STUDIA PHONOLOGICA VI (1972)

Speech Analysis-Synthesis and Recognition System

Toshiyuki SAKAI and Kenji OHTANI

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 7 bit PCM coding. Intelligibility of the synthesized speech was ascertained to be rather good. 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 discriminats. Up to 86.9% of the unvoiced consonants could be identified correctly.

Besides, 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.

1 Introduction

In order to check the sufficiency of the 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 consideration has not been taken into acount 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.(1)(2)(3)(4)(5) Anoother one has been studied with the object of recognizing phonemes in any spoken words or sentences.(6)(7)(8)(9)(10) In both of these recognition schemes, the constants or the threshold values which are used in the processes of parameter ex-

Toshiyuki SAKAI (坂井利之): Professor, Department of Information Science, Kyoto University.
Kenji OHTANI (大谷謙治): Assistant, Department of Information Science, Kyoto University.

traction, segmentation and identification are optimized by an inspection or simple calculation. No recognition shemes 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 must 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 paper.(11)

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

(1) vocoder method developed by IBM(12)

(2) PARCOR speech-synthesis method by Itakura(13)

(3) speech analysis-synthesis by linear prediction by B. S. Atal(14)

In these systems, original speech is transformed to compressed information, 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 chapters accompanies the segmentation process in the analysis part and the recognition part aims to identify the phoneme-like units in spoken words.

2 CONSTRUCTION OF THE SYSTEM(15)(16)

A schematic block diagram of the system is shown in Fig. 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

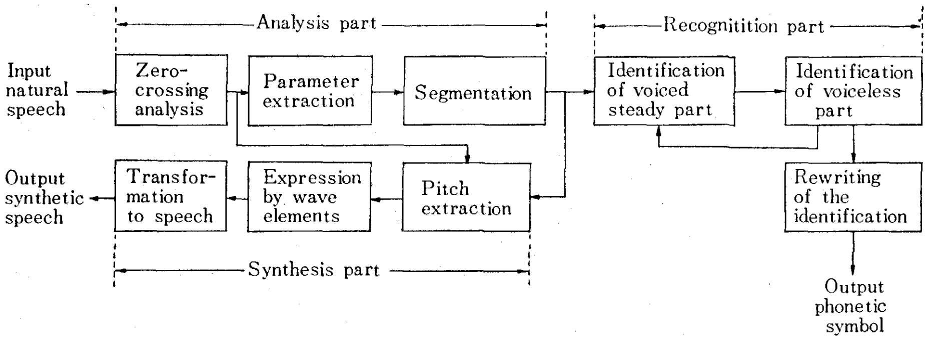

Fig. 1. Construction of the system.

the resultant sequence of the parameters is segmented and classified 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.

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 has a large amount of redundancy. 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-leement). 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.

3 Speech Analysis Part

3.1 Outline of the Speech Analysis Part

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 divided into a fricative (F) and a nonfrictive (C), 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.

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 to pick up the frequency ranges of the first formant (CHI) 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 4800 Hz.

The speech sound wave is, first, pre-emphasized by the RC high-pass filter (6 db/oct., 1.6 kHz 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 210–1100 Hz and 850–4800 Hz respectively. Next, the output waves are multiplexed and converted to a digital code by A/D converter, whose sampling frequency is 10 kHz. Each sample is coded to 11 bits, and is transmitted to the auxiliary storage of a computer.

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

$\bar{\tau}_1$: Average zero-crossing interval of CH1

$\bar{\tau}_2$: Average zero-crossing interval of CH2

Parameter "R" is the amplitude ratio of CH1 to CH2. From $\bar{\tau}_1$ and $\bar{\tau}_2$, parameters f_1 and f_2 in a frequency domain are obtained by

$$f_1 = 1/(2\bar{\tau}_1)$$
$$f_2 = 1/(2\bar{\tau}_2)$$

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, ${}_if_1$, ${}_if_2$.

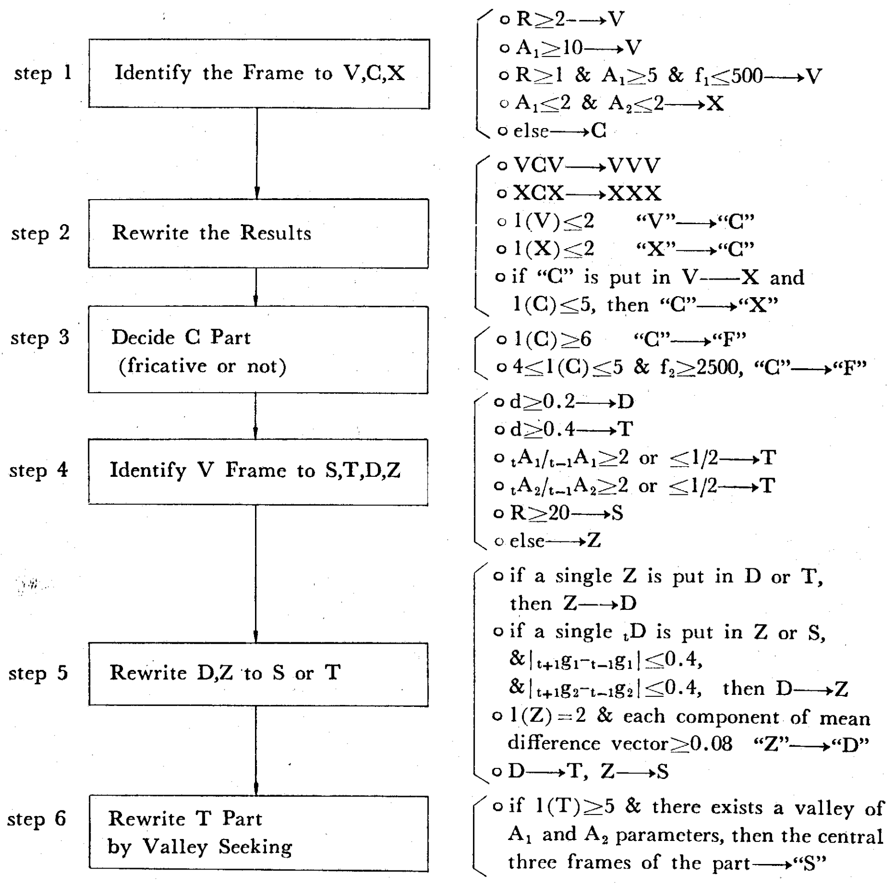
step 1
Identify the Frame to V,C,X
o $R \geq 2 \longrightarrow V$
o $A_1 \geq 10 \longrightarrow V$
o $R \geq 1$ & $A_1 \geq 5$ & $f_1 \leq 500 \longrightarrow V$
o $A_1 \leq 2$ & $A_2 \leq 2 \longrightarrow X$
o else ⟶ C
step 2
Rewrite the Results
o VCV ⟶ VVV
o XCX ⟶ XXX
o $l(V) \leq 2$ "V" ⟶ "C"
o $l(X) \leq 2$ "X" ⟶ "C"
o if "C" is put in V——X and $l(C) \leq 5$, then "C" ⟶ "X"
step 3
Decide C Part (fricative or not)
o $l(C) \geq 6$ "C" ⟶ "F"
o $4 \leq l(C) \leq 5$ & $f_2 \geq 2500$, "C" ⟶ "F"
step 4
Identify V Frame to S,T,D,Z
o $d \geq 0.2 \longrightarrow D$
o $d \geq 0.4 \longrightarrow T$
o ${}_tA_1/{}_{t-1}A_1 \geq 2$ or $\leq 1/2 \longrightarrow T$
o ${}_tA_2/{}_{t-1}A_2 \geq 2$ or $\leq 1/2 \longrightarrow T$
o $R \geq 20 \longrightarrow S$
o else ⟶ Z
step 5
Rewrite D,Z to S or T
o if a single Z is put in D or T, then Z ⟶ D
o if a single ${}_tD$ is put in Z or S, & $|{}_{t+1}g_1 - {}_{t-1}g_1| \leq 0.4$, & $|{}_{t+1}g_2 - {}_{t-1}g_2| \leq 0.4$, then D ⟶ Z
o $l(Z) = 2$ & each component of mean difference vector ≥ 0.08 "Z" ⟶ "D"
o D ⟶ T, Z ⟶ S
step 6
Rewrite T Part by Valley Seeking
o if $l(T) \geq 5$ & there exists a valley of A_1 and A_2 parameters, then the central three frames of the part ⟶ "S"

Fig. 2. Flowchart of the segmentation process

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

3.4 Segmentation

Process of segmentation is shown in Fig. 2. 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, ($R \geqq 2 \rightarrow V$) 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 results of the decision at the step 1 is rewritten according to its duration "l" and the results of the neighboring frames; for example, VCV→VVV means that C frame put between V part 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.

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 and g_2 are logarism of f_1 and f_2 normalized by their respective variances.

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 of ${}_tA_j$ ($j=1, 2$) 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.

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 abbrebiated 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.

3.6 Results and their Discussion

Results of the segmentation of consonants are listed in Table 1 and 2. The Table 1 shows the results of consonants in the top position of words, while the Table 2 shows the results of consonants in the inside position of words. In Table 1, the score contained in "ø" row of the first column means that vowels were segmented as sounds in the form of voiceless consonant+vowel, and the scores contained in "ø" column except in the first row mean that the consonants failed to be detected. The column "silence+voiceless" in Table 2 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 1. 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, dʒ	5.1	12.8*	82.1*		
h	46.2*	23.1*	18.1		12.7

Table 2. 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
g̃					66.4*	12.0		21.6
z, dʒ	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 [g̃] 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 [dʒ], the voiceless parts were detected.

4 Speech Synthesis Part

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.

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 two 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.

4.2 Pitch Extraction

In the voiced part, pitch intervals are extracted so as to be used for the wave-element expression.

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~1.3T 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^1 with OXI^J.

4.3 Wave-Element Expression

In Table 3, the wave-element expression of the segment is listed. 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 $\boldsymbol{a}$ means a set of amplitude parameters of CH1 and CH2.

The silence part (X) is expressed by repetitions of a wave-leement 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 3.

Table 3. Wave-element expression

part	wave-element expression
C	$\boldsymbol{a}_1 \mathrm{P}_{c1}{}^1 \cdot \boldsymbol{a}_2 \mathrm{P}_{c2}{}^1 \cdots\cdots \boldsymbol{a}_r \mathrm{P}_{cr}{}^1$
F	$(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots\cdots, \boldsymbol{a}_r)\ \mathrm{P}_f{}^{r*}$
X	$(\overbrace{\mathrm{O}, \mathrm{O}, \cdots\cdots, \mathrm{O}}^{r})\mathrm{P}_x{}^r$
S	$(\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots\cdots, \boldsymbol{a}_r)\mathrm{P}_s{}^r$
T	$\boldsymbol{a}_1 \mathrm{P}_{t1}{}^1 \cdot \boldsymbol{a}_2 \mathrm{P}_{t2}{}^2 \cdots\cdots \boldsymbol{a}_r \mathrm{P}_{tr}{}^1$

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 this difference that the wave-element in the part corresponds to a pitch interval.

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 two sequences of digital samples are converted to analogue value by two sets of D/A converters, and filtered by the low-pass filter (900 Hz cut-off, 12 db/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, 12 db/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 (6 db/oct., 1.6 kHz cut off).

Sonagrams of the input natural speech and its synthetic speech are presented in Fig. 3 [ijaku] (胃弱).

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 segmenta-

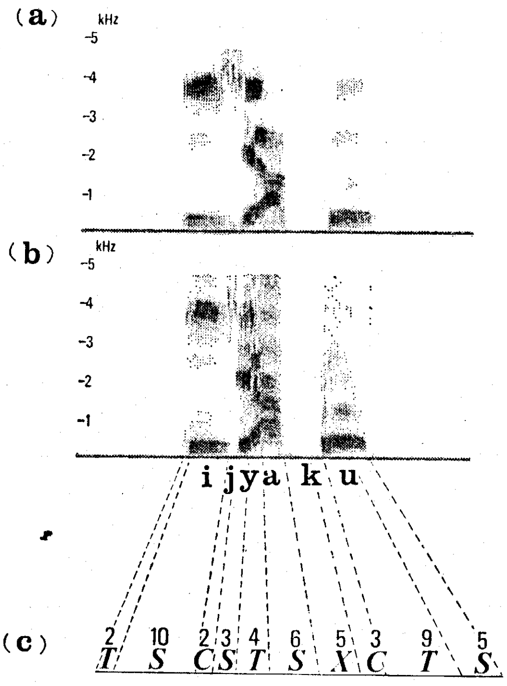

Fig. 3. Synthesized speech [ijaku] (胃弱)
(a) sonagram of natural input speech
(b) sonagram of synthesized speech
(c) result of segmentation

tion process are also presented. The number attached to the identified symbols shows the number of frames in the segments.

4.5 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 bits × 10,000 samples), it corresponds to about 1/10 information-compression ratio.

5 Speech Recognition Part

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 Section 3.

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; for example, 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/, /g̃/, (2) voiced plosive /b/, /d/, /g/
(3) liquid-like /r/ (4) voiced fricative /z/, /dʒ/
(5) voiceless plosive /p/, /t/, /k/
(6) africative /c/
(7) voiceless fricative /s/, /ʃ/, /h_1/
(8) aspirated /h/

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

5.1 Outline of the Recognition Part

The flowchart of the recognition system is shown in Fig. 4. 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 ampli-

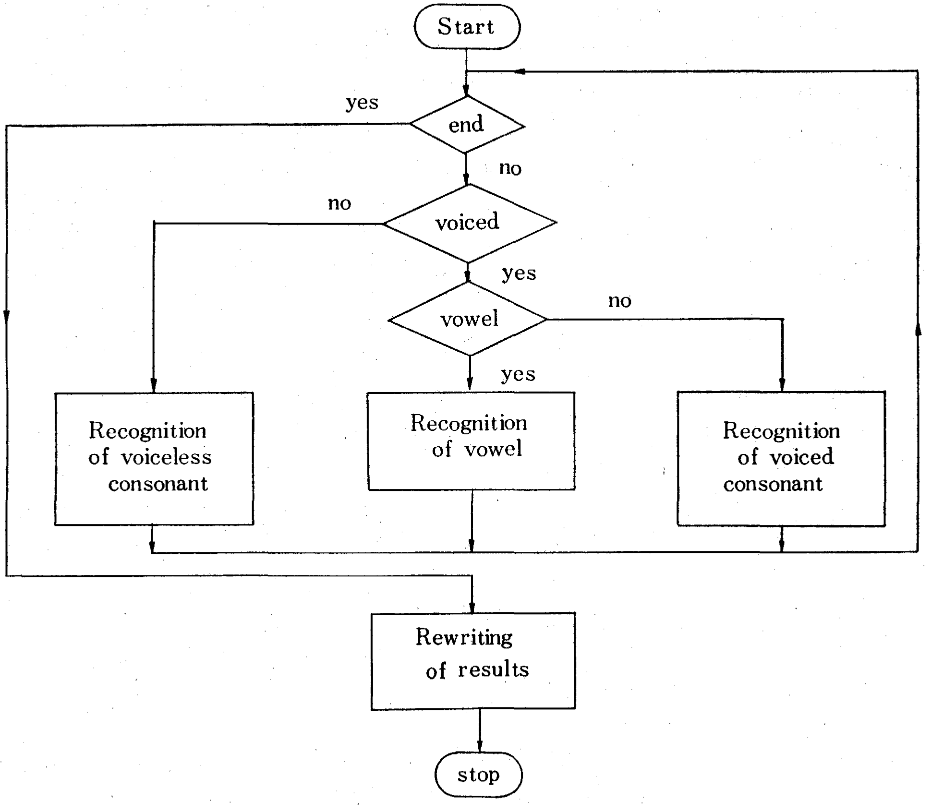

Fig. 4. Flowchart of the recognition system.

tude) 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 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 chennels 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.

5.2 Discrimination of Vowels and Voiced Consonants

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 gained by averaging the abutting two samples of CH2).

Let $\boldsymbol{x}$ be the vector expression in log scale of the above-mentioned four parameters which are averaged in a voiced steady part.

Bayes' discriminant functions are given by

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

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 $\boldsymbol{x}$ is normal and that the frequency of appearance of each category is even, the function h_i of $\boldsymbol{x}$ is the discripminant function of the optimum classifier. The input pattern $\boldsymbol{x}$ 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.

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

$$\alpha = \max\{{}_{j+1}A_1,\ {}_{j+2}A_1,\ {}_{j+3}A_1\}/{}_jA_1$$

$$\beta = \max\{{}_{j-1}A_1,\ {}_{j+1}A_1\}/{}_jA_1$$

Here, the "j"th frame has the minimum A_1 parameter in the voiced steady part. 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].

5.3 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. 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).

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

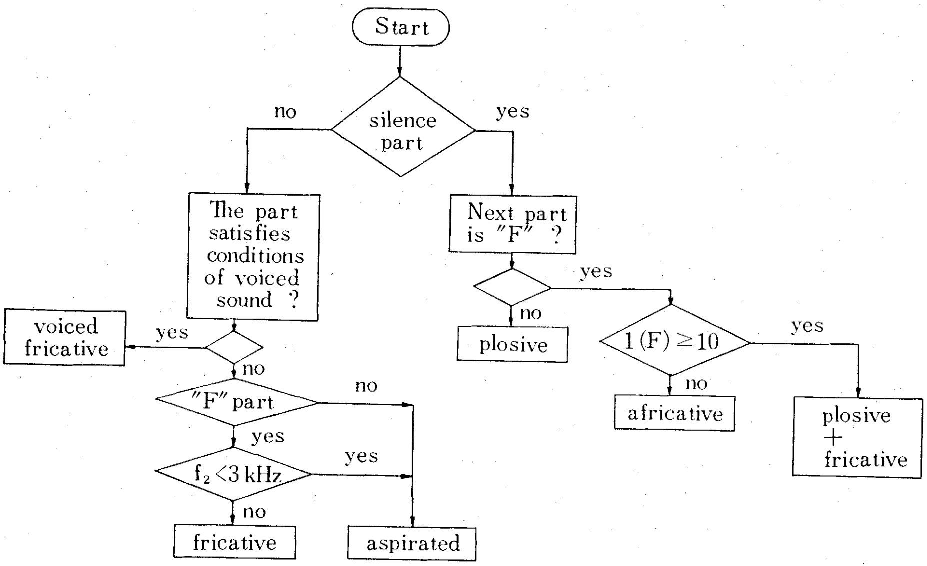

Fig. 5. Recognition tree of voiceless consonants.

sounds depends on whether the average f_2 parameter in the part is greater than the threshold (3 kHz) 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 2), 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.

5.4 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_1.

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.

5.5 Recognition Experiments of input Speech

The following three recognition experiments were conducted with the 1900 words spoken by five male announcers shown in section 3.5 and with 500 words by ten male students.

5.5.1 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 4 (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

Table 4. Recognition results in experiment (I)
5 announcers, 1900 words ; voiced sounds are recognized
by personally adjusted Bayes' discriminant functions

Recognized as

phoneme in spoken words	/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 (%)

Recognized as

phoneme in spoken words	/b, d, g/	/r/	/z, dʒ/	/m, n, g̃/	voiceless	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, h_i/	/c, k_i/	/h/	others	the number of samples
/p, t, k/	91.4	1.5	4.9	2.2		294
/s, ʃ, h_i/	2.2	85.3	2.2	9.8		225
/c, k_i/	5.6		92.0	2.4		115
/h/	4.0			78.8	17.2	99

(c) Recognition results of voiceless consonants (%)

many of [o] sounds of the speaker "NK" were confused with /u/. On the other hand, [u] by the speaker "NK" was apt to be judged to be /o/. Errors of recognition of [e] were found in the words uttered by "UT".

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 misidentified 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.

5.52. 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.

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.

5.5.3 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.

Results of the recognition are shown in Table 5. 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.

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.

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

Table 5. Recognition results in experiment (III)
10 students, 500 words ; the same discriminant functions are used as in experiment (II)

Recognized as

phoneme in spoken words	/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 (%)

Recognized as

phoneme in spoken words	/b, d, g/	/r/	/z, dʒ/	/m, n, g̃/	voiceless	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, dʒ/		3.0	63.6	1.5	22.7	9.1	66
/m, n, g̃/	1.5	7.4	5.9	54.4		30.9	68

(b) Recognition results of voiced consonants (%)

Recognized as

pooneme in spoken words	/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 (%)

(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.

6 Conclusion

The composite speech processing system was described, and also results of the segmentation, speech synthesis and speech recognition were shown.

The fundamental idea of the analysis-synthesis method consisted in a description of speech by the wave-element expression, that is, by elements (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) differing from the speech synthesis by rule,[11] sequences of the zero-crossing intervals cut out from the input natural speech are used as the wave-elements for that word.

(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.

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.[8]

BIBLIOGRAPHY

1. King J.H. and Tunis C.J., "Some experiments in spoken word recognizer", IBM J. 10, p. 65 (1966)
2. Koda M., Hashimoto S. and Saito S., "Spoken digit mechanical recognition system", Jour. of IECEJ* 55-D, p. 186 (1972)
3. Sakoe H. and Chiba S., "Recognition of continuously spoken words based on time-normalization", Jour. of Acous. Soc. of Japan 27, p. 483 (1971)
4. Sholtz P.N. and Bakis R., "Spoken digit recognition using vowel-consonant segmentation", JASA 34, p. 1 (Jan. 1962)
5. Teature C.F., et al., "Experimental, limited vocabulary, speech recognizer", IEEE Trans. AU-15, p. 127 (1967)
6. Denes P., "The design and operation of the mechanical speech recognizer at Univ. College London", Jour. Brit. IRE 19, p. 219 (1959)
7. Forgie J.W. and Forgie C.D., "Results obtained from a vowel recognition computer program", JASA 31, p. 1480 (1959)
8. 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)
9. Reddy D.R., "Computer recognition of connected speech", JASA 42, p. 329 (1967)

10. Sakai T. and Doshita S., "The automatic speech recognition system for conversational sound", IEEE Trans. EC-12, p. 835 (1963)
11. Sakai T., Doshita S., Nagao M. and Ohtani K., "On line speech synthesis through I/0 interface unit", STUDIA PHONOLOGICA V (1970)
12. 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)
13. 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)
14. Atal B.S. and Hanauer S.L., "Speech analysis and synthesis by linear prediction of the speech wave", JASA 50, p. 637, (1971)
15. 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)
16. 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)

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

(Aug. 31, 1972, received)