /si/ which are written in the same way but are pronounced quite differently. Then these sequences of phoneme symbols are transformed into sequences of wave-elements.

---

* This magnetic drum has 64 bands, each band consisting of 40 sectors; each sector has 128 characters; the average access time is 8.3 ms.

A given phoneme may be pronounced differently depending upon the phoneme preceding or following it. This situation is referred to as "coarticulation effect" and must be taken into consideration in the synthesis; otherwise, the result cannot be expected to have naturalness or clarity. Thus, this system is also designed to determine wave-elements for "transitional" parts in consideration of the connection of a vowel, a consonant, and a vowel, the so-called V–C–V context. The wave-elements corresponding to a sentence to be synthesized are processed in this manner and are read into the core memory from the wave-element dictionary in sequence, then the wave-element data along with their auxiliary information are written in the magnetic drum. The auxiliary information contains the amplitude for each wave-element and also the number of times a wave-element to be repeated.

As explained in detail in the following section, the transformation of sustained vowels and voiced consonants into speech is achieved by assigning constant amplitudes to them and by repeating the wave-element equivalent to a pitch. Hereafter, a set of wave-elements and their auxiliary information is called segment.** Sequences of segments are transmitted to the synthesizer and transformed into speech waveforms in real time via the core memory buffer area.

3.3 Compilation units

Speech sound can be roughly segmented into sustained parts and transitional parts. The former includes vowels, nasals, liquid-like sounds and other fricative sounds whose characteristics are uniform in their wave form or energy concentration. The latter includes bursts and glides whose characteristics vary with time. Among the sustained sounds, vowels, nasals and liquid-like sounds are nearly periodic. Fricative sounds have no periodicity, but their durations are

** Wave-elements only contain information concerning zero-crossing time intervals. However, since each wave-element has an amplitude assigned to it at the time of synthesis, the synthesized sound is not a perfect zero-crossing wave, but it is a square wave which changes its level in each wave-element.

long and the frequencies of energy concentration are rather stable. That is, speech sounds in this category have a large amount of redundancy with respect to their wave form characteristics. Therefore in this system they are expressed in simple forms described in detail later, which are similar to "regular expression" pertaining to the automata theory. The characteristics of transitional parts of speech sounds are constrained by adjacent phonemes. They can be described by the rule for synthesis in that the contextual effects can be introduced.

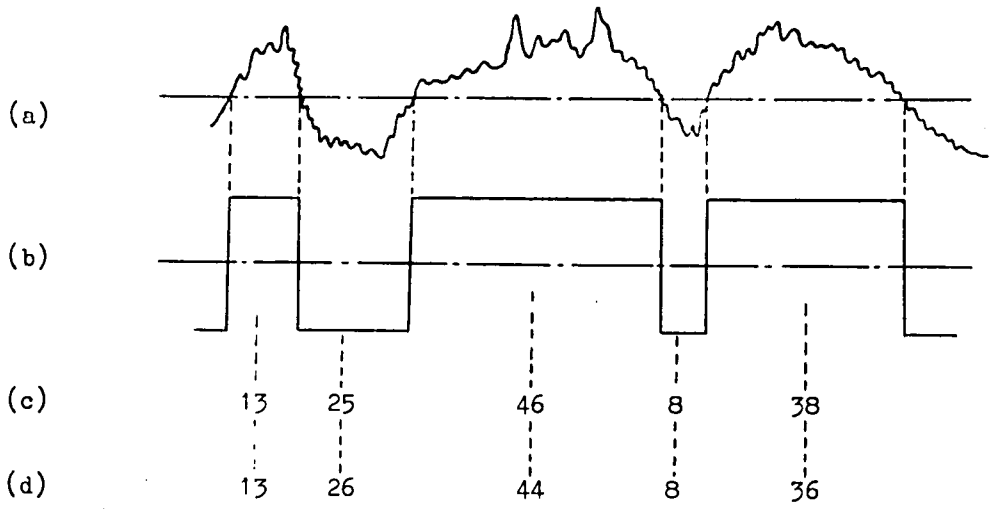
3.3.1 Wave-Element as a Compilation Unit

In our studies speech is entirely described by zero-crossing waves, since it is desired that wave-elements be digitized for storage by a digital computer and that as small a unit of information as possible be used to express speech. In the present study the following rules were used to quantize the zero-crossing time intervals (Fig. 3.2 (C)) of zero-crossing waves (Fig. 3.2 (b)) [35]. Letting τ denote the zero-crossing interval measured by a 20-kHz clock, a_n and b_n are given as follows:

$$a_n = \left[\frac{20000}{2n\triangle f}\right], \quad b_n = \left[\frac{20000}{2(n+\frac{1}{2})\triangle f}\right] \qquad (3.1)$$

where $\triangle f$=50Hz. (Here, [a] is the Gaussian symbol indicating the largest integer which does not exceed a.) For τ such that $a_{n+1} < \tau < a_n$ the quantized zero-crossing interval τ' was put as $\tau' = b_n$. The correspondence between τ and τ' is shown in Fig. 3.2 (e). This is equivalent to nearly uniform quantization with unit step of $\triangle f$=50 Hz in the range where $a_n \neq a_{n+1}$ on the frequency axis. In the end, the zero-crossing time intervals are distributed among 26 different numbers (symbols) as shown in Fig. 3.2 (d).

The preparation of wave-elements was carried out by analysis of zero-crossing waves obtained by the above method from Japanese syllables pronounced by a male speaker (whose pitch period was about 9 ms). First, speech sound waves are sampled by a 20-KHz clock, their amplitudes coded by an A–D converter



(e)

τ	1	2	3	4	5	6	7	8	9	10
τ'	1	2	3	4	5	6	7	8	9	10

τ	11	12	13	14	15	16 ~ 17	18 ~ 19	20 ~ 21	22 ~ 24	25 ~ 27
τ'	11	12	13	14	15	17	18	21	23	26

τ	28 ~ 33	34 ~ 39	40 ~ 49	50 ~ 65	66 ~ 99	100 ~
τ'	30	36	44	57	80	133

Fig. 3.2 Transformation of speech into zero-crossing wave

with accuracy of sign +10 bits, and then these codes are stored in the magnetic tape of the computer. Speech sounds stored in the magnetic tape are divided into groups of 30 samples each, each group is then read out, all the points of plus and minus code conversion are detected, and finally the sounds are transformed into zero-crossing waves.

After the zero-acrossing widths are assigned to 26 distinct values, the sequence of scattered zero-crossing intervals is printed out by the line printer. Then a characteristic

pattern of each phoneme is selected from these speech analysis data and is used as information on wave-elements for expression of phonemes as explained in the following section. (In the case of voiced speech, for instance, the data correspond to the information on one pitch interval in the sustained part.)

The wave-elements for the so-called "transitional" part between a consonant and a vowel or between vowels, where the formant shifts quite rapidly, is hard to be obtained from natural human speech. The present system is therefore designed in such a way that this part can easily be prepared mechanically no matter what combination the formants have. The wave-elements for the "transitional" parts are obtained by simulating the terminal analog model, as shown in Fig. 3.3 by means of a computer. Here the sampling frequency for the amplitude is 20 KHz. Moreover, F_1 and F_2 correspond to the first and second formant frequencies, respectively, whereas the third and fourth formants, F_3 and F_4, are fixed. The simulation was carried out at 50-Hz intervals for combinations of frequencies (F_1 and F_2) between 300 Hz and 950 Hz and between 700 Hz and 2500 Hz. Then the speech sound corresponding to one pitch in the sustained part was transformed into a zero-crossing wave and the zero-crossing width was quantized into one of 26 numerical numbers.

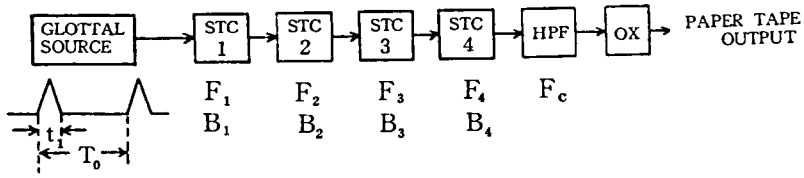


Fig. 3.3 Terminal analog model for production of wave-elements for transitional part

STC: single-tuned band-pass filter

HPF: RC one-stage high-pass filter

F_i: center frequency; F_1, F_2 variable F_3=2700Hz, F_4=3500Hz

B_i: band-width; B_1=70Hz, B_2=100Hz
B_3=200Hz, B_4=350Hz

F_c: cut-off frequency of HPF; 1600Hz , 6 db/oct.

T_0: pitch period; 9msec

t: 100μsec

3.3.2 Expression of Phonemes by Segments

From the standpoint of speech synthesis, the elements in the expression of each speech wave would be desirable to be small. A vowel presents a fairly uniform periodic waveform and can therefore be expressed as repetitions of the pitch waveform which is the fundamental period of the vowel. The voiced plosives can be expressed formally according to the rule that each plosive consists of a buzz bar arising due to vibration of vocal cords and of a plosive which follows it. By means of these methods of expression, a speech wave corresponding to a phoneme can be expressed as a sequence of three parameters. These parameters are a wave-element "P" consisting of a sequence of discrete symbols, the amplitude information "a" and the number of required repetitions "r" of the wave-element. This set of three parameters is called a segment and denoted by aP^r. Table 3.1 (a) shows some examples of wave-element expressions of some phonemes, while Table 3.1 (b) gives the sequence of symbols (the sequence of zero-crossing widths) of P_a.

Table 3.1 (a) Expression of Phonemes by segments

(b) Sequence of numerical zero-crossing intervals of Pa

(a)

PHONEME	EXPRESSION	PHONEME	EXPRESSION
/a/	$17P_a^8$	/d/	$3P_d^5 \cdot 3P_t^1$
/e/	$13P_e^8$	/g/	$3P_g^5 \cdot 3P_k^1$
/m/	$3P_m^7$	/s/	$1P_s^{10}*$
/n/	$3P_n^7$	/z/	$2P_{z1}^4 \cdot 2P_{z2}^2 \cdot 2P_s^4*$
/p/	$OP_{sp}^9 \cdot 2P_p^1$	/ch/	$OP_{sp}^7 \cdot 2P_{ch}^6*$
/t/	$OP_{sp}^9 \cdot 2P_t^1$	/ts/	$OP_{sp}^7 \cdot 2P_{ts}^6*$
/k/	$OP_{sp}^9 \cdot 2P_k^1$	/h/	$1P_h^2*$
/b/	$3P_b^5 \cdot 3P_p^1$	/r/	$2P_r^5$

(b) Pa : 8 8 12 3 3 8 9 10 9 10 11 1
4 10 8 10 8 10 6 7 8 8 9 7

The duration of a vowel sound is, as a rule, set about 70 ms and, if necessary, the duration is regulated by the number of times the pitch is repeated. The nasals /m/ and /n/ are consonants, but each is made by repetitions of a single wave-element as in the case of vowels. The unvoiced plosives /p/, /t/ and /k/ have two segments each. The first segment, namely, OP_{sp}^{9} , is an interval in which no effective information arises since the amplitude parameter is 0. This interval is about 90ms. The wave-elements P_p, P_t and P_k correspond to plosive parts. The voiced plosives /b/, /d/, and /g/ consist of buzz bars P_b, P_d and P_g, and plosive parts P_p, P_t and P_k, respectively, as shown in the table.

The fricative /s/ consists of a single segment. Since this fricative is noise-like, its sequence of zero-crossing intervals should be random and hence a periodic expression such as given above should not be used for this fricative. Thus, unlike other segments, the wave-element P_s is not a repetition of a single unit found in the dictionary but is repeated ten times as a random sequence of the numerical values (about 100 in P_s). This operation is distinguished from other operations by the symbol "*" found in Table 3.1 (a).

3.3.3 Expression of the Transitional Part by Segments

In order to pronounce two different kinds of phonemes continuously in a consonant-vowel, vowel-consonant or vowel-vowel combination, movement of the speech organs must necessarily occur. Such movement and the resulting sound is called a "transition." The transitions between phonemes must be synthesized by the use of those segments which were chosen to make the formant loci continuous.

It is known [36] [39] that in various vowel-consonant-vowel combinations the formant frequency at the point where second formant appears in the transition from consonant to vowel (i.e., the "off-transition" point) and the formant frequency at the point where the second formant disappears in the transition from vowel to consonant (the "on-transition point) are closely related. Moreover, it is considered that the three-phoneme group is sufficient for consideration of such contexts. In the present study, analysis of natural sound by sonagrams led to measurement of change in the off-transition from a consonant

to a vowel and in the on-transition from a vowel to a consonant (see Fig. 3.4).

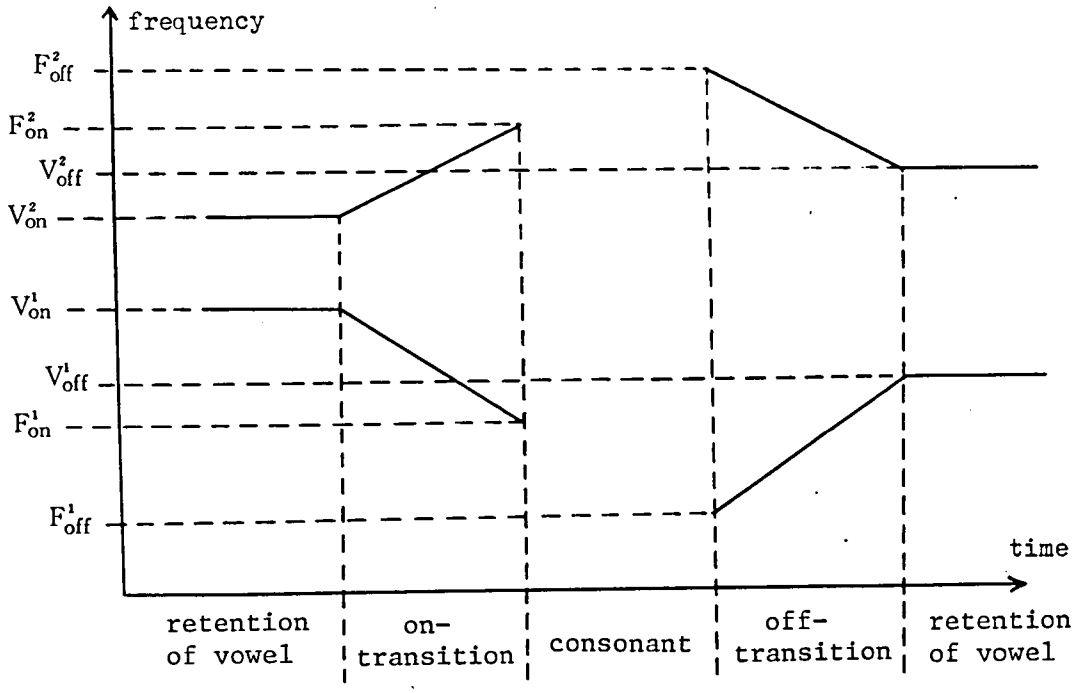


Fig. 3.4 Formant loci in V–C–V context

As a result, the following empirical equations were obtained:

$$F^i_{on} = A^i_{on} \times V^i_{off} + B^i_{on} \times V^i_{on} + C^i_{on}$$
$$F^i_{off} = A^i_{off} \times V^i_{off} + B^i_{off} \times V^i_{on} + C^i_{off} \qquad (3.2)$$

where i is 1 or 2 (the first or second formant), F_{on}^i and F_{off}^i denote the formant frequencies at the on-transition point and the off-transition point, and V_{on}^i and V_{off}^i denote the formant frequencies of the preceding and succeeding vowels, respectively. A, B, and C are characteristic constants of each consonant. Some examples are given in Table 3.2. Here, T^2_{OFF} in the table is calculated by;

$$F^2_{OFF} = T^2_{OFF} + A^2_{OFF} \times (V^2_{OFF} - T^2_{OFF}) + B^2_{OFF} \times (V^2_{ON} - T^2_{OFF}), \qquad (3.3)$$

T^2_{OFF} is considered to be the so-called target frequency. The entries of the right-most column represent the calculated examples of F^2_{OFF} in [a–C–a] context.

Linear approximation is carried out between these calculated F_{off}^1, F_{off}^2, F_{on}^1, F_{on}^2, and their respective V_{off}^1, V_{off}^2, V_{on}^1, and V_{on}^2 and the first

and second formants of the wave-elements to be interpolated are decided. The transitional parts between phonemes are obtained by these wave-elements with

		off-set formant frequency (Hz)						on-set formant frequency (Hz)						target freq.	computed F^2_{off} in a-C-a (Hz)
		first formant			second formant			first formant			second formant				
		A^1_{off}	B^1_{off}	C^1_{off}	A^2_{off}	B^2_{off}	C^2_{off}	A^1_{on}	B^1_{on}	C^1_{on}	A^2_{on}	B^2_{on}	C^2_{on}	T^2_{off}(Hz)	
consonant	/t/	0.45	0	240	0.45	0.18	690	0	0.45	240	0	0.55	750	1860	1450
	/d/	0	0	400	0.27	0.22	850	0	0	400	0	0.44	880	1670	1440
	/b/	0.28	0	220	0.53	0.17	200	0	0.28	220	0	0.70	300	670	1040
	/r/	0	0	300	0.20	0.20	970	0	0	300	0.02	0.50	790	1620	1450
	/m/	0.40	0	270	0.69	0	220	0	0.40	270	0	0.69	220	710	1050

Table 3.2 Coefficients necessary for calculation of transitional formant frequencies

appropriate amplitudes assigned to them. These wave-elements are previously prepared by simulation program of the afore-mentioned terminal analog model. Each is repeated only once.

3.4 Speech Synthesis Program

Figure 3.5 shows the flow chart of the speech synthesis program. A composite sentence are interpreted and transformed into speech in order. After the Japanese sentence is read into the computer, it is converted to a sequence of phonemes. Then this sequence of phonemes is expressed as sequences of segments as described in Table 3.1 (a). Moreover the wave-element for the transitional part between two single sounds and the corresponding amplitude are determined by the rule described in the preceding section. Once a given sentence is totally expressed as a sequence of necessary segments, the wave-element consisting of the sequence of numbers is read out of the wave-element dictionary and written in the drum. When this is completed, the sequence of numbers stored in the magnetic drum is immediately transimitted to the synthesizer through the core memory. (see Fig 3.1)

Sentences divided by periods or question marks are coded into a paper tape and then transformed into sequences of phonemes, as explained in Sect. 3.2, with

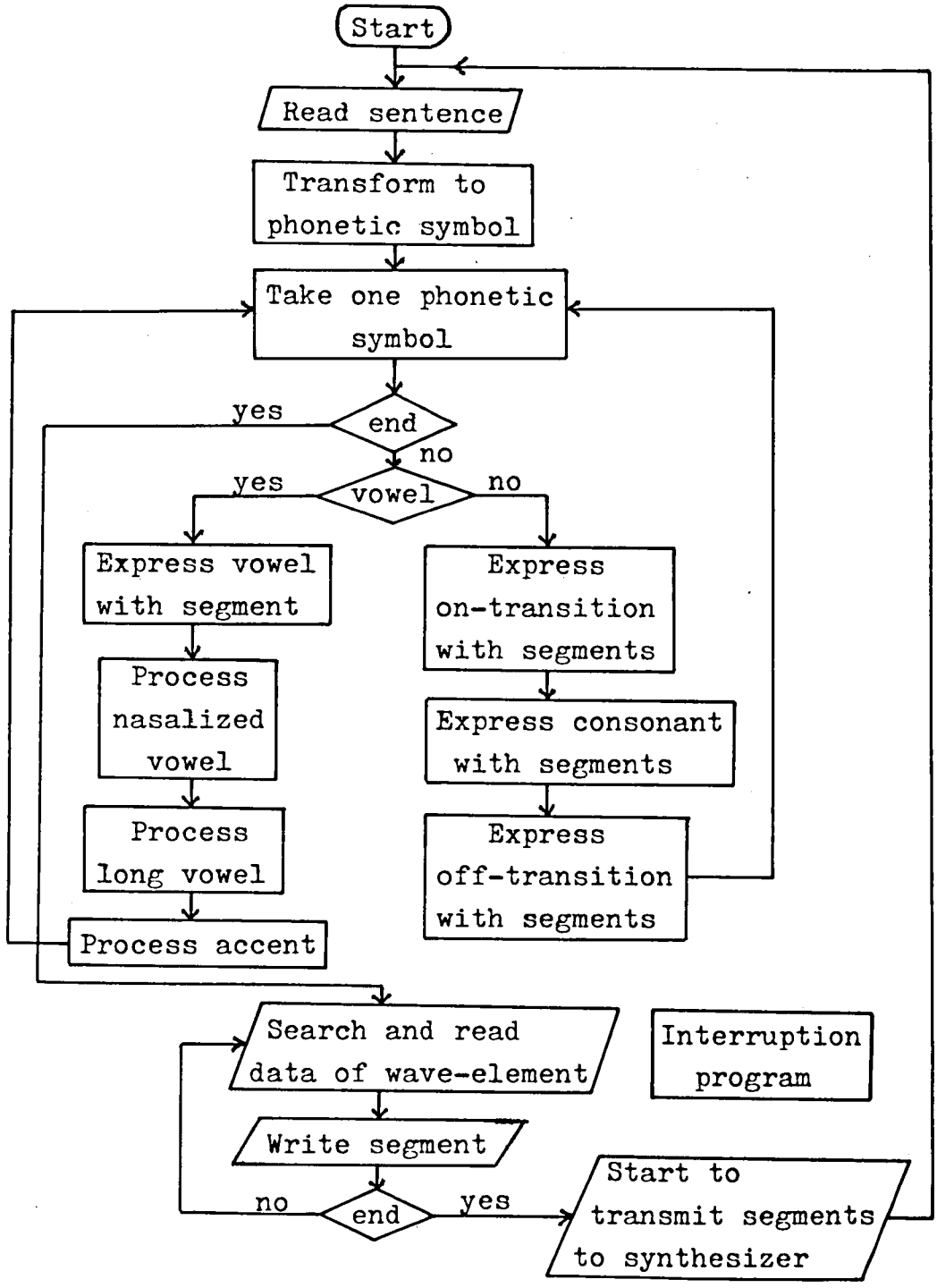


Fig. 3.5 Flowchart of speech synthesis program

accent symbols interpreted. An accent symbol is punched on the tape above or below the appropriate phonemes. The pitch frequency of a sustained vowel is changed by manipulating the sequence of zero-crossing intervals of the wave-elements which comprise one pitch. Thus, to raise the pitch, some data at the rear of the wave-element are erased. This affects the spectral structure only slightly.

The sequence of phonemes is processed one at a time. In the case of a vowel, the phoneme is first expressed by segments following the procedure on the left in Fig. 3.5. In the case of a vowel following a nasal like /m/ or /n/, the phoneme is replaced by the symbol for the wave-element obtained from other nasalized sounds that were prepared beforehand for the purpose. Then, the number of above-mentioned sequences of phonemes is determined with consideration of accent symbols.

In the case of a long vowel, the number of repetitions of the wave-element is increased. In the case of a consonant, the on-transition from the sustained vowel to the consonant is first expressed as a sequence of segments by the rule specified in Sect. 3.3.3 following the procedure on the right in Fig. 3.5. Then the consonant part is replaced by the segment expression given in Table 3.1. Transition from this consonant to the succeeding vowel is expressed by the calculation described in Sect. 3.3.3. For instance, the initial part "da" of phonetic string "daigaku" is transformed into such a sequence of segments as, in Fig. 3.6 (b). Where (400, 1450) means a wave-element which is a computed one pitch wave whose first and second formant is 400 (Hz) and 1450 (Hz) respectively. Here the value 400 and 1450 and so on are computed by the linear interpolation between the formants of vowel /a/ and offset formant frequencies from /d/ derived by the formula presented in Section 3.3.3, setting the variables V_{on}^{i} to the predetermined constants in this C–V context.

(a) $\underbrace{da}igaku$

(b) $3P_d^5 \cdot 3P_t^1 \cdot 9(400,1450)^1 \cdot 12(600,1400)^1 \cdot 15(750,1350)^1 \cdot 17P_a^8$

(c) $3(2,3,\text{----},23)^5 \cdot 3(7,4,\text{-----},3)^1 \cdot 9(\text{------})^1 \cdot 12(\text{-------})^1 \cdot$
$15(\text{-------})^1 \cdot 17(8,8,\text{-----};7)^8$

Fig. 3.6 Transformation of input symbols to a sequence of segments

In Fig. 3.7 the configuration of the segment is illustrated. Each OXI of the wave element is coded to six bits which correspond to a "character" of the used computer. Two parameters "a" and "r" along with the wave element are coded to lower five bits of two characters. In the case of a consonant segment, one bit indicating whether it is voiced part or not and one bit indicating if it is a plosive are added as indexes to the characters for "a" and "r" and are used as switching signals when the amplitude envelope of the synthesized wave is properly transformed by the synthesizer. This synthesizer will be described later.

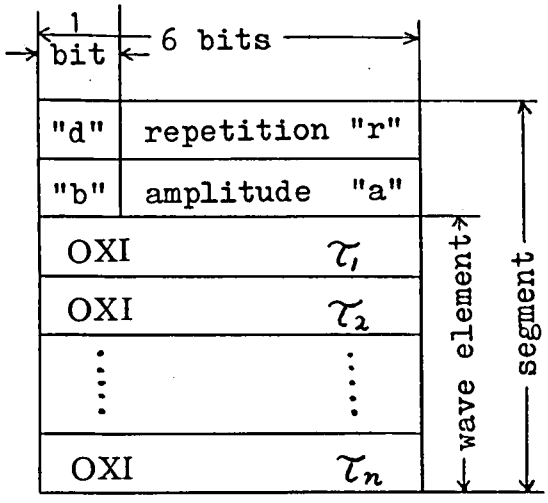


Fig. 3.7 Construction of a segment

When the sequence of segments for the sentence at hand is completely stored in the drum, it is transferred to the buffer area of the core memory a few segments at a time and then put into the speech synthesizer. This speech synthesizer is connected to the data channel of the computer. As soon as the first of the segments comprising a sentence comes out to the synthesizer, the program sequence begins to interpret the next sentence. Consequently, the delivery of new segments from the magnetic drum to the buffer core memory is treated in an interruption program. The timing for this operation is regulated by a special signal transmitted by the synthesizer. Data from the core memory can be fed into the synthesizer continuously by using the two sub-areas of the core memory (A and B in Fig. 3.1) for supply and output alternatively. Thus real-time synthesis of speech is accomplished. This process is roughly sketched in Fig. 3.1. Phases (1) and (2) are interchanged at appropriate moments.

The program for synthesis is written in the NEAC-2200 assembly language (EASY CODER). The program consists of about 730 steps (about 4.5K characters) and the data area is about 9.5K characters.

3.5 Speech Synthesizer

The speech synthesizer is connected to the data channels through a peripheral adaptor. This adaptor is an I/O interface unit designed to connect all kinds of auxiliary equipment to the NEAC–2200 computer. It facilitates the flow of data and control signals between the computer and the external equipment by 53 coaxial cables. The speech synthesizer uses 17 of these cables.

The synthesizer selects the digital data comprising the segments and forms sequences of rectangular waves having the required amplitude, each data segment being repeated the required number of times. These sequences are then transformed into speech waveforms in real time.

Figure 3.8 shows a block diagram of the speech synthesizer. In Fig. 3.9, the synthesizer is illustrated in more detail. Since the synthesizer is not equipped

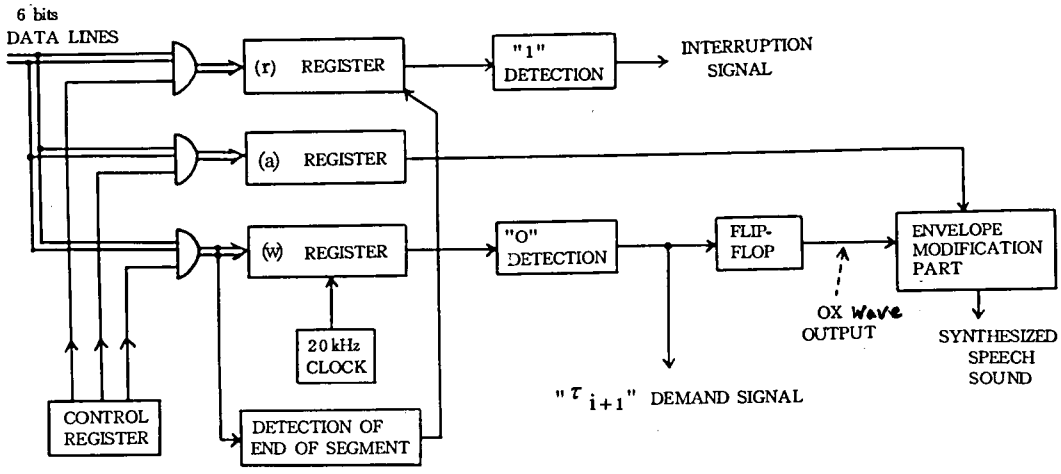


Fig. 3.8 Block diagram of speech synthesizer

with a memory device capable of storing all the data of a segment at the same time, scanning of each wave-element is repeated as many times as is required. Each segment has three kinds of information in the following order: the number of repetions "r", the amplitude "a", and the zero-crossing sequences "τ_i" (i=1,2...) which compose the wave-element P. The numerical values of ("r" and "d"),

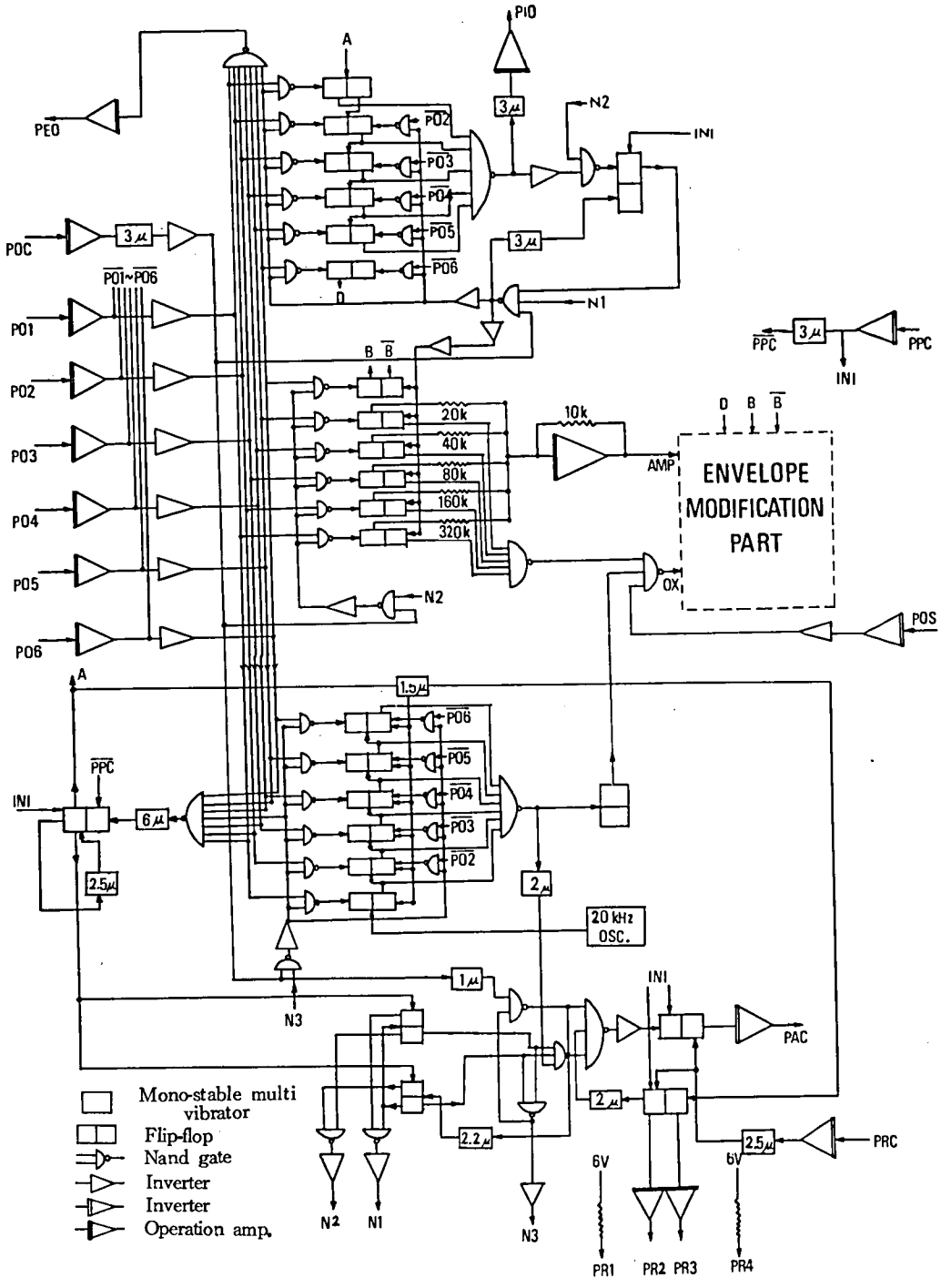


Fig. 3.9 Detailed block diagram of the synthesizer

("a" and "b"), and "τi" are set in 6-bit registers (r), (a), and (w). The numerical values of "τ_i" set in register (w) are reduced by 1 by a 20-KHz clock every 50 μs. Since the sign of a synthesized wave (square wave) is assumed to change when 0 is detected by "0" detection circuit, the flip-flop state is reversed and, at the same time, the numerical value "τ_{i+1}", which follows "τ_i", is requested from the computer. Thus the output of the flip-flop is the zero-crossing wave whose synthesis is desired. The content of register (a) is D–A converted and gives an amplitude to the obtained zero-crossing wave. The value set in register (r) is reduced by 1 whenever the end of the segment is reached. When this figure reaches 1, the register prepares for switchover to another segment and, if necessary, transmits a interrupting signal to the computer for the supply of data from the drum. Thus, when the last scanning on the segment is completed, switchover to another segment is carried out. If the content of register (r) is not 1, then as soon as the end of a segment is detected, the same data of τ_i (i=1,2.···) is supplied from the beginning of the segment. The control register regulates the timing for the storage of data in registers (r), (a), and (w).

The function of the ... envelope modification part in Fig. 3.10 is to assign an appropriate amplitude to a zero-crossing wave and change its envelope to resemble the envelope of the natural sound as much as possible. As its input, this part receives the output of register (a), the synthesized zero-crossing waves, and also the upper 1 bit of registers (r) and (a), namely, the information indicating whether the input sound is voiced or not and whether it is plosive or not. A detailed circuit diagram of this amplitude envelope modification section is shown in Fig. 3.11.

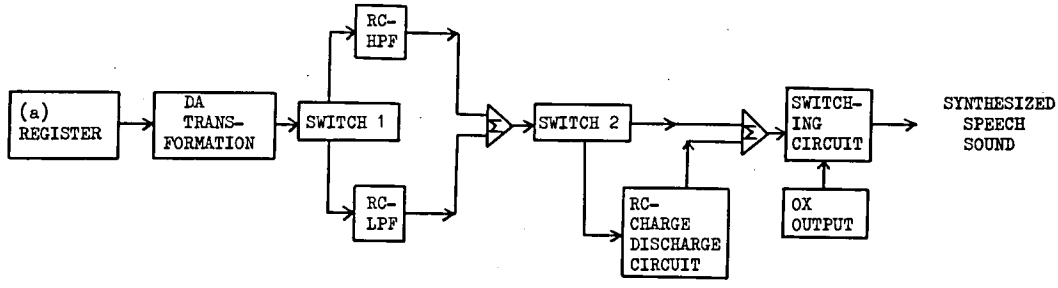


Fig. 3.10 Block diagram of envelope modification part

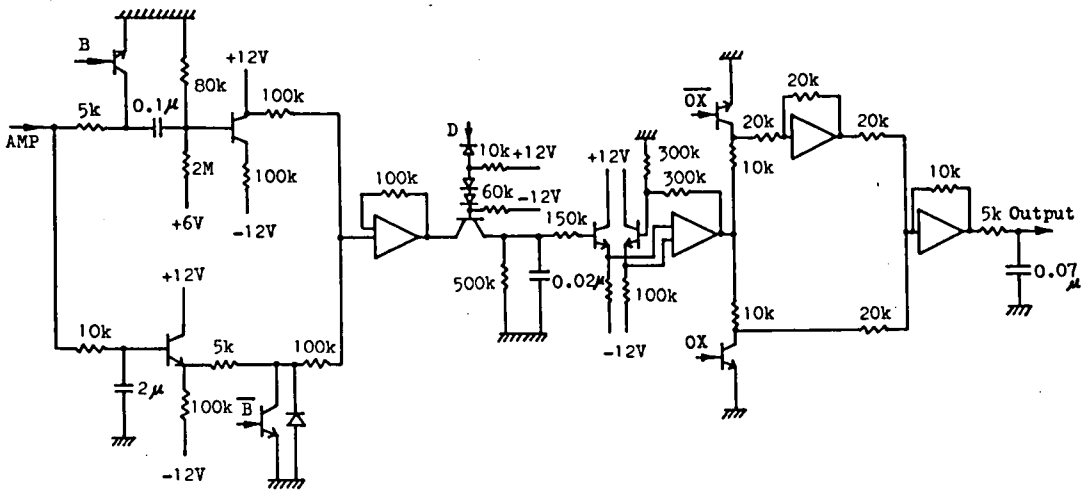


Fig. 3.11 Detailed circuit diagram of the envelope modification part

The value in register (a) is converted to an analog quantity and then sent by Switch 1 through the RC one-stage HPF (T=30 ms) if the bit for plosive indicates "1" and through the RC one-stage LPF (T=30 ms) if it indicates "0". If it is a vowel-like sound (a sustained vowel or its transitional part) or a voiced consonant, the resulting wave is assigned the damping which is characteristic of voiced sounds by charging or discharging the RC charge-discharge circuit (T=7ms) for each pitch. The resulting amplitude envelope is switched positively or negatively by the synthesized zero-crossing waves. Thus synthesized speech waves are obtained.

3.6 System Evaluation

3.6.1 Synthesized Speech Sound

The sonagram of the synthesized speech [daigaku] (大学) is shown in

Fig. 3.12 (a), while (b) shows the sonagram of the corresponding natural sound. In Fig. 3.13 the synthesized speech wave form of that word is presented.

Up to second formant frequency region, the frequency spectrum of the synthesized speech is well matched to the original one. Noise-like pattern in the high frequency region is due to the distortion by zero-crossing wave. The formant loci of the synthesized speech consist of well-creased straight lines; while the natural speech has smoothly continuous formant loci. The amplitude parameters of consonants are slightly over-estimated, in order to improve the intelligibility of consonants. In Fig. 3.13 amplitude envelope of consonant [d] represents gradually rising characteristic; while the plosive part of [k] shows a abrupt rising of amplitude. Damping characteristic of voiced part is also shown in the synthesized speech sound wave in Fig. 3.13.

3.6.2 Amount of Information

Approximately 550 wave-elements were prepared as compilation units for the speech synthesis. Each wave-element is composed of about 25 sequences of zero-crossing intervals on the average. For example, if each interval is assigned one of 26 different numerical symbols and is assumed to be coded in 5 bits, then the information in the dictionary amounts to about 70 K bits. This memory size is sufficiently small to be stored on an auxiliary memory such as a drum memory.

3.6.3 Time Required for Synthesis

We shall next discuss time required for synthesis. A sentence which takes about 10 seconds to read aloud (for example, reading all 48 Japanese alphabet letters) has about 400 segments. Since the average access time of the magnetic drum is 8.3 ms and the readout speed of the device is 80 K characters per second, it takes (8.3 ms + 1.5 ms) x 400=3.9 seconds to read wave-elements out of the dictionary (each sector stores one wave-element). If five segments are written into the drum at a time, then the total access time is approximately 8.3 ms x 90, and the write-in time required for the core memory is approximately 1.5 ms x 450, which does not exceed 1.5 seconds. The capacity

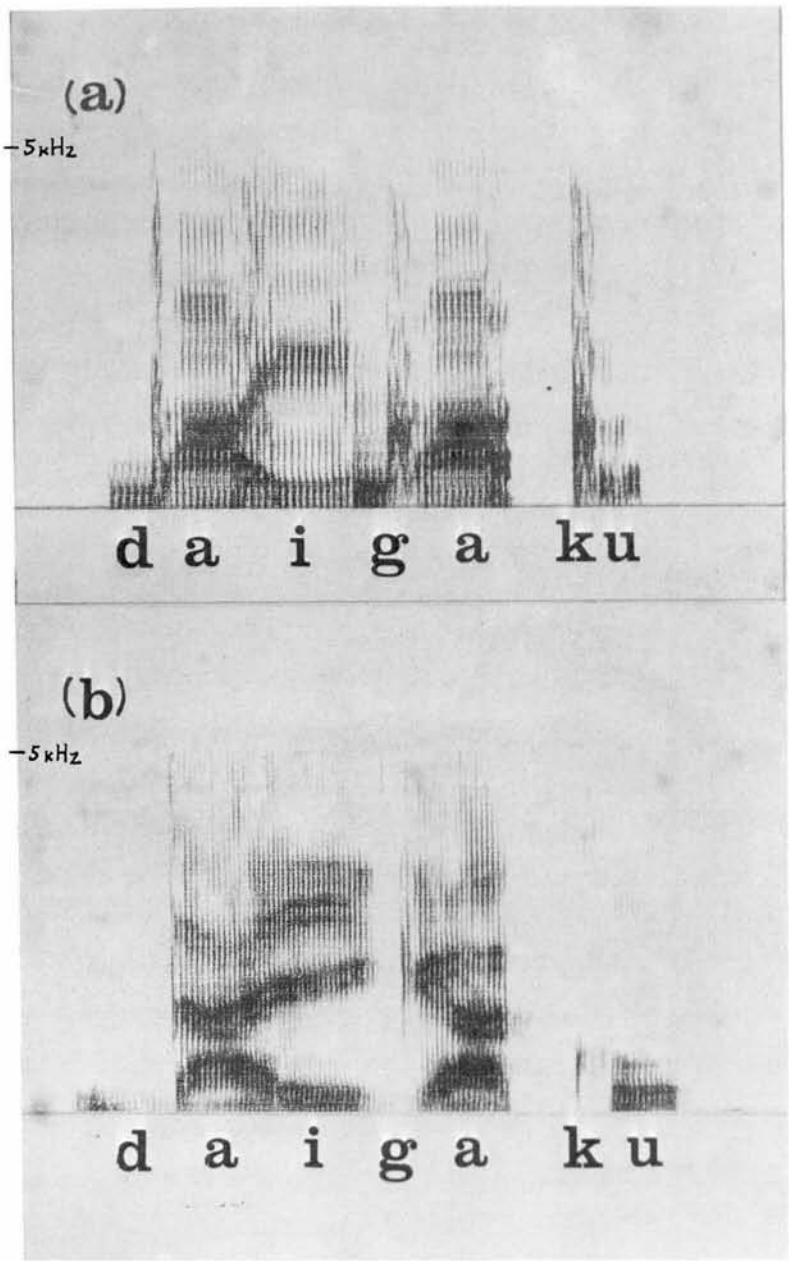


Fig. 3.12 Sonagram of synthetic speech [daigaku] (大学)

(a) synthetic speech sound

(b) natural speech sound

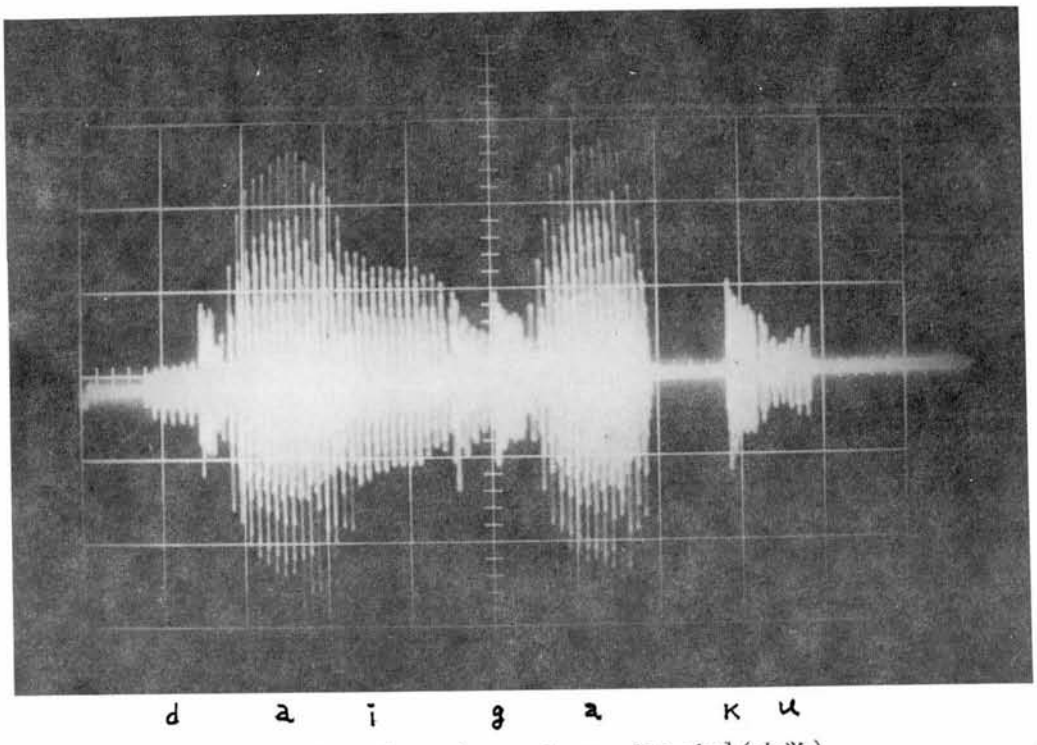


Fig. 3.13 Synthesized speech sound wave [daigaku] (大学)

of the synthesis program is such that it takes about 6 to 7 seconds to express an input sentence as a sequence of segments and to achieve random rearrangement of noise-like segments such as P_S. As a result, the entire synthesis takes 12 to 13 seconds. But this synthesis time can be reduced to less than a third by editing the segments for the transitional parts beforehand. This fact was confirmed by further experiments on the speech synthesis of child and female. It is expected that if the wave-element dictionary is stored in the core memory instead of the drum, the time needed for synthesis may be further reduced.

3.6.4 Intelligibility of Synthesized Speech[47]

Sixty-seven syllables were synthesized by the method described above and a hearing test was conducted by using a tape having random arrangements of these syllables. Speech was heard directly from the loudspeaker in an ordinary

laboratory room which was not anechoic. The participants were three males who had not received any special training. The tape had two different arrangements of the syllables with 3-second intervals between syllables.

The hearing results for consonants are given in Table 3.3. The rate of correct response was 75%. The column "ϕ" in the table indicates the mishearing of a consonant and a vowel as a single vowel. Two instances of this type of mishearing were made in [he] by the same person. Fairly good results were obtained for /k/, /g/, /h/, /s/, /z/, /c/ (ch, ts), and /y/. However, a high rate of mishearing between /m/ and /n/ was found. The rate of correct response to single vowels was 100%, but a few instances of mishearing of vowels were found in the case of a nasal plus a vowel. Although /h/ was frequently misheard for a plosive in earlier tests using the zero-crossing waves of natural sounds[35], the present tests showed a considerably better rate of correct response. This may be due to the fact, as described before, that the amplitude envelopes were put through LPF. Moreover, many instances of mishearing of a vowel for a consonant plus a vowel, which had been reported in a number of previous tests, were not found at all in the present synthesized syllables. There are still some difficulties with regard to such pairs of phonemes as /p/ and /t/ or/m/ and /n/, but it is considered that such difficulties will be solved by improving the data on wave-elements.

3.7 Multiple-Speech Output System[50][51][59]

3.7.1 Introduction

The multiple speech output system can be utilized for the automatic information service such as the information retrieval system and the calculation service. The computer provides the multiple-speech outputs to several kinds of terminals at the same time, with no delay. Lee and Mulvany[28] have demonstrated the voice-recording-reproducing type multiple-speech outputs system. This system, in principle, can reproduce synthetic speech of very high quality. However, its capability is naturally limited because the large storage capacity is

Table 3.3 Confusion matrix for synthesized consonants hearing test

consonant sent \ consonant received	p	t	k	b	d	g	h	s	z	n	m	r	c	y	w	φ
p	20	7	2			1										
t	1	14	3													
k	2	3	25													
b	4	1		14	4	2				2	1	2				
d		1		3	11	3										
g					2	27						1				
h							25	1	2							2
s								30								
z									30							
n	2			2		2				12	9	3				
m										11	19					
r				2	2	9				1		16				
c													12			
y														18		
w	2			2											2	

required for spoken word data, and the device for the generation of connected speech becomes complicated as the variety of messages to be synthesized is increased. Buron[2] has developed the vocoder type speech synthesis system which is suited for multiple speech output. This speech output was presented at Expo 67 in Montreal. Nakata[34] et al have published the results of their research on a new multiple speech output system, which is based on compilation of damped sinusoids. It is said that the multiplicity of this system is about 100, in principle. On the other hand, Matsui[30] has proposed a new multiple speech output method which is based on compilation of impulse responses of vocal tract.

In this section, we will describe a new multiple speech output scheme based on the same compilation method as the one used in speech synthesis by rule, as presented in the preceding section. The fundamental compilation units

are wave-elements; however, at the speech-synthesis stage, syllables corresponding to (V–C) context (V; vowel, C; consonant), (V–V) context, or (C–V) context are used as the actual compilation units in order to save the necessary compilation time. These are constructed with a sequence of segments in advance.

3.7.2 System Configuration

In Fig. 3.14, the block diagram of the system is presented. The actual compilation units correspond to (C–V), (V–V) and (V–C) parts (henceforth, called a di-gram). A sequence of segments $3P_d^5 \cdot 3P_t^1 \cdot$ 9(400,1450)$^1 \cdot$12(600,1400)$^1 \cdot$ 15(750,1350) in Fig. 3.6 is one example of di-grams (C–V). Total size of memory of the actual compilation units is about 170 kbits.

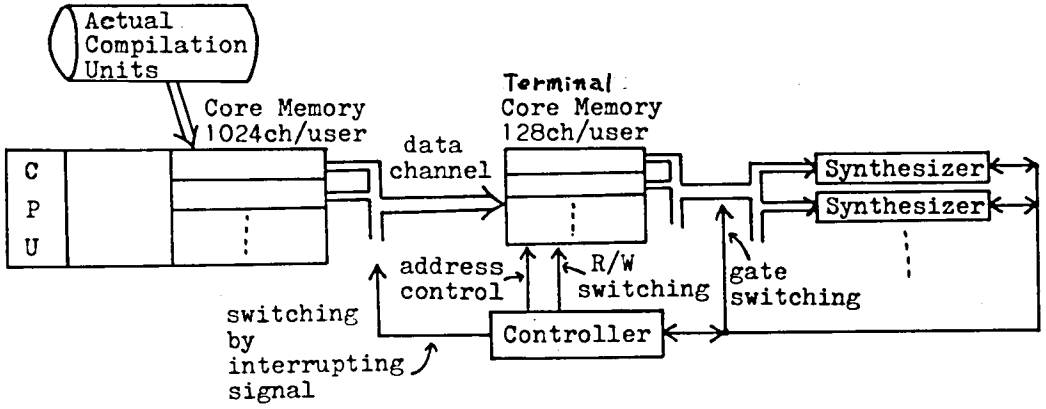


Fig. 3.14 Block diagram of multiple speech output system

The computer program operates in the same way as the so-called multi-programming mode, in which the background job is concurrently operated with the speech synthesis job. The program control is transfered from the batch processing routine to the compilation routine by demanding signal of the speech output which is transmitted from the terminal. In the compilation routine, a di-gram, that is, a sequence of segments is read out into the core memory (512 characters×2/user) from the drum memory. Then, the sequence of segments is transmitted to the terminal core memory (64 characters×2/user) through the data channel one by one, whenever the interrupting signal is transmitted from the terminal synthesizer. Data input/output between the terminal core memory and

the synthesizer is identical with that of the single speech output system. The control part is related with a switching of data transmission between the computer and the terminal core memory and between the terminal memory and the synthesizers. And also, address designation and read/write switching of the terminal core memory are controlled by this part. Both of the control part and user's speech synthesizer consist respectively of about 60 integrated circuits.

3.7.3 Multiplicity of Output Terminals

Data transmission rate between the computer and the terminal core memory is about 1.5 k character/sec. This is about 1/16 of PCM coding (7bits/sample). On the other hand, the capacity of the data channel of the used computer, NEAC 2200/200, is 166 k characters/sec. Then, this capacity is sufficient for the rather many speech outputs.

The CPU time rate spent on the compilation routine is about 0.05∿0.065 per a user. This rate restricts the multiplicity to 15∿20.

The multiplicity mainly depends on the mean access time of the auxiliary memory, that is, the drum memory. The mean di-gram access time is about 10msec in this computer, and the average duration of di-grams is about 110msec. Then, the multiplicity is limited to 11, on an average.

3.8 Conclusion

A discussion has been given of an instantaneous speech synthesis system which compiles and synthesizes speech by using vocal fragments called wave-elements. In connection with this system, the article has also discussed a method of expressing phonemes, a program for synthesis, a synthesizer having an on-line connection with a computer, and the results of tests on the clarity of synthesized speech. In section 3.7, the multiple-speech output system was discussed, and the availability of this scheme was shown from the viewpoint of multiplicity.

Some characteristics of this system are:

(1) little time is required for synthesis,

(2) any sentence can be synthesized because the compilation unit is small, and the coarticulation effect can be considered.

(3) the system matches the digital processing of a computer very well since the vocal units (wave-element) are quantized,

(4) the required program is simple, and

(5) this system can easily be connected to any other information processing system because of the features given in (2), (3) and (4).

On the other hand, a defective aspect of this system is the lack of naturalness in the synthesized sounds. This problem seems to stem mainly from the fact that information on the zero-crossing intervals of speech is used as the compilation units and also that no intonation is taken into account in the sentence analysis.

74 項欠

Chapter 4 Speech Analysis-Synthesis System

In order to examine the effectiveness of parameters in speech recognition, it is preferable to reproduce speech sound from the extracted parameters by a so-called speech analysis-synthesis system.

From this point of view, we have constructed a composite speech research system which consists of a speech analysis-synthesis part and a speech recognition part. Both of them are based on the zero-crossing analysis method.

Experiments were conducted with 1900 words produced by five adult males. The average information compression ratio of the analysis-synthesis part was about 1/10 compared with PCM coding. Intelligibility of the synthesized speech was ascertained to be rather good by a hearing test. In the recognition part, every segmented part of speech is recognized as either one of the 5 vowels and 8 consonant groups. About 95.6% of the vowels and 69.0% of the voiced consonants were recognized correctly by personally adjusted discriminants. Up to 86.9% of the unvoiced consonants could be identified correctly.

More extended recognition experiments were conducted and in these experiments common discriminant functions were applied to all the subjects and other ten male students. The results obtained made clear the efficiency and the merit of the recognition scheme.

In this chapter, the speech analysis-synthesis part of the composite system is described, and an introductory description of the system is also given.

A detailed description of the recognition part will be given in Chapter 5.

4.1 Introduction

It is one of the most important problems in the study of the automatic speech synthesis and recognition to distinguish the parameters which are necessary and sufficient for the expression of the linguistic information contained in the observed speech.

An objective of speech synthesis is to reproduce speech from a simple and essential parameters by machine or some rule expressed by a computer program. If satisfactorily intelligible synthetic speech is obtained, then such parameters can be found to be sufficient for expressing the contents of speech.

On the other hand, in automatic speech recognition, parameters are extracted from input speech and segmented into such units as phonemes. Then, the segmented parts are recognized as one of phonemes. When the vocabulary used is limited, sequence of the extracted parameters or features is normalized on a time axis without the segmentation process; then, the sequence can be recognized as a word by pattern matching method. In either case, the parameters or the results of the segmentation used in such recognition are not necessarily sufficient conditions for original speech in the mathematical sense. In short, the information conveyed by original speech is not sufficiently reserved in the sequence of the extracted parameters or features.

In order to check the sufficiency of these interim results for speech recognition, it is necessary to reproduce speech from them. That is, speech analysis-synthesis experiment is necessary to be conducted. Both speech synthesis research and speech recognition research are inter-related and complementary, one to the other.

Up to now, however, such a closely connected consideration method has not been taken into account in speech recognition. The proposed schemes of automatic speech recognition are classified into two large groups. One of them has been studied with the object of recognizing a limited vocabulary; ten figures, for instance.[25][26][53][55][58] Another one has been studied with the object of recognizing phonemes in any spoken words or sentences.[8][16][21][41][43] In both of these recognition schemes, the constants or the threshold values which are used in the processes of parameters extraction, segmentation and identification are optimized by an inspection or simple calculation. No recognition schemes have yet been constructed in a way to make it possible to optimize these values by the aid of the synthesis system. It is one of the most promising means to utilize the understanding of the speech perception mechanism by man for the

development of a powerful automatic speech recognition scheme by machine.

In order to improve the capability of the recognition system by the aid of perception experiments of synthetic speech, we would be desirable to construct some synthesis system based on the same processing scheme as the recognition system.

From the above-mentioned considerations, we were trying to construct a flexible composite speech research system consisting of a speech analysis-synthesis part and a speech recognition part based on the zero-crossing analysis methods. The fundamental idea of the speech analysis-synthesis part of this system is similar to the idea of wave-element expression used in the speech synthesis system by rule as described in the preceding chapter.

Particular attention has been paid to the following speech analysis-synthesis systems.

(1) vocoder method developed by IBM [2]
(2) PARCOR speech-synthesis method by Itakura [22]
(3) speech analysis-synthesis by linear prediction by B. S. Atal [1]

In these systems, original speech is transformed to compressed form, such as the output of filter banks and prediction codes. The compressed information is passed through the synthesizer whose operation is just the inverse of the input speech processing, and the speech sound is reproduced. These systems have many superior features concerning the quality of the output speech and the information compression ratio. The speech compression methods used in these systems can also be utilized in the pre-processing of speech recognition.

The main difference between these systems and our system is that the former three schemes do not include the segmentation process; while the latter composite system as will be described in the following two chapters accompanies the segmentation process in the analysis part and the recognition part aims to identify the phoneme-like units in spoken words.

4.2 Construction of the system(38)(52)

A schematic block diagram of the system is shown in Fig. 4.1. The speech analysis part is connected both to the speech analysis-synthesis and to the recognition part. Both the parameter extraction and the segmentation process are executed in the analysis part. The audio signal is pre-emphasized and fed into a computer through two parallel band pass filters. The output signal in each band is subjected to an infinite peak clipping wave by the computer. In this part, parameters are extracted every 12 msec (hereafter, called a frame), and the resultant sequence of the parameters is classified segmented and labeled as the voiced part, voiceless part and silence part and so-on. The voiced parts are further segmented into two; steady parts and transitional parts. The function of this stability detection is based on the measurement of time change characteristics of the extracted parameters.

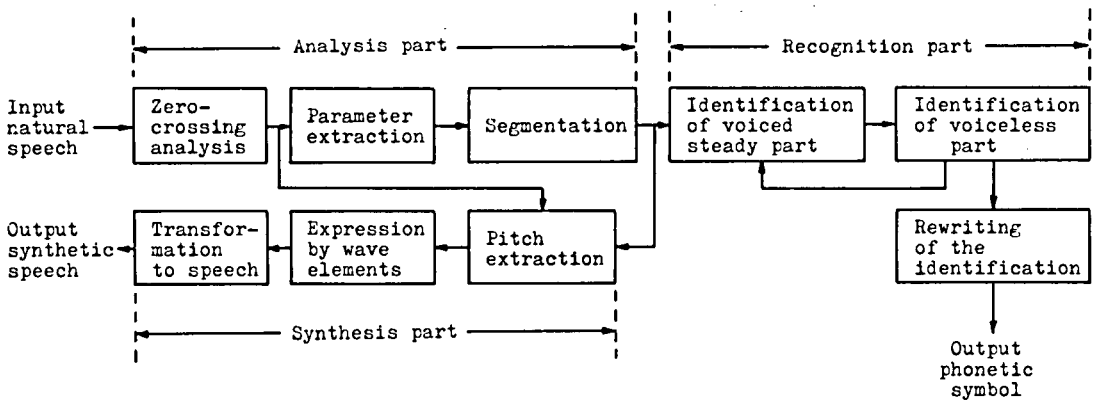


Fig. 4.1 Construction of the system

Speech analysis-synthesis is, in principle, speech reproduction from the compressed information extracted from an input sound. In retentive parts of a voiced sound, similar sequences of a zero-crossing interval (OXI) repeat at the pitch frequency. A fricative sound is noise-like and does not have any periodicity. The important feature of this sound is not the sequence order of

OXI, but the distribution of OXI. In other words, sequence of OXI in speech is very much redundant. From this point of view, we have tried in this system to compress the information of the sequence of OXI by substituting the sequence with the repetition of a sequence of OXI in a short time range such as a frame or a pitch (wave-element). The transformation of the compressed information into speech is achieved by assigning an amplitude to every wave-element and by repeating it. Here, the amplitude information is based on the average amplitude in the corresponding frame.

The speech recognition scheme as will be described in the following chapter is based on the phoneme recognition in spoken words or a sentence; this is necessarily accompanied with a segmentation process. The segmented part is recognized as either one of five vowels and eight consonant groups used. The voiceless consonants are identified as one of the four groups, according to the duration, average zero-crossing interval and so-on. A steady part of a voiced sound is, at first, decided on whether it is a vowel or a voiced consonant, according to its duration and the existence of a minimum amplitude in the part. The voiced sounds are recognized by Bayes' discriminant functions. Parameters such as the mean zero-crossing interval and the amplitude ratio of two channels, are used in the recognition of the vowels. In the case of the voiced consonants, time change characteristics of the amplitude parameters are added to the parameter vector.

The main information flow in the composite speech processing system is illustrated in Fig. 4.2. The connectors (boxes) in Fig. 4.2 correspond to the boxes (connectors) in Fig. 4.1. Input speech is first transformed to a sequence of zero-crossing intervals. From this sequence and other amplitude information, parameter extraction is executed. By using the sequence of extracted parameter sets, segmentation processing is executed and the sequence of parameters are classified segmented and labeled into one of several groups of phoneme-like units (gross feature). Based on this gross feature, in the speech synthesis part, wave-element expression is produced from the sequence of zero-crossing intervals; while, in the speech recognition part, phoneme recognition is performed by using the parameter sets.

Wave-element expression (result of the synthesis process) and sequence of

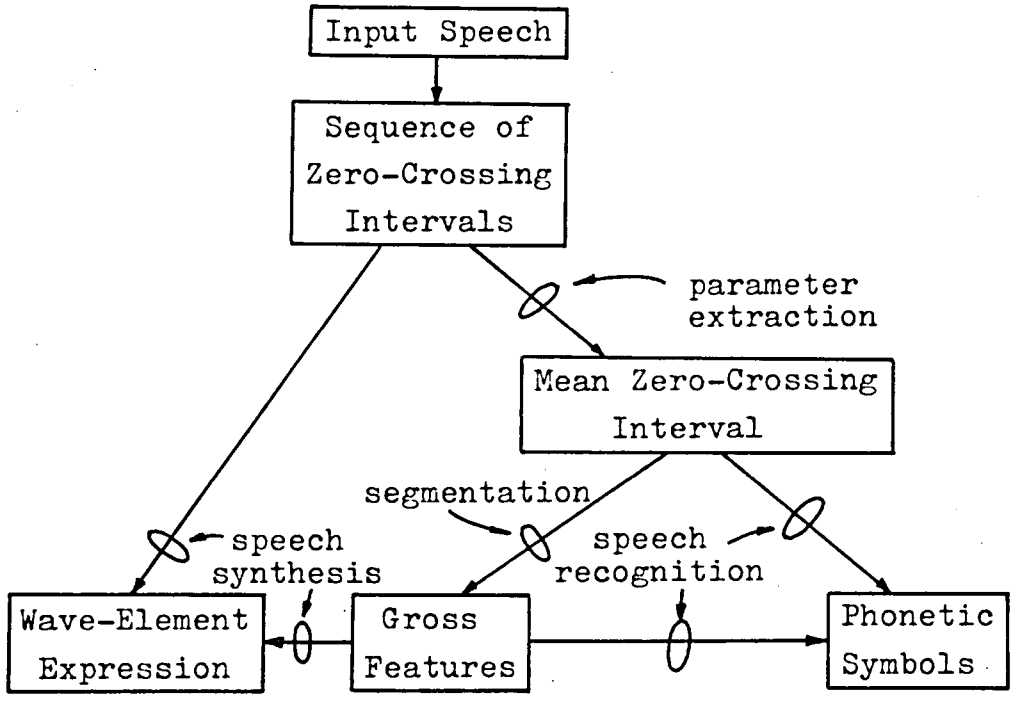


Fig. 4.2 Flow of information in the system

phonetic symbols (result of the recognition process) are related with each other through the gross features (result of segmentation) derived by the segmentation process. Intelligibility of the synthetic speech may much depends on preciseness or goodness of the gross feature, in the same way as the correctness of the sequence of phonetic symbols does. Therefore, the intelligibility or quality of synthetic speech becomes to be a sufficient condition of gross features for the identification of phonemes string in speech recognition. Besides, the extracted parameter by the zero-crossing analysis is indirectly affirmed by the quality of synthesized speech.

In order to modify the gross feature by some rule, it is hoped that the perception experiments of the synthetic speech will give an insight into a new speech recognition scheme.

4.3 Speech Analysis Part

4.3.1 Outline of the Speech Analysis Part

An outline of this part is illustrated in Fig. 4.3. In the analysis part, the audio signal filtered into two channels is fed into a computer and transformed to a zero-crossing wave. To extract the prosodic features used in the synthesis stage, a sequence of peak amplitude (a sequence of maximum amplitudes between two successive zero-crossing points) is also obtained from the input speech. Parameters such as the mean zero-crossing interval and the ratio of the rectified amplitude of two channels, are extracted every 12 ms (the time interval is called a frame). Next, the sequence of parameters is segmented into each of the voiced part (V), voiceless part (C) and silence part (X). Furthermore, the voiceless part is decided to be a fricative (F) or not, according to the mean zero-crossing interval. By the time change characteristics of parameters, the voiced parts are segmented into steady parts (S) and transitional parts (T). Gross features, that is, the results of this segmentation are used in the synthesis part and also in the recognition part.

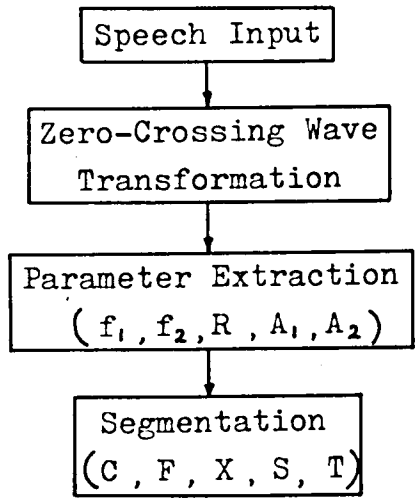


Fig. 4.3 Speech analysis part

4.3.2 Speech Input

It is well known that the formant frequencies are very important factors for the vowels and also for some voiced consonants to be identified. Especially, the five vowels of Japanese can, in most cases, be distinguished by the lower two formants. In this system, input speech sounds are passed through the circuits, as shown in Fig. 4.4, to pick up the frequency ranges of the first formant (CH1) and the second formant (CH2), prior to the zero-crossing analysis by computer. To cover the frequency region of the voiceless part, the highest range of the band-pass filter of CH2 is settled to 4800Hz.

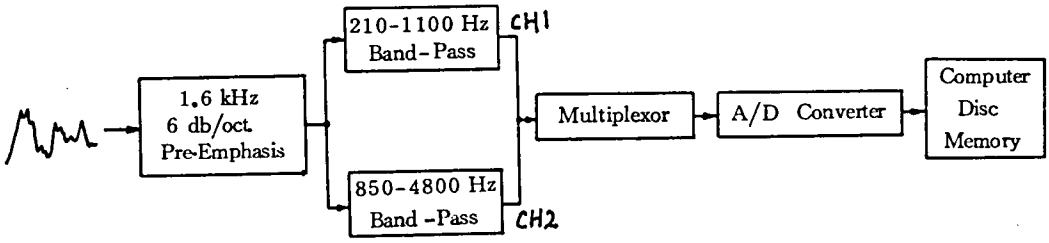


Fig. 4.4 Speech input circuit

The speech sound wave is, first, pre-emphasized by the RC high-pass filter (6db/oct., 1.6kHz cut off) to level off the amplitude of the frequency spectrum envelope, and is passed through the two sharp cut off band-pass filters whose cut off frequencies are 210–1100Hz and 850–4800Hz respectively. Next, the output waves are multiplexed and converted to a digital code by A/D converter, whose sampling frequency is 10kHz. Each sample is coded to 11 bits, and is transmitted to the auxiliary disk storage of a computer.

4.3.3 Parameter Extraction

Input speech waves are transformed to zero-crossing waves by a computer program. Following parameters are computed at regular time intervals (12 msec).

A_1: Average amplitude of rectified speech of CH1

A_2: Average amplitude of rectified speech of CH2

R: A_1/A_2

$\overline{\tau_1}$: Average zero-crossing interval of CH1

$\overline{\tau_2}$: Average zero-crossing interval of CH2

Parameter "R" is the amplitude ratio of CH1 to CH2. From $\overline{\tau_1}$ and $\overline{\tau_2}$, parameters f_1 and f_2 in a frequency domain are obtained by

$$f_1=1/(2\overline{\tau_1})$$
$$f_2=1/(2\overline{\tau_2}) \qquad (4.1)$$

Parameters f_1 and f_2 might give a measure of the center of the spectral energy present in the passband. Approximately, these correspond to the formant frequencies.

Hereafter, to indicate the frame number "i", these parameters are expressed as, for example, ${}_if_1$, ${}_if_2$.

4.3.4 Segmentation

Process of segmentation is shown in Fig. 4.5. In steps 1 and 2 in the figure, every frame is classified as the voiced frame (V), voiceless frame (C) and silence frame (X). In step 3, the voiceless part ("part" is used as a sequence of frames which are identified as the same category) is decided as to whether it is a fricative (F) or not. The voiced part is divided into a steady part (S) and a transitional part (T) in steps 4,5 and 6. Here, "A ➡ B" means that a frame satisfying the condition A is assigned to symbol B; while "A,B → C" means that if condition A is satisfied, then B is rewritten to C.

In steps 1 and 2, parameters A_1, A_2 and R are mainly used. Each frame is identified to be either V, C or X by the process of step 1; for example, ($R \geqq 2$ ➡ V) means that a frame whose R parameter is more than 2 is given by symbol V. Parameter R of the voiced frame V, except in the vowel [a], is ordinarily larger than any other frame. The silence frame is detected by the fact that both of parameters A_1 and A_2 are less than the pre-set threshold value. In step 2, the result of the decision at the step 1 is rewritten

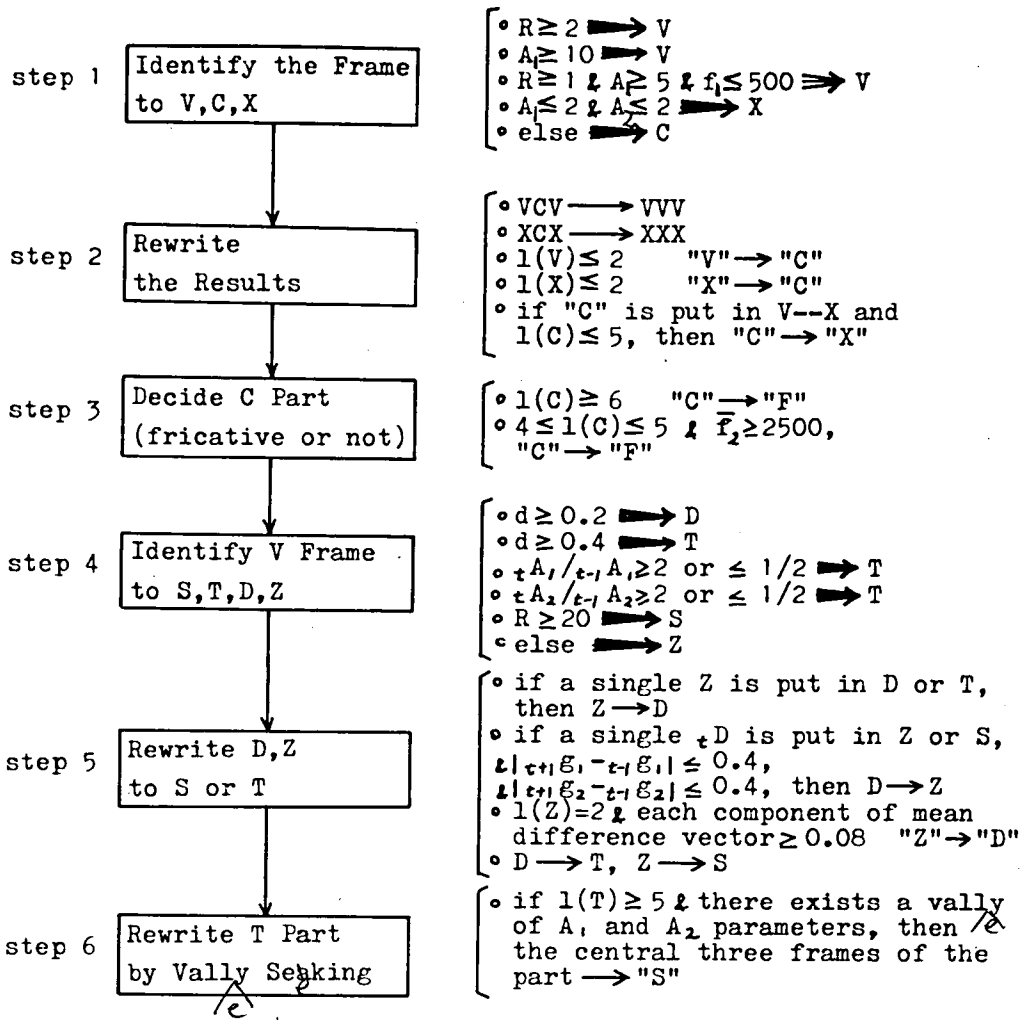


Fig. 4.5 Flowchart of the segmentation process

V: voiced frame "V": voiced part
C: voiceless frame "C": voiceless part
F: fricative frame "F": fricative part
X: silence frame "X": silence part
S: Z: voiced steady frame "S": "Z": voiced steady part
T: D: voiced transient frame "T": "D": voiced transient part
l: time length of the part
d: distance from preceding frame

according to its duration "ℓ" and the results of the neighboring frames; for example, VCV $\rightarrow$VVV means that C frame put between V parts is rewritten to V frame. By the process of step 3, the voiceless part is judged as to whether it is a fricative (F) or not, according to its duration and the value of f_2 parameter. If the duration of C part is more than 6, or if its duration is 4 or 5, and average f_2 parameter in the part is more than 2500, the C part is rewritten to F part.

In step 4,5 and 6, the voiced part is divided into steady parts and transitional parts. Euclidean distance "d_i" of two consecutive vectors ($_ig_1$, $_ig_2$) and ($_{i-1}g_1$, $_{i-1}g_2$) is used for this stability detection. Here, g_1, g_2 and d_i are computed by;

$$\begin{cases} g_1 = \log f_1/\sigma_1 \\ g_2 = \log f_2/\sigma_2 \end{cases} \qquad d_i = \sqrt{(_ig_1 - _{i-1}g_1)^2 + (_ig_2 - _{i-1}g_2)^2}. \tag{4.2}$$

The denominators σ_1 and σ_2 are the respective variance of logarism of f_1 and f_2 in the voiced parts, which are computed from many speech samples in advance.

Here, in order to remove the random variation of g_1 and g_2 parameters, sequences of g_1 and g_2 parameters are passed through low pass [illegible] operation and the remaining peak values in the smoothed sequence of parameters are substituted with the average of values of the neighboring frames.

In step 4, a voiced frame is classified as a steady frame (S), (Z) or a transitional frame (T), (D), according to the value "d_i" and the time change characteristic of the amplitude. If d_i is more than 0.2, then the frame is assigned to D. And if d_i is more than 0.4 or the amplitude time change ratio $_tA_j/_{t-1}A_j$ (j=1,2) is more than 2 or less than 0.5, then the frame is assigned to T. A frame having large R parameter ($\geqq 20$) is assigned to S, which generally corresponds to a steady part of voiced consonant. Others are assigned to symbol Z.

Next, in step 5, Z and D frames are rewritten as S or T frames, according to their duration and the contexts of the results obtained by step 4; for example,

a single Z frame put in D or T part is rewritten to D frame. It is generally difficult to detect the voiced consonants as steady parts by the process of steps 4 and 5. Therefore, in step 6, if there is a frame whose amplitude parameters A_1 and A_2 are minimum in the transitional part whose duration is more than 5, three central frames of that part are rewritten as steady frames.

4.3.5 Speech Materials used in the Experiments

Speech analysis-synthesis and recognition experiments were conducted with the 1900 Japanese words which had been recorded on a magnetic tape. These words were uttered by five announcers (hereafter abbreviated as SG, KB, NR, NK and UT). All of [vowel$_1$-vowel$_2$] and [vowel$_1$-consonant-vowel$_2$] contexts in Japanese are included in 380 words for each speaker. These words are listed in Table 4.1.

Table 4.1 List of words used in the experiments

V-V		V-y-V		V-w-V		V-p-V		V-t-V	
naate	名宛	kaya	蚊帳	kawa	川	papa	パパ	atama	頭
hai	灰					papirusu	パピルス		
au	会う	ayu	鮎			napukin	ナプキン		
mae	前					gapen	鉛ペン	atena	宛名
kao	顔	tayori	便り			aporo	アポロ	ato	後
kiatsu	気圧	iya	嫌	kiwa	きわ	baipasu	バイパス	itai	痛い
suii	推移					taipisuto	タイピスト		
jiu	滋雨	hiyu	比喩			paipu	パイプ		
chie	智恵					gi-pen	ギーペン	iteki	医学
shio	塩	iyo	伊予			nipondo	ニポンド	ito	糸
suashi	素足	uyamuya	うやむや	kuwa	鍬	jamupan	ジャムパン	uta	歌
ruiji	類似					karupisu	カルピス		
suu	吸う	huyu	冬			arupusu	アルプス		
sue	末					bo-rupen	ボールペン	uten	雨天
uo	魚	uyoku	右翼			ku-pon	クーポン	utoi	うとい
teate	手当	heya	部屋	kewashi	険しい	depa-to	デパート	geta	下駄
teire	手入れ					kepin	ケピン		
neuchi	値うち	ideyu	出湯			epuron	エプロン		
maee	前へ							ete	得て
teoke	手桶	seyo	せよ			repo-to	レポート	seto	瀬戸
soaku	粗悪	oya	親	kowai	こわい	popai	ポパイ	hotaru	蛍
koi	恋					kopi	コピー		
ou	追う	oyu	お湯	kuwawaru	くわわる	popura	ポプラ		
koe	声					opera	オペラ	kote	小手
anooka	あの丘	oyogu	泳ぐ			reopon	レオポン	kotori	小鳥

Table 4.1 List of words used in the experiment
(continued 2)

V-k-V		V-h-V		V-s-V		V-ʃ-V		V-ch-V	
aka	赤	haha	母	asa	朝	basha	馬車	amacha	甘茶
aki	秋	ahiru	あひる			ashi	足	hachi	蜂
aku	灰汁	ahure	溢れ	hasu	蓮	kashu	歌手	kachu	渦中
akebi	あけび	ahen	阿片	ase	汗				
tako	凧	aho	阿呆	aso	阿蘇	basho	場所	kacho	花鳥
ika	いか	ihai	位牌	isamu	勇む	isha	医者	mugicha	麦茶
iki	息	jihi	慈悲			ishi	石	ichi	一
iku	行く	ihuku	衣服	isu	椅子	bishu	美酒	ichu	意中
ike	池	ihen	異変	ise	伊勢				
ikoi	憩い	ihon	異本	iso	磯	isho	遺書	icho	銀杏
ukai	迂回	uha	右派	usagi	兎	musha	武者	mucha	無茶
uki	浮き	nuhi	奴婢			mushi	虫	uchi	内
uku	浮く	ruhu	流布	musume	娘	mushuku	無宿	uchu	宇宙
ukeru	受ける	uhen	右辺	usemono	失物				
ukon	右近	muhon	謀反	uso	嘘	keimusho	刑務所	ucho-ten	有頂天
sekai	世界	ehagaki	絵葉書	esa	えさ	eshaku	会釈	kechappu	ケチャップ
eki	駅	ehime	愛媛			deshi	弟子	echigo	越後
ekubo	えくぼ	ehu	絵符	tesuri	手すり	geshunin	下手人	meichu	命中
seken	世間	tehen	手偏	seseragi	せせらぎ				
neko	猫	ehon	絵本	heso	臍	kesho	化粧	techo	手帳
oka	丘	kohaku	琥珀	osa	長	dosha	土砂	ocha	お茶
oki	沖	hohitsu	補筆			hoshi	星	tochi	土地
oku	奥	sohu	祖父	osu	押す	toshu	徒手	tochu	途中
oke	桶	hohei	歩兵	osen	汚染				
soko	底	toho	徒歩	osoi	遅い	kosho	古書	kocho	胡蝶

Table 4.1 List of words used in the experiment
(continued 3)

V-ts-V		V-b-V		V-d-V		V-g-V		V-r-V	
		kaba	河馬	tada	ただ	taga	たが	sara	皿
		kabi	かび			hagi	萩	kari	狩
atsui	暑い	kabu	株			kagu	家具	taru	樽
		kabe	壁	tade	たで	hage	はげ	kare	彼
		kabocha	カボチャ	kado	かど	ago	あご	maronie	マロニエ
		ibara	茨	idai	偉大	jiga	自我	irai	依頼
		ibiki	いびき			igi	異議	kiri	霧
itsu	何時	ibuki	息吹			kigu	器具	chiru	散る
		kiben	詭弁	iden	遺伝	igen	威厳	kire	きれ
		ibo	いぼ	ido	井戸	igo	囲碁	iro	色
		uba	乳母	udaru	うだる	sugata	姿	ura	裏
		kubi	首			kugi	釘	uri	瓜
utsu	打つ	ubu	うぶ			hugu	ふぐ	uru	うる
		subete	全て	ude	腕	kuge	公家	mure	群
		kubomi	くぼみ	udo	うど	sugoi	すごい	kuro	黒
		tebata	手旗	eda	枝	kega	けが	tera	寺
		ebi	えび			negi	ねぎ	eri	襟
tetsu	鉄	ebumi	絵踏			megumi	恵み	keru	蹴る
		tebento	手弁当	medetai	めでたい	kegen	怪訝	terebi	テレビ
		eboshi	烏帽子	edo	江戸	negoto	寝言	tero	テロ
		oba	叔母	sodatsu	育つ	togame	咎め	sora	空
		obi	帯			nogiku	野菊	ori	檻
kotsu	こつ	kobu	こぶ			kogu	こぐ	oru	折る
		nobe	野辺	sode	袖	toge	刺	kore	これ
		sobo	祖母	odori	踊り	hogo	反故	doro	泥

Table 4.1 List of words used in the experiment
(continued 4)

V-z-V		V-dʒ-V		V-m-V		V-n-V	
aza	あざ	dajare	駄洒落	ama	海女	ana	穴
		aji	味	ami	網	ani	兄
azuki	小豆	kaju	果樹	tamuke	手向け	tanuki	狸
aze	あぜ			ame	雨	ane	姉
nazo	謎	majo	魔女	kamo	鴨	kano	彼の
hiza	ひざ	ijaku	胃弱	ima	居間	ina	いな
		iji	意地	imi	意味	hinichi	日にち
izu	伊豆	mijuku	未熟	imu	忌む	inu	犬
izen	以前			hime	姫	ine	稲
mizo	溝	bijo	美女	imo	いも	mino	みの
muzan	無残	kujaku	孔雀	uma	馬	unagi	うなぎ
		uji	宇治	umi	海	uni	うに
suzu	鈴	mujun	矛盾	umu	生む	kunugi	くぬぎ
huzei	風情			ume	梅	une	うね
buzoku	部族	hujo	婦女	kumo	雲	uno	宇野
mezashi	目ざし	tejaku	手酌	tema	手間	tenarai	手習い
		ejiki	餌食	zemi	セミ	zeni	銭
kezuru	削る	tejun	手順	temukai	手向い	tenukari	手抜り
zeze	膳所			seme	攻め	genetsu	解熱
ezo	蝦夷	gejo	下女	memo	メモ	enoki	榎
tozan	登山	amanojaku	あまのじゃく	koma	駒	kona	粉
		oji	叔父	tomi	富	oni	鬼
mozu	百舌鳥	nojuku	野宿	nomu	飲む	konu	来ぬ
sozei	相撲			kome	米	one	尾根
hozo	枸	hojo	補助	tomo	友	ono	谷

As objects of speech analysis and/or speech recognition, special groups of words or sentences such as digit numbers, station names, and programming statements have been frequently used. However, these materials can not cover many contexts which appear in a natural language. We hope that words in Table 4.1 which we selected will be one good set of materials when a experimental system to be tested with arbitrary context.

4.3.6 Results and their Discussion

Results of the segmentation of consonants are listed in Table 4.2 and 4.3. The Table 4.2 shows the results of consonants in the top position of words, while the Table 4.3 shows the results of consonants in the inside position of words. In Table 4,2, the score contained in "ϕ" row of the first column means that vowels were segmented as sounds in the form of voiceless consonant+vowel, and the scores contained in "ϕ" column except in the first row mean that the consonants failed to be detected. The column "silence+voiceless" in Table 4.3 means that the consonants were segmented as the silence part and the succeeding voiceless part (C) or (F). The contents of positions having star(*) symbols denote the correct segmentation.

Table 4.2 Results of segmentation
(consonants in forefront of words)
star symbols * denote the correct segmentation

output / input	voiceless part		voiced part		ϕ
	C part	F part	S part	T part	
ϕ	2.9				97.1 *
p, t, k	71.5 *	22.0 *			6.5
s, ʃ, c	8.4	90.5 *			1.1
b, d, g, r	8.8		80.4 *	10.8	
m, n	0.6		78.3 *	20.0	1.1
z, d_3	5.1	12.8 *	82.1 *		
h	46.2 *	23.1 *	18.0		12.7

Table 4.3 Results of segmentation
(consonants in inside of words)
star symbols * denote the correct segmentation

output / input	voiceless part		silence part + voiceless part		voiced part			others
	C part	F part	X + C part	X + F part	S part	T part	S+C, F part	
p, t, k	2.2	1.5	91.4*	4.9				
h	23.2*	60.8*	4.8		10.5	0.7		
s, ʃ		95.0*		5.0				
c	2.4		5.6	92.0*				
b, d	7.2				87.0*	0.5	5.3	
r					88.8*	3.2		8.0
m, n					82.8*	7.6		9.6
$\tilde{g}$					66.4*	12.0		21.6
z, d_z	26.5*	16.3*			40.5*	3.5	8.2*	5.0

Main results are summarized as follows.

(1) Errors of segmentation of the voiceless consonants are mainly caused by detection error of the silence parts, and these error rates largely depend on individual speakers.

(2) About 18% of [h] in the top position of words and 10% of [h] in the inside position of words are misjudged as voiced sounds. All of them are not erroneous, because [h] is, in some cases, uttered with voicing. This confusion is one of the most difficult problems in automatic speech recognition.

(3) Segmentation of the voiced parts is generally more difficult than that of the voiceless consonants. Up to 1/3 of nasal [$\tilde{g}$] sounds can not be segmented from the neighboring vowels, particularly, from the vowel [u].

(4) Some of the burst parts in [b] and [d] were judged as the voiceless part,

almost all of which are in the utterances by NR and NK. On the other hand, in about half the voiced fricative [z] and [d_3], the voiceless parts were detected.

The results of segmentation or gross feature detection heavily depend upon individual speakers and the intensity level of their utterances. The performance of segmentation will be improved by personal adjustment of parameters, particularly, R used in the analysis part. All the scores in the positions having "*" symbols in Table 4.2 and 4.3 are, on the average, 86.3% and 86.5% respectively.

4.4 Speech Synthesis Part

4.4.1 Outline of the Speech Synthesis Part

In this section, we will discuss the speech information compression method by wave-element expression. The speech sound is reproduced by the compressed information, that is, a sequence of wave-elements and their attached parameters. Here, a wave-element corresponds to two sequences of zero-crossing intervals (CH1 and CH2) in a frame or a pitch period.

A flow chart of the synthesis part is presented in Fig. 4.6. In the synthesis part, first, pitch extraction in the voiced parts is executed and the sequence of extracted pitch intervals is smoothed by the low-pass operation. Next, every section segmented by the analysis program is expressed by a sequence of sets of wave-element, its amplitude and repetition parameter. A wave-element expression of a voiced steady part is transformed into speech by repetition of a wave-element, the amplitude of the wave being modified at every repetition. A wave-element consists of two sequences of zero-crossing intervals. Repetition parameter is identical to both channels. However, the amplitude parameters of these channels have generally distinct values. Also the pitch interval of the synthesized speech is controlled by manipulation of the length of the sequence of zero-crossing intervals.

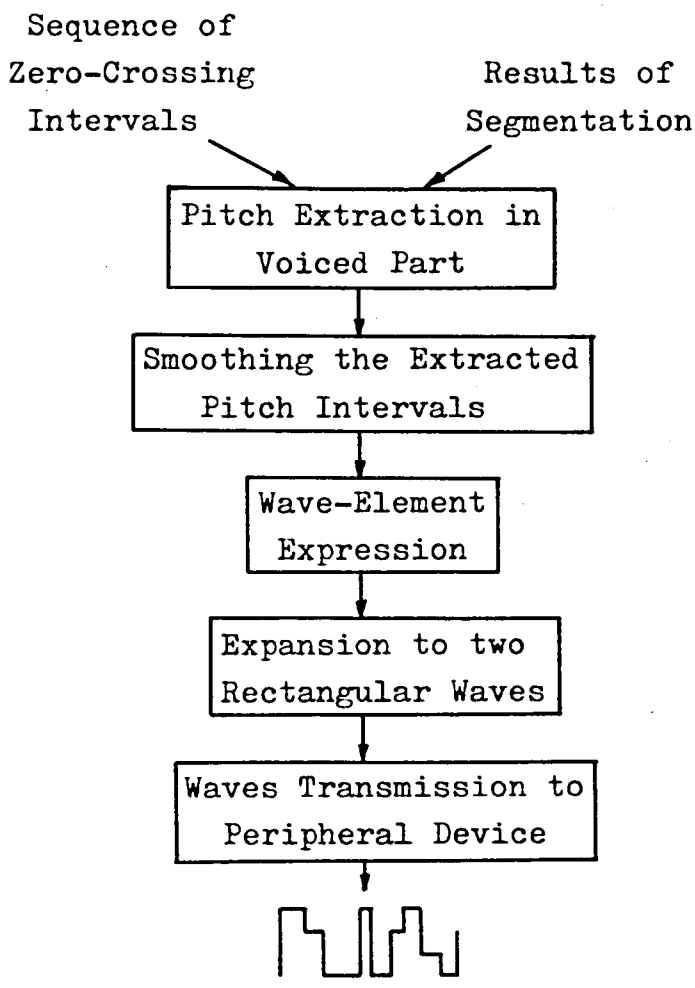


Fig. 4.6 Speech synthesis part

4.4.2 Pitch Extraction

In the voiced part, pitch intervals are extracted so as to be used for the wave-element expression. The extraction procedure is similar to the method as described in section 2.3.2.

At first, the zero-crossing interval (OXI^i) with maximum peak amplitude is taken out from a leading frame in a voiced part. Next, the zero-crossing interval (OXI^j) with maximum peak amplitude is taken out from the sequence in a time range $0.7T \sim 1.3$ T counted from the OXI^i (T is the average of three

preceding pitch intervals). The time interval between OXI^i and OXI^j is the calculated pitch interval. We can obtain the sequence of pitch intervals in the voiced part, repeating the operation by substituting OXI^i with OXI^j.

4.4.3 Wave-Element Expression

In Table 4.4, the wave-element expression of the segment is listed. Wave-element expression is similar to that in Table 3.1. Here, P is a wave-element and r is the corresponding repetition parameter. Star symbol attached to r parameter means an operation of random shuffling of the wave-element. Vector expression α means a set of amplitude parameters of CH1 and CH2.

The silence part (X) is expressed by repetitions of a wave-element with zero amplitude. For the voiceless part (C), all of the sequences of zero-crossing intervals in the corresponding part of original speech are used as they are. In other words, all of the repetition parameters are one. The amplitude information is generally changed, according to the A_1 and A_2 parameters of the frame. The fricative part (F) is expressed as repetitions of a single wave-element consisting of the sequences of zero-crossing intervals in the center frame of the corresponding part of original speech. The sequences of zero-crossing intervals are shuffled at random at every repetition in order that the periodicity may not appear in the sequence. This special manipulation is distinguished by the star symbol in Table 4.4.

Table 4.4 Wave-element expression

part	wave-element expression
C	$\alpha_1 P^1_{c1} \cdot \alpha_2 P^1_{c2} \cdots\cdots \alpha_r P^1_{cr}$
F	$(\alpha_1, \alpha_2, \cdots\cdots, \alpha_r) P_f^{r*}$
X	$(\overbrace{\emptyset, \emptyset, \cdots\cdots, \emptyset}^{r}) P_x^r$
S	$(\alpha_1, \alpha_2, \cdots\cdots, \alpha_r) P_s^r$
T	$\alpha_1 P^1_{t1} \cdot \alpha_2 P^1_{t2} \cdots\cdots \alpha_r P^1_{tr}$

The steady voiced part of speech is made by repetitions of a wave element which consists of the sequences of zero-crossing intervals (CH1 and CH2) of the central pitch period in the part. In order to reserve the prosodic information of the input speech, the time length of the wave-element in the steady voiced part is matched to the extracted pitch interval smoothed by the low-pass operation. Thus, to raise the pitch frequency, the tailing zero-crossing

intervals are erased, and in order to lower the pitch frequency, the leading zero-crossing intervals in the pitch are added to the end of the pitch. The amplitude of the zero-crossing wave is modified by A_1 and A_2 parameters at every repetition. The transitional voiced part (T) is processed in the same manner as the voiceless part (C) only with the difference that the wave-element in the part corresponds to a pitch interval.

4.4.4 Speech Output

Two rectangular waves are obtained from the sequence of wave-element, amplitude information and number of required repetitions. The amplitudes of the rectangular waves are changed at the unit of a frame or a pitch interval. The waves are transformed into a speech sound wave in an on-line mode by the peripheral device attached to the computer.

The peripheral device is shown in Fig. 4.7. The two sequences of digital samples are converted to analog value by two sets of D/A converters, and filtered by the low-pass filter (900 Hz cut-off, 12db/oct.) succeeded by a high-pass filter whose cut-off frequency is 200 Hz (CH1), or by the high-pass filter (1100 Hz cut-off, 12db/oct.) preceded by a low-pass filter whose cut-off frequency is 4800 Hz (CH2). These waves are added and the high frequency components are de-emphasized by the RC low-pass filter (6db/oct., 1.6kHz cut off), which compensates the pre-emphasis characteristic of the speech input device shown in Fig. 4.4. The speech filtering circuit can remove a certain degree of distortion included in the rectangular waves.

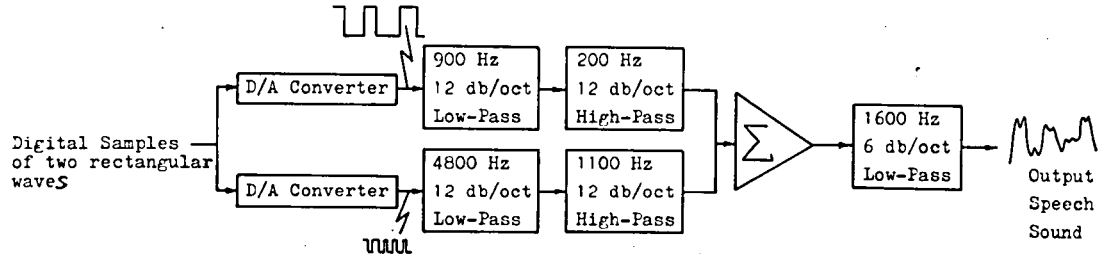


Fig. 4.7 Peripheral speech output circuit

4.4.5 Synthesized Speech Sounds

Sonagrams of the input natural speech and its synthetic speech are presented in Fig. 4.8 [ijaku](胃弱), Fig. 4.9 [kiben](詭弁), Fig. 4.10 [sode](袖), and Fig. 4.11 [tebento](手弁当).

The synthetic speech is reproduced by the sequence of zero-crossing intervals. However, the spectral pattern of the synthetic speech shows fairly good consistency with that of the input speech. The main reason for the fact is due to the usage of two channels of zero-crossing waves. In these figures, results of the segmentation process are also presented. The number attached to the identified symbols shows the number of frames in the segments.

In Fig. 4.8, phoneme [d_3] is segmented as C part whose duration is two. And [y] consists of voiced steady part and transitional part. [k_i] in Fig. 4.9 is constructed with F frame which is repeated eight times with random shuffling of the wave-element. Syllabic nasal [η] consists of a sustained S part. Steady part is detected in [d] of [sode] in Fig. 4.10. In Fig. 4.11, steady part missed to be detected from phoneme [e] in the second syllable of the word [tebento].

4.4.6 Intelligibility of Synthesized Speech

The afore-mentioned 1900 words were synthesized. The intelligibility of synthetic speech was tested by using a tape having random arrangements of these words. In the test, we used the speech materials which were synthesized not by using the above-mentioned speech output circuit, but by a computer program which added directly two rectangular waves and converted them to analog signals. The output waves were de-emphasized by the low-pass filter. Each word was arrayed with one second interval and presented twice to the listeners in a quiet room. The listeners were eight males who had not received any special training.

Results of the word articulation test are given in Table 4.5. The rate of correct response was about 82.6%. These results showed about 8% difference among the listeners. Of course, the words which have errors in the segmentation

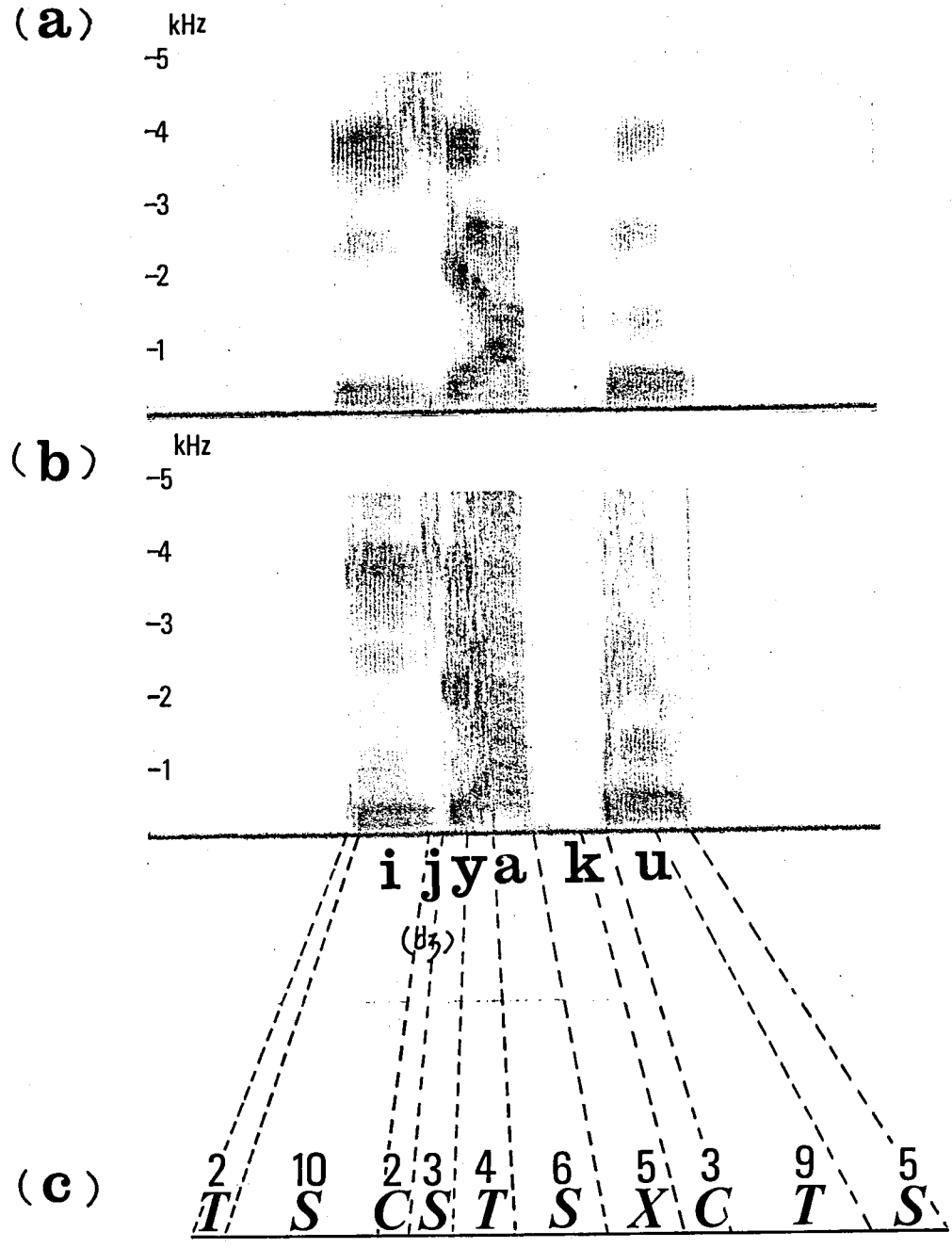


Fig. 4.8 Synthesized speech [ijaku] (胃弱)

(a) sonagram of natural input speech

(b) sonagram of synthesized speech

(c) result of segmentation

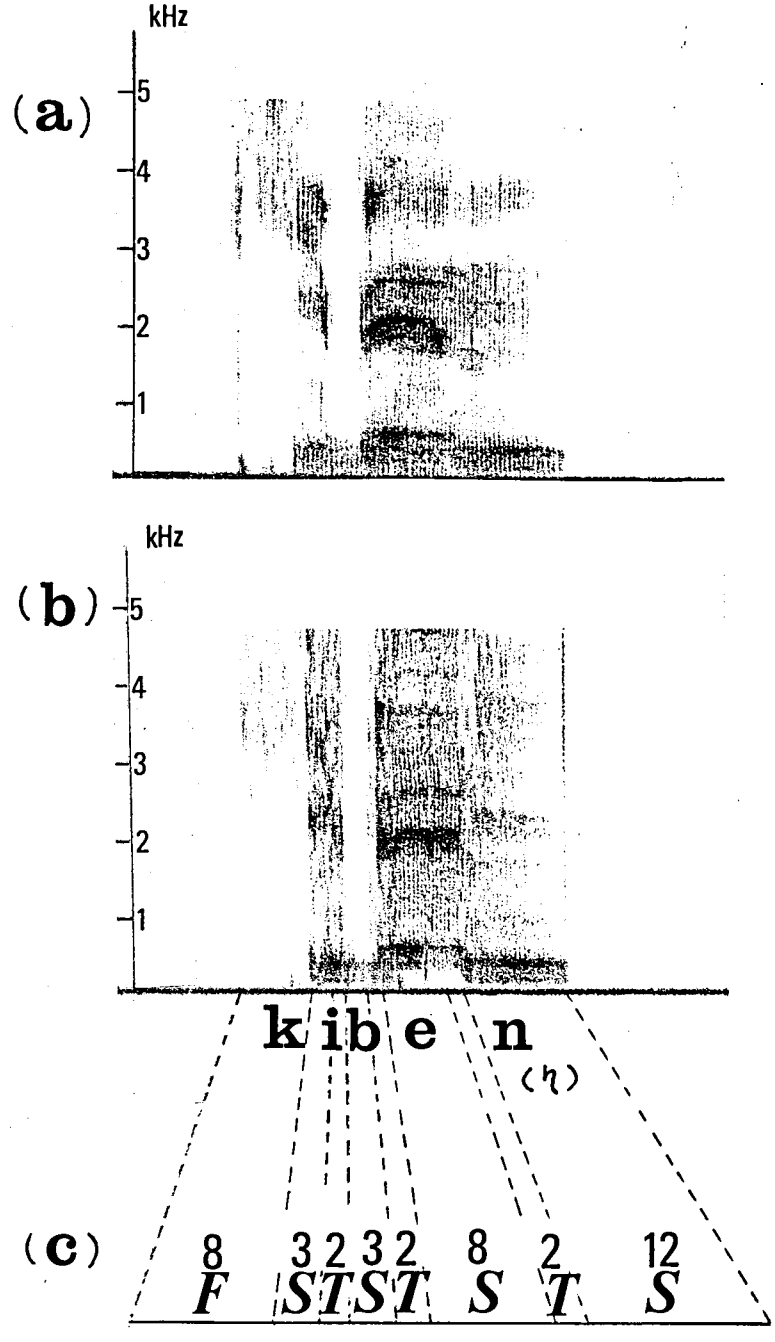


Fig. 4.9 Synthesized speech [kiben] (詭弁)

(a) sonagram of natural input speech

(b) sonagram of synthesized speech

(c) result of segmentation

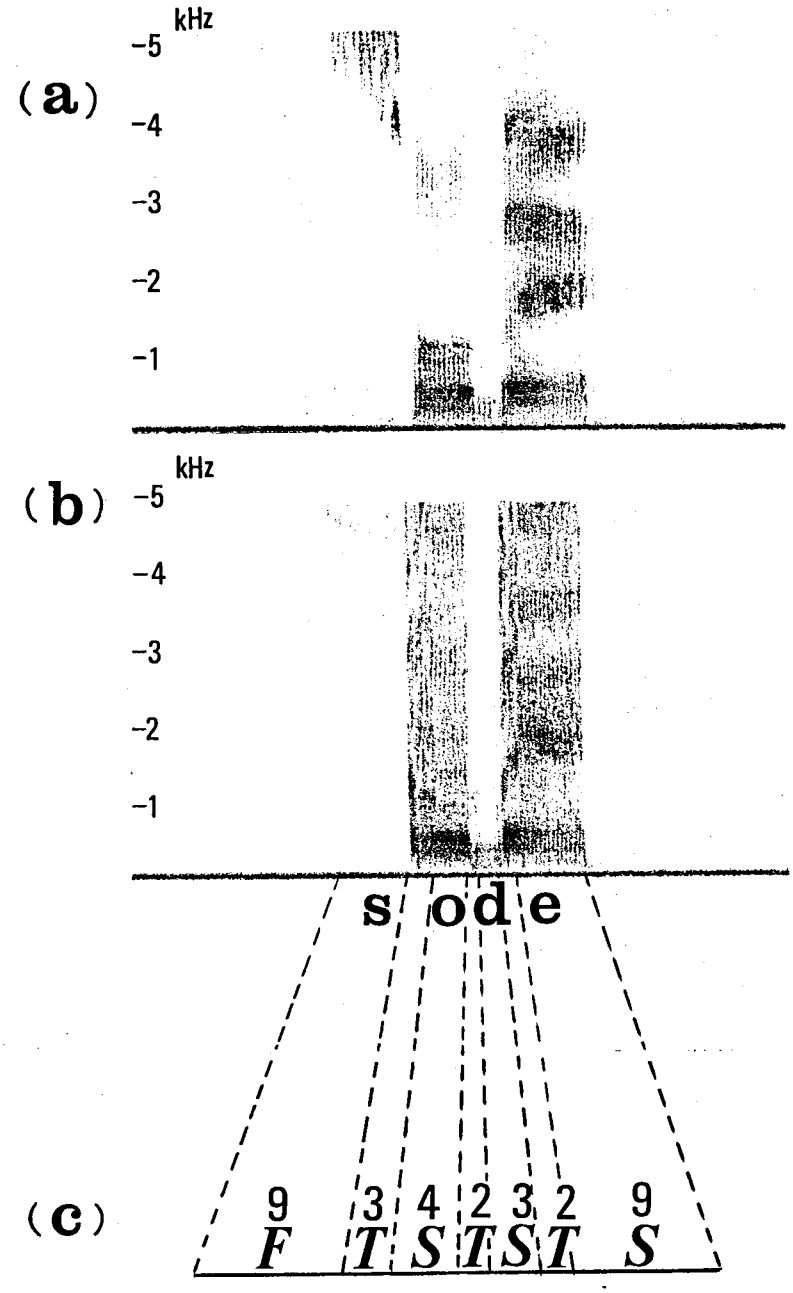


Fig. 4.10 Synthesized speech [sode] (袖)

(a) sonagram of natural input speech

(b) sonagram of synthesized speech

(c) result of segmentation

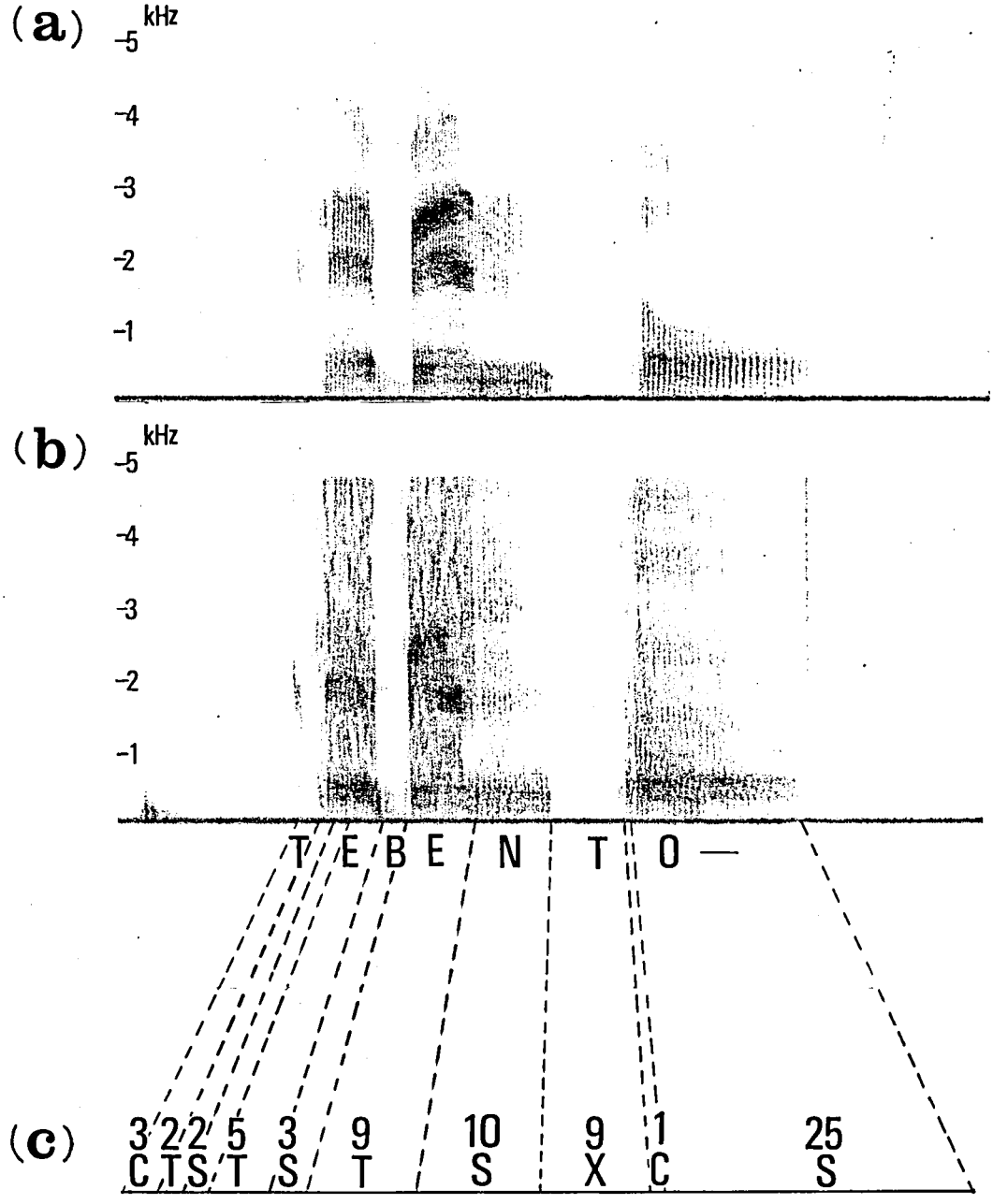


Fig. 4.11 Synthesized speech [tebento-] (手弁当)

(a) sonagram of natural input speech

(b) sonagram of synthesized speech

(c) result of segmentation

process were apt to be mis-understood.

Table 4.5 Results of hearing test of the synthesized speech

speaker \ participant	KT	KN	KW	AM	NA	HS	KS	TG	mean
SG	83.9	87.8	83.1	81.7	80.3	81.9	79.2	77.2	81.9
KB	89.1	89.6	84.4	84.2	84.4	80.1	84.7	81.4	84.7
NR	87.8	86.7	86.4	81.4	80.9	83.5	78.2	80.1	83.1
NK	85.4	81.6	86.4	82.4	79.5	82.9	77.1	78.9	81.8
UT	88.9	85.2	85.0	78.0	81.9	78.0	77.7	78.3	81.6
mean	87.0	86.2	85.1	81.5	81.4	81.3	79.4	79.2	82.6

The results of the hearing test were not so good as we expected. However, we can obtain more intelligible speech by using the peripheral speech output circuit shown in Fig. 4.7. **though we do not have ascertained it precisely.**

4.4.7 Results and their Discussion

Rather intelligible speech was obtained by the speech analysis-synthesis method in the present system. Japanese words were synthesized and their intelligibility was ascertained to be good. By this system, we can also construct a sentence speech synthesizer. The highly intelligible speech in Japanese and also in English were synthesized by the system.

Information rate of the wave-element expression is about 7.25 kbits/sec. on the average with regard to the 1900 words used in the experiment. This corresponds to about 30% of the information rate of the original zero-crossing wave. Compared with PCM coding (7 bitsX10,000 samples), it corresponds to about 1/10 information-compression ratio.

4.5 Conclusion

The speech analysis-synthesis part in the composite speech processing system was described, and also results of the segmentation and speech synthesis were shown.

The fundamental idea of the analysis-synthesis method consisted in a description of speech by the wave-element expression, that is, by element (wave-element) and their concatenating rules. The high degree of intelligibility and also naturalness were obtained, no matter how the synthetic speech wave was reproduced from the zero-crossing waves. The reasons are that,

(1) **wave-elements which are cut out from a input natural word in the analysis program are used as the wave-elements for synthesis of that word in the synthesis program.**

(2) the intonation of the input speech is applied to the synthetic speech as it is.

(3) sequences of the zero-crossing intervals separated into two channels are used in the synthetic speech.

By concatenating the synthetic words, we can obtain a sentence speech output system with high quality. Because the information quantity of the synthetic words is about 1/10 of the natural speech, the small memory capacity is sufficient for the actual usage of this speech output scheme.

102 項欠

Chapter 5 Speech Recognition System

5.1 Introduction

We will describe a speech recognition system which aims at the recognition of a vocabulary of rather many words. Phonemes string in a spoken word is recognized according to the results of segmentation in the speech analysis part as described in Chapter 4.

It is possible to recognize a limited set of words or sentences, even if the phonemes in the word or sentence cannot be identified uniquely. Besides, the judgement of the places of articulation, as used in discriminating between /m/ and /n/, is more difficult than discriminating other phonemic features such as the manner of articulation. Therefore, in this system, we classified the Japanese consonants into following eight groups, according to the manners of articulation, and did not distinguish the phonemes in a group.

(1) nasal /m/, /n/, /$\tilde{g}$/, (2) voiced plosive /b/, /d/, /g/
(3) liquid-like /r/ (4) voiced fricative /z/, /d_z/
(5) voiceless plosive /p/, /t/, /k/
(6) africative /c/
(7) voiceless fricative /s/, /ʃ/, /h_i/
(8) aspirated /h/

Phonemes in spoken words are classified into each of the five vowels, the syllabic nasal [η] and the eight consonant groups. Groups (5) and (8) and groups (6) and (7) are not distinguished in the forefront of words.

5.2 Outline of the Recognition Part[38][52]

The flowchart of the recognition system is shown in Fig. 5.1. Results of the segmentation and the extracted parameters are used in this part.

The voiced steady part is first judged as to whether it belongs to the vowel group or the voiced consonant group. If there exists a valley (minimum

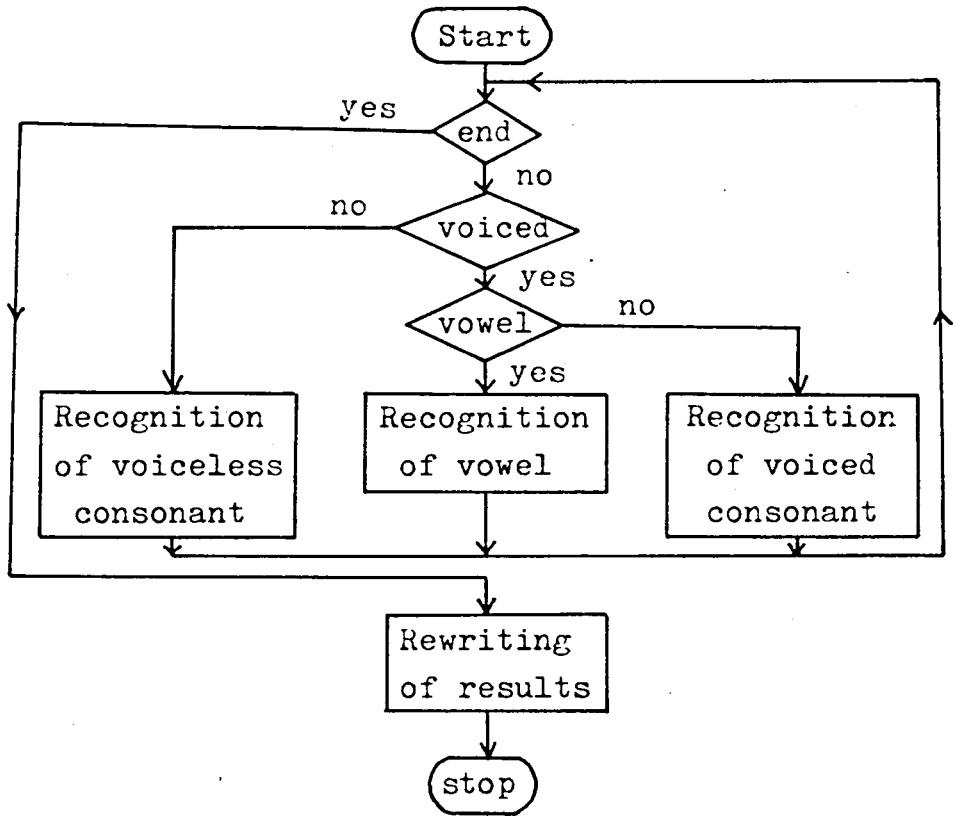


Fig. 5.1 Flowchart of the recognition system

amplitude) of parameters A_1 and A_2 in a steady part in inside of a word and the duration of the part is less than a certain threshold value, the part is decided to be one of the voiced consonant groups. In the case of the steady part in the forefront of a word, parameters A_1, R, f_1 are used in the judgement. If average A_1 and f_1 are smaller than certain thresholds ($A_1 < 30$, $f_1 < 500$) and R is larger than a certain threshold ($R > 2$), the part is judged as a voiced consonant.

The voiced parts are recognized by Bayes' discriminant functions. The parameters, such as the mean zero-crossing interval and the amplitude ratio of two channels as will be described in next section are used in the recognition of vowels. In the case of voiced consonants, time change characteristics of amplitudes are included in the parameter vector.

Voiceless consonants are identified as one of the four groups, according to the gross classification obtained from the segmentation process. Discrimination

of a aspirated sound from fricative sounds is based on its duration and the average value of f_2 parameter.

Recognition results thus obtained are re-written in the last step of the recognition program.

5.3 Recognition Method

5.3.1 Discrimination of Vowels

For recognition of the vowels and syllabic nasal [η] are used logarism of parameters R, f_1, f_2 and f_S (calculated in the same manner as in f_1 and f_2 from the wave which is obtained by averaging the abutting two samples of CH2).

Let x be the vector expression in log scale of the above-mentioned four parameters which are averaged in a voiced steady part. As described in section 2.4.5, Bayes' discriminant functions are given by

$$h_i(x) = -(x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i) - \log |\Sigma_i| \qquad (i = 1, 2, \cdots\cdots 6) \qquad (5.1)$$

Here, the subscript "i" represents the categories. μ_i is the average of vector of the category "i". Σ_i represents the covariance matrix of the category "i". Assuming that the distribution of the pattern vector x is normal and that the frequency of appearance of each category is even, the function h_i of x is the discriminant function of the optimum classifier. The input pattern x is recognized as the category "i" which has the minimum value of $(-h_i)$. Exceptionally, we add the value 2 to (h_i) of syllabic nasal [η], because the frequency of appearance of [η] is much smaller than any other vowels.

As an example, μ_i and Σ_i calculated from selected frames in about 150 words by five announcers are shown in Table 5.1 and 5.2. **Contents of Table 5.2 show values 10^4 times as much as the actual values for convenience.**

R parameter of [a] is smaller than any other vowels. Vowel [i] has the smallest f_1 and largest f_2 parameter of all categories. f_S parameters have middle values between f_1 and f_2. Distribution of parameter R is found to be

large in Table 5.2; while distributions of parameters f_1 and f_2 are rather small, except those of f_2 of [u] and [o]. According to Table 5.2 parameters of syllabic nasal [η] distribute in a wide range because they heavily depend on speaker's voice traits.

These were used in the recognition experiment (II) to be described in the following section.

Table 5.1 Average of parameters of vowels

category	parameters ($\log_e$)			
	100 x R	f_1	f_2	f_s
a	4.0791	6.7775	7.1138	7.0757
i	5.1684	5.6976	7.9710	7.4736
u	5.9037	5.8902	7.3985	7.1795
e	5.0856	6.1846	7.6638	7.5086
o	6.0186	6.2076	7.1087	6.8654
η	6.0128	5.7018	7.5036	7.2389

5.3.2 Discrimination of Voiced Consonants

In recognizing voiced consonants, parameters α and β which represent the time change characteristics of A_1 parameter are used, in addition to the R, f_1 and f_2. The discriminant functions of such voiced consonants are similar to those of vowels, except that the set of parameters is different.

Parameters α and β are defined as

$$\alpha = \max\{{}_{j+1}A_{1,\,j+2}A_{1,\,j+3}A_1\}/{}_jA_1$$
$$\beta = \max\{{}_{j-1}A_{1,\,j+1}A_1\}/{}_jA_1 \qquad (5.2)$$

Table 5.2 Covariance matrices of vowels

Σ_a

$100 \times (\log_e)$	100 x R	f_1	f_2	f_S
100 x R	799.929	-79.887	-53.273	-85.066
f_1		206.103	-58.446	-56.818
f_2			111.671	75.014
f_S				87.722

Σ_i

100 x R	3641.047	-34.499	-310.561	96.850
f_1		186.632	43.383	26.489
f_2			182.696	-91.863
f_S				119.679

Σ_u

100 x R	4586.907	-112.062	-468.303	-547.284
f_1		281.425	-128.412	-50.349
f_2			922.003	408.051
f_S				401.225

Σ_e

100 x R	2517.865	-84.601	31.704	-95.003
f_1		108.590	3.921	1.080
f_2			113.526	15.209
f_S				54.733

Σ_o

100 x R	3699.937	-343.536	795.193	100.821
f_1		168.995	-40.737	-18.165
f_2			1254.006	206.768
f_S				140.721

Σ_η

100 x R	6595.045	-954.104	-0.512	-348.296
f_1		525.368	-131.658	-59.227
f_2			362.979	245.384
f_S				369.375

Here, the "j"th frame has the minimum A_1 parameter in the voiced steady part. Average vector of parameters in log scale, μ_j, are shown in Table 5.3. Voiced plosive and voiced fricative sounds take, in general, a comparatively large number of parameter α. On the contrary, liquid-like sound takes a large value of parameter β. In nasals, parameters α and β are also smaller than in the other voiced consonants. Voiced plosive and fricative sounds take small values of f_1. The f_2 parameter of fricative sound is larger than that of vowel [i]. Σ_j for voiced consonants used in the recognition experiment (II) are shown in Table 5.4. Values in almost all entries are larger than those of vowels. **Here, the contents of Table 5.4 show values 10^4 times as much as the actual value for convenience.**

Table 5.3 Average of parameters of voiced consonants

Parameters ($\log_e$)

category	f_1	f_2	100 x R	100 x α	100 x β
b, d, g	5.73570	7.33975	6.30615	7.19243	5.46023
r	5.96576	7.45168	4.96200	6.21784	5.55495
z, d_3	5.74621	8.01924	4.28919	6.69024	5.23378
m, n, $\tilde{g}$	5.85831	7.49081	5.67737	5.03728	4.82445

5.3.3 Discrimination of Voiceless Consonants

Recognition of voiceless consonants depends mainly on the results of segmentation. A flowchart of the recognition of voiceless consonants at the middle part of the words used is shown in Fig. 5.2. In a right half of the flowchart, plosive, africative, and fricative preceded by silent sounds are recognized. For example, voiceless plosives consist of voiceless part (C) preceded by the silence part (X). An africative sound consists of a short fricative part (F) preceded by the silence part (X). When a fricative part whose duration is longer than 10 frames follows a silence part, it is judged as plosive sound succeeded by fricative sound. This case is found in, for example, [arupsu](アルプス).

Table 5.4 Covariance matrices of voiced consonants

$\Sigma_{b,d,g}$

$100 \times (\log_e)$	f_1	f_2	100 x R	100 x α	100 x β
f_1	895.005	-210.841	-1515.316	300.529	688.712
f_2		1341.094	-288.580	33.350	157.292
100 x R			13131.367	-269.932	-4171.460
100 x α				6448.902	1877.530
100 x β					4000.582

Σ_r

f_1	954.964	-429.489	-850.315	-191.366	157.829
f_2		627.612	212.942	311.511	-24.985
100 x R			3354.432	25.180	-748.967
100 x α				3498.117	1295.431
100 x β					1870.583

Σ_{z,d_3}

f_1	690.548	-85.180	-171.344	315.333	204.090
f_2		602.504	-1513.678	479.924	292.353
100 x R			8760.529	-5297.305	-2133.724
100 x α				10491.137	2548.249
100 x β					2069.434

$\Sigma_{m,n,\tilde{g}}$

f_1	487.437	-126.008	-720.239	-12.309	67.103
f_2		484.303	-29.428	-34.672	-58.487
100 x R			4142.906	-203.275	-162.833
100 x α				845.718	145.287
100 x β					395.507

The C or F part which does not follow the silence part is judged as fricative sound or aspirated sound. Discrimination of fricative sounds from aspirated sounds depends on whether the average f_2 parameter in the part is greater than the threshold (3kHz) or not. F part whose f_2 parameter is greater than 3 kHz is recognized as fricative sound.

Because voiced fricative sound is frequently segmented as C or F part

(see Table 4.3), possibility of the voiceless part to be voiced fricative is tested in this routine. When all f_1 parameters are between 200 Hz and 500 Hz and also R parameter are more than a certain threshold (0.2) in the part, the part is identified as a voiced fricative sound.

When a voiceless consonant is to be recognized in the leading position of a spoken word, the existence of the silence part (X) is not used. Therefore, the voiceless consonant is recognized to be a fricative or not by its length and f_2 parameter.

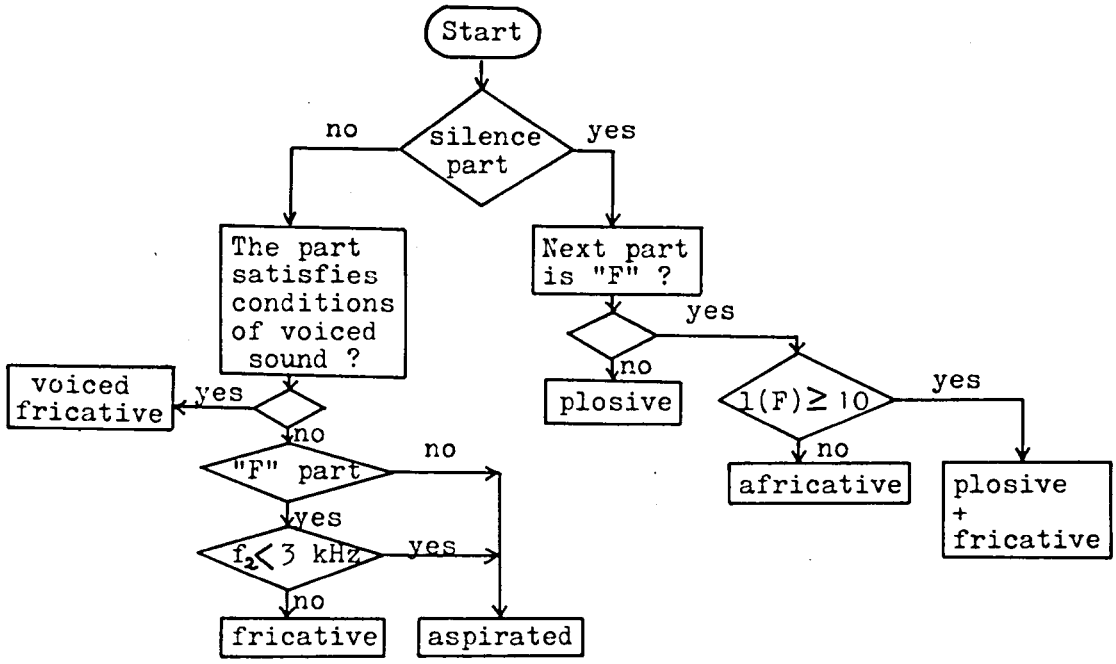


Fig. 5.2 Recognition tree of voiceless consonants

5.3.4 Rewriting of Results

In the rewriting part, adjoining vowels which were identified as the same category are combined into a single vowel. Besides, if two vowels are adjoining

and the identified category of one vowel is the secondary preferable category of the another vowel part and if the total length of the adjoining vowels is less than a certain threshold, the adjoining two parts are combined into a single vowel which has a larger value of h_i.

If there exists a (C) part or a (F) part in a succeeding close region of a voiced fricative consonant, the identified result of the voiceless part is eliminated.

5.4 Recognition Experiments of input Speech

The following three recognition experiments were conducted with the 1900 words spoken by five male announcers shown in section 4.3.5 and with 500 words by ten male students.

5.4.1 Recognition by Personally Adjusted Discriminant Functions (Experiment I)

In this experiment, average vectors and covariance matrices used for the discriminant functions of vowels and voiced consonants were calculated from the selected words (25 words) for each person. Bayes' discriminant functions prepared by spoken words of a particular person were applied to recognition of the words uttered by that same person.

The confusion matrix for the automatic recognition of vowels, voiced consonants and voiceless consonants are shown in Table 5.5 (a), (b) and (c). Scores in the column labeled "others" represents the errors, which were mainly caused by the segmentation errors.

Averaged scores of vowels is about 95.6%. Most of vowels [a] and [i] were recognized correctly. All the subjects showed a conspicuous mutual confusion between [u] and [o] in common with each other. Especially, comparatively many of [o] sounds of the speaker "NK" were confused with /u/. On the other hand, [u] by the speaker "NR" was apt to be judged to be /o/. Errors of recognition of [e] were found in the words uttered by "UT".

Table 5.5 Recognition results in experiment (I)
5 announcers, 1900 words; voiced sounds are recognized by personally adjusted Bayes' discriminant functions

phoneme in spoken words \ Recognized as	/a/	/i/	/u/	/e/	/o/	/ŋ/	others	the number of samples
/a/	99.5						0.5	823
/i/		99.0	0.3	0.7				650
/u/			91.9	0.4	3.0	2.5	2.2	732
/e/			3.1	94.6			2.3	701
/o/	1.7		5.0		92.7	0.3	0.3	794
/ŋ/			19.4			77.0	3.6	82

(a) Recognition results of vowels (%)

phoneme in spoken words \ Recognized as	/b,d,g/	/r/	/z, dʒ/	/m,n,g̃/	voice-less	others	the number of samples
/b,d,g/	76.1	2.2	3.0	6.0	8.6	4.1	268
/r/	10.5	56.6	6.3	6.3	2.1	18.2	143
/z, dʒ/	5.3	2.6	71.4	1.8	16.7	2.2	227
/m,n,g̃/	2.2	1.0	0.6	68.2		28.1	506

(b) Recognition results of voiced consonants (%)

phoneme in spoken words	/p,t,k/	/s, ʃ,hi/	/c,ki/	/h/	others	the number of samples
/p,t,k/	91.4	1.5	4.9	2.2		294
/s,ʃ,hi/	2.2	85.3	2.2	9.8		225
/c,ki/	5.6		92.0	2.4		115
/h/	4.0			78.8	17.2	99

(c) Recognition results of voiceless consonants (%)

Semi-vowel [j] is segmented into the voiced steady part and is recognized as /i/ or /e/ in most cases when it is put between the back vowels. On the contrary, semi-vowel put between front vowels is, generally, segmented into the transitional part whose duration is rather long.

About 68.1% of the voiced consonants were recognized correctly. Conspicuous examples of confusion were the following.

(1) Voiced plosive and voiced fricative sounds were confused with voiceless sounds.
(2) Liquid-like sounds were recognized as vowels.
(3) Nasals failed to be detected by the segmentation process, or were mis-identified as vowels (especially /u/). The former cases of errors mainly occurred in the utterances of speakers "UT", "NR" and "NK".

The scores of the voiceless consonants at the middle of words were about 86.9%. Confusion was conspicuous between aspirated sounds and fricative sounds.

5.4.2 Recognition by Common Discriminant Functions (Experiment II)

Lumping together the 125 words used in settling the discriminant functions of the experiment (I), we prepared one set of discriminant functions. All the 1,900 words used were recognized by the single set of functions.

Scores obtained from this recognition are shown in Table 5.6. About 94.1% of all the vowels were recognized correctly. Compared with the results of Experiment (I), the error rate of [u] increased by 5%; however, scores of the other vowels did not change so much. The score of the syllabic nasal was about 45.1%.

The recognition rate of all the voiced consonants was about 60.9%. A decline in scores of the voiced consonants was greater, as compared with the vowels.

Table 5.6 Recognition results in experiment (II)
5 announcers, 1900 words; voiced sounds
are recognized by common Bayes'
discriminant functions

phoneme in spoken words \ Recognized as	/a/	/i/	/u/	/e/	/o/	/ɳ/	others	the number of samples
/a/	98.7			0.1	0.4		0.9	823
/i/		97.2	0.8	1.1		0.2	0.8	650
/u/		3.0	86.9	3.7	1.5	1.0	4.0	732
/e/	0.1	1.1	1.6	93.4	0.3	0.1	3.3	701
/o/	0.8		2.9		92.9	0.1	3.3	794
/ɳ/			44.0		1.2	45.1	9.7	82

(a) Recognition results of vowels (%)

phoneme in spoken words \ Recognized as	/b,d,g/	/r/	/z,dʒ/	/m,n,g̃/	voice-less	others	the number of samples
/b,d,g/	70.5	5.6	5.6	5.6	8.6	4.1	268
/r/	15.4	49.7	7.7	4.9	2.1	20.3	143
/z,dʒ/	6.6	6.2	61.7	4.4	16.7	4.4	227
/m,n,g̃/	1.7	2.1		61.8		27.4	519

(b) Recognition results of voiced consonants (%)

5.4.3 Extension to Other Speakers (Experiment III)

Finally, the discriminant functions used in the experiment (II) were tested by 500 words spoken by ten male students, including the vowels and the consonant groups with equal frequency. Words used in this experiment are shown in Table 5.7.

Results of the recognition are shown in Table 5.8. Scores of all the vowels were about 89.5%. The recognition rate of the voiced consonants fell down to about 52.4%. Especially, scores of [u] and [r] fell heavily.

Table 5.7 Word list used in experiment III

b	河　馬	kaba	s	朝	asa
r	狩	kari	p	パピルス	papirusu
z	小　豆	azuki	c	熱　い	atsui
m	雨	ame	h	阿　片	ahen
d	角	kado	ʃ	場　所	basho
r	依　頼	irai	h	位　牌	ihai
dʒ	意　地	iji	p	タイピスト	taipisuto
n	犬	inu	c	い　つ	itsu
b	詭　弁	kiben	s	伊　勢	ise
r	色	iro	h	異　本	ihon
z	無　残	muzan	k	迂　回	ukai
m	海	umi	c	内	uchi
b	う　ぶ	ubu	ʃ	無　宿	mushuku
r	群	mure	h	右　辺	uhen
dʒ	婦　女	hujo	t	疎　い	utoi
n	手習い	tenarai	c	ケチャップ	kechappu
b	え　び	ebi	ʃ	弟　子	deshi
r	け　る	keru	p	エプロン	epuron
z	ぜ　ぜ	zeze	h	手　偏	tehen
m	メ　モ	memo	c	手　帳	techo
d	育　つ	sodatsu	s	長	osa
r	お　り	ori	k	恋	koi
dʒ	野　宿	nojuku	h	祖　父	sohu
n	屋　根	one	k	桶	oke
b	祖　母	sobo	c	誇　張	kocho

On the other hand, about 93.2% of the voiceless consonants in the inside part of the words used were recognized accurately. These scores were preferably better than those obtained from Experiment (I). The fricative sounds and the aspirated sounds were well separated by the values of f_2 parameter.

Table 5.8 Recognition results in experiment (III)
10 students, 500 words; the same discriminant functions are used as in experiment (II)

phoneme in spoken words \ Recognized as	/a/	/i/	/u/	/e/	/o/	others	the number of samples
/a/	92.4		0.5		4.7	2.4	211
/i/		89.3	4.0	1.2		5.6	253
/u/		1.3	80.0	6.1	6.5	6.1	230
/e/	1.0		0.5	97.4		1.0	196
/o/	2.3	0.5	3.7	0.9	88.5	4.2	217

(a) Recognition results of vowels (%)

phoneme in spoken words \ Recognized as	/b,d,g/	/r/	/z, $d_{\overline{3}}$ /	/m,n,$\widetilde{g}$/	voice-less	others	the number of samples
/b,d,g/	56.6	10.5	5.3	5.3	7.9	14.5	76
/r/	15.0	35.0	25.0	8.3	3.3	13.3	60
/z,$d_{\overline{3}}$ /		3.0	63.6	1.5	22.7	9.1	66
/m,n,$\widetilde{g}$/	1.5	7.4	5.9	54.4		30.9	68

(b) Recognition results of voiced consonants (%)

phoneme in spoken words \ Recognized as	/p,t,k/	/s, ʃ,h_i/	/c,k_i/	/h/	others	the number of samples
/p,t,k/	98.8		1.2			83
/s, ʃ,h_i/		96.1		1.3	2.6	77
/c,k_i/	4.1		95.9			73
/h/		1.8		82.1	16.1	56

(c) Recognition results of voiceless consonants (%)

The reasons why the recognition rate of the voiced sounds are worse than that seen in Experiment (II) are the following.

(1) In Experiment (II), the speakers whose words were input for recognition agreed with those who uttered the words for settling the discriminant functions. On the contrary, in Experiment (III), those two groups of speakers were different.

(2) The speakers in the Experiment (III) were non-professional.

5.5 Conclusion

As a result of this recognition scheme, we have rather high scores of recognition of the vowels and the consonant groups in any contexts of spoken words. Three recognition experiments have made clear the dependency of speakers and discriminant functions upon an accurate recognition rate.

However, there remains a problem of recognizing the vowel [u], liquid-like sounds and nasals more accurately. The identification of the semi-vowel [j] and [w] is an open problem.

Recognition errors of the voiced consonants are caused mainly by segmentation errors. Such feedback mechanism will be necessary as can direct the analysis part to try again the segmentation process. If the vocabulary used is limited to some extent, it will be possible to utilize the contexual information and to improve the segmentation accuracy.(21)

There were many common aspects between the results of an audition test of the synthesized speech described in Chapter 4 and results of its recognition. This fact may be considered to suggest that the composite speech processing experiments are a very important means by which a new speech recognition scheme can be developed. The essential information or features for the automatic speech recognition will be made clear by an audition test of the synthetic speech modified by some rules.

118 項欠

Chapter 6 Conclusion

In this thesis, we described the speech synthesis, analysis-synthesis and recognition system by the zero-crossing analysis method.

Pattern processing has come to be one of the most important fields in the recent computer technology. It has been ardently hoped that patterns such as speech, character and picture might be used as the *interactive communication* medium between man and machine.

This thesis was devoted to the description of new speech processing systems, such as speech synthesis by rule in chapter 3, speech analysis-synthesis system in chapter 4 and speech recognition system in chapter 5.

In chapter 2, we discussed the speech analysis methods by zero-crossing wave. Fundamental properties of zero-crossing intervals were examined by using the acoustical speech-synthesis model. Visual representation method of the zero-crossing wave was *studied* and a pitch extraction method was shown, in which the auto-correlation measurement of the sequence of zero-crossing intervals was used. On the other hand, formant extraction was attempted by measurement of the mean zero-crossing interval. The method was ascertained to be useful for vowel classification.

In chapter 3, a new speech synthesis system by using the zero-crossing wave was presented. In this system, we proposed a new information compression method which we call wave-element expression. A wave-element consists of a sequence of zero-crossing intervals in such a pitch period. A speech sound wave is synthesized by a sequence of wave-elements and their attached parameters, that is, the repetition times and the amplitude information of the wave-element. By the synthesizer, we could obtain any spoken sentence in real time, and could apply the system to the output of any other information processing system, such as the information retrieval by computer or the automatic translation from English to Japanese. In the final section of this chapter, we also showed the multi-channel speech output system based on the same speech synthesis method.

In chapter 4, we described a new speech analysis-synthesis system based on

the same idea as was used in the speech synthesis by rule described in chapter 3. The audio signal is fed into a computer through two parallel band-pass filters, and the output signal in each band is subjected to an infinite peak clipping wave. Parameters such as the mean zero-crossing interval and the ratio of amplitudes are extracted in every constant time window. The input speech is segmented into phoneme-like units according to the extracted parameters. Every resultant segment is expressed by the wave-element expression, from which the speech sound is reproduced. Information compression ratio of this expression was 1/10 of PCM coding, and rather intelligible speech could be obtained. This system may be put into practical use as the speech output system with a limited set of vocabulary.

In chapter 5, we showed a speech recognition system which was connected to the speech analysis-synthesis system described in chapter 4. By using results of the segmentation and mean values of parameters, phoneme-like units are recognized as one of the five vowels, syllabic nasal and eight consonant groups. The vowels, syllabic nasal and the voiced consonants are classified by the Bayes' discriminant functions, and the voiceless consonants are classified mainly by the results of segmentation and the parameters of average zero-crossing intervals. Recognition experiments were conducted with 1,900 words spoken by five male announcers and 500 words by ten male students. Rather good recognition accuracy was obtained. The composite speech research system which include the analysis part, the synthesis part and the recognition part will be a very powerful means for the development of a new speech recognition scheme, in which the speech recognition mechanism by machine can be directly contrasted with the speech perception mechanism by man.

ACKNOWLEDGEMENT

The author wishes to express his hearty gratitude to Professor Toshiyuki Sakai for his guidance and constant support given him ungrudgingly during the course of this study.

He also wishes to thank Assistant Professor Shuji Doshita (Tokyo Institute of Technology), Assistant Professor Yasuhisa Niimi (Kyoto Univ. of Industrial Arts and Textile Fibers) for giving him much needed and informative help.

Thanks are also due to Assistant Professors Makoto Nagao, Hidenosuke Nishio, Shigeharu Sugita, Mr. Koichi Tabata, Shinji Tomita and members of Professor Sakai's Research Group for helpful discussion, co-operation and various conveniences offered in time of need.

122 項欠

BIBLIOGRAPHY

(1) Atal B. S. and Hanauer S. L., "Speech analysis and synthesis by linear prediction of the speech wave", JASA 50, p. 637, (1971)

(2) Buron R. H., "Generation of a 1000 word vocabulary for a pulse-excited vocoder operating as an audio response unit", IEEE Transe. AU-16, p.21 (1968)

(3) Chang S. H., et al, "The intervalgram as a visual representation of speech sound", JASA 23, p. 675 (1951)

(4) Chang, et al., "Representations of speech sounds and some of their statistical properties", Proc. IRE 39, p. 147 (1951)

(5) Coker C. H., "Synthesis by rule from articulatory parameters", Paper A-9 Proc. 1967 Conf. Speech Communication and Processing, Cambridge (1967)

(6) Cooper F. S., et al., "The interconversion of audible and visible patterns as a basis for research in the perception of speech", Proc. Nat. Acad. Sci. 37, p. 318 (1951)

(7) Davis K. H., et al., "Automatic recognition of spoken digits", JASA 24, p. 637 (1952)

(8) Denes P., "The design and operation of the mechanical speech recognizer at Univ. College London", Jour. Brit. IRE 19, p. 219 (1959)

(9) Denes P. and Mathew M. V., "Spoken digit recognition using time-frequency pattern matching", JASA 32, p. 1450 (1960)

(10) Dixon N. R. and Maxey H. D., "Terminal analog synthesis of continuous speech using the diphone method of segment assembly", IEEE Trans. AU-16 p.40 (1963)

(11) Doshita S., "Studies on the analysis and recognition of Japanese speech sounds", Doctoral Thesis, Kyoto University (Sept. 1965)

(12) Ewing G. D. and Tailor J. F., "Computer recognition of speech using zero-crossing information", IEEE Trans. AU-17, p. 37 (1969)

(13) Fant G., "Acoustic theory of speech production", Mouton & Co.'s Gravenhage (1960)

(14) Fant G., "Automatic recognition and speech research", STL-QPSR 1(1970)

(15) Flanagan J. L., et al., "Synthetic voices for computers", IEEE Spectrum, p. 22 (October 1970)

(16) Forgie J. W. and Forgie C. D., "Results obtained from a vowel recognition computer program", JASA 31, p. 1480 (1959)

(17) Fry D. B., "Theoretical aspect of mechanical speech recognition", Jour. Brit. IRE 19, p. 211 (1959)

(18) Hashimoto K., et al., "Editing and hearing experiments on speech for telephone directory assistance", Jour. of Acous. Soc. of Japan 20, p. 241 (1964)

(19) Honda T. and Ogawa Y., "Analog synthesizer for speech and singing", Symposium of the Acous. Soc. of Japan, p. 383 (Nov. 1966)

(20) Horita T. and Shiraishi T., "Speech response system", Record of Joint Convention of IECEJ*, 169 (Oct. 1967)

(21) Itahashi S. and Kido K., "Speech recognition - spoken word recognition using dictionary and phonological rule", Jour. of Acous. Soc. of Japan 27, p. 473 (1971)

(22) Itakura F. and Saito S., "Speech information compression based on the maximum likelihood spectral estimation", Jour. of Acous. Soc. of Japan 27, p. 463 (1971)

(23) Kato Y. and Ochiai K., "Terminal analog speech synthesizer", Record of the Joint Convention of the Acous. Soc. of Japan 1-2-14 (Nov. 1966)

(24) Kato Y., et al., "Spoken digit recognizer", IECEJ*47, p. 1319 (1964)

(25) King J. H. and Tunis C. J., "Some experiments in spoken word recognizer", IBM J. 10, p.65 (1966)

(26) Koda M., Hashimoto S. and Saito S., "Spoken digit mechanical recognition system", Jour. of IECEJ* 55-D, p. 186 (1972)

(27) Koenig W., et al., "The sound spectrograph" JASA 18, p. 19 (1946)

(28) Lee L. H. and Mulvany R. B., "Now a talking computer answers inventory inquiries", Electronics, p. 30 (Aug. 16, 1963)

(29) Licklider J. C. R. and Pollack I., "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech", JASA 20, p. 42 (1948)

(30) Matsui E., "A multiple-speech output system by pitch synchronous compilation of vocal units", Record of the Joint Convention of the IECEJ* and others, 2547 (1968)

(31) Matsui E., et al., "Dynamic analog speech synthesizer", Record of the Joint Convention of the Acous. Soc. of Japan 2-2-9 (Nov. 1965)

(32) Matsui E., et al., "A time series analysis based on the Kalman filtering theory", Record of Joint Convention of Acous. Soc. of Japan, 3-2-14 (May 1972)

(33) Mattingly I. G., "Experimental methods for speech synthesis by rule", IEEE Trans. AU-16, p. 198 (1968)

(34) Nakata K. and Miura T., "A method of speech synthesis for multiplexed audio responses", Jour. of IECEJ* 52-C, p. 579 (1969)

(35) Niimi Y., "Studies on programming system for block diagram simulation and its applications", Doctoral Thesis, Kyoto Univ. (Dec. 1967)

(36) Öhman S. E. G., "Coarticulation in VCV utterances: Spectrographic measurements", JASA 39, p. 151 (1966)

(37) Ohtani K., Tomita S. and Sakai T., "Multiple speech output system by zero-crossing wave", Record of the Joint Convention of IECEJ* S.3-10 (1970)

(38) Ohtani K. and Sakai T., "A study on speech synthesis and recognition", Technical Report of the Professional Group on Speech of Acous. Soc. of Japan (Nov. 1971)

(39) Pickett J. M. and Coulter D. C., "Statistics of F2 adjacent to consonants and prediction of F2 onsets", JASA 39, p. 953 (1966)

(40) Rabiner L. R., "Speech synthesis by rule; An acoustic domain approach", Bell Syst. Tech. J. 47, p. 17 (1968)

(41) Reddy D. R., "Computer recognition of connected speech", JASA 42, p. 329 (1967)

(42) Reddy D. R., "Speech recognition: Prospects for the seventies", Proceedings of IFIP Congress 1971, Ljubliana, 1 (1971)

(43) Sakai T. and Doshita S., "The automatic speech recognition system for conversational sound", IEEE Trans. EC-12, p. 835 (1963)

(44) Sakai T. Niimi Y. and Ohtani K., "Speech synthesis by zero-crossing wave", Record of Joint Convention of IECEJ*, 161 (1967)

(45) Sakai T., Niimi Y. and Ohtani K., "Various characteristics of speech viewed from computer processing", Technical Report of the Professional Group on Information Theory and Automata of the IECEJ*(Apr. 1968)

(46) Sakai T., Ohtani K. and Yamagata S., "Speech synthesis of English sentence by computer", Technical Report of the Professional Group on Automaton and Information Theory of the IECEJ*(Sep. 1971)

(47) Sakai T. and Ohtani K., "Consideration of speech synthesis system by editing of wave-elements", Record of National Convention of the Information Processing Society of Japan, p. 11 (1968)

(48) Sakai T., Doshita S., Nagao M. and Ohtani K., "On line speech synthesis through I/0 interface unit", STUDIA PHONOLOGICA V (1970)

(49) Sakai T. and Ohtani K., "A study of speech synthesis system using zero-crossing wave", Jour. of IECEJ* 52-C, p. 704 (1969)

(50) Sakai T., Ohtani K. and Tomita S., "On-line, real-time, multiple-speech output system", Proceedings of IFIP Congress 1971, Ljubliana, TA-4 (1971)

(51) Sakai T., Ohtani K. and Tomita S., "Multiple speech output system by zero-crossing wave", Technical Report of the Professional Group on Automaton and Information Theory of the IECEJ* (June 1970)

(52) Sakai T. and Ohtani K., "Speech analysis-synthesis and recognition system", Record of Joint Convention of the Acous. Soc. of Japan 3-2-10 (May 1972)

(53) Sakoe H. and Chiba S., "Recognition of continuously spoken words based on time-normalization", Jour. of Acous. Soc. of Japan 27, p. 483 (1971)

(54) Scarr R. W. A., "Zero-crossings as a means of obtaining spectral information in speech analysis", IEEE Trans. AU-16, p. 247 (1968)

(55) Sholtz P. N. and Bakis R., "Spoken digit recognition using vowel-consonant segmentation", JASA 34, p. 1 (Jan. 1962)

(56) Stevens K. N., et al., "An electrical analog of the vocal tract", JASA 25, p. 732 (1953)

(57) Takeda T., et al., "Recording-type speech synthesizer", Record of Joint Convention of the IECEJ*, 76 (1966)

(58) Teature C. F., et al., "Experimental, limited vocabulary, speech recognizer", IEEE Trans. AU-15, p. 127 (1967)

(59) Tomita S., Ohtani K. and Sakai T., "On-line, real-time, multiple speech output system", Jour. of IECEJ* 55-D, p. 149 (1972)

IECEJ*: The Institute of Electronics and Communication Engineers of Japan