STUDIA PHONOLOGICA XVII (1983)

Phoneme Recognition for Speaker-Independent Connected Word Recognition

Kiyoshi MAENOBU, Yasuo ARIKI and Toshiyuki SAKAI

SUMMARY

In this paper, a method of phoneme recognition is described for speaker-independent connected word recognition. The phoneme recognition is carried out by two processes, namely, segmentation and phonemic labeling.

In the segmentation, insensitivity is desirable to speaking rate or structural difference of speech production organs. To achieve such insensitivity, we propose an optimal segmentation technique by the variance-minimization which globally divides the input speech pattern into segments by minimizing the sum of the variance of each segment, in stead of local and sequential division.

In the phoneme labeling, segment labels are determined on the basis of the frame labels which are determined by template matching between phonemic reference patterns and the frames. It is devised that the phonemic reference patterns to be matched with the frames are reduced using the inter-segment and intra-segment information.

1. INTRODUCTION

In researches of speaker dependent connected word recognition, 99.6% recognition accuracy is achieved by the two-level DP-matching technique which carries out the word matching and word sequence decision simultaniously.[1] The reason for the high recognition accuracy is attributed to the following three points.

(1) Words are employed as the recognition unit so that the coarticulation effect can be more avoided in comparison with phoneme or syllable-unit-based approach.

(2) The amount of information contained in words is greater than that of phoneme or syllables so that the recognition accuracy increases compared with phoneme or syllable-unit-based approach.

(3) Segmentation and recognition are carried out simultaniously by two-level DP-matching algorithm (segmentation free) so that the recognition error decreases which arises from preliminary segmentation.

Speaker dependent connected word recognition exactly shows such high recognition

Kiyoshi MAENOBU (前信潔) : Yasuo ARIKI (有木康雄) : Assistant Professor, Department of Information Science. Kyoto University.
Toshiyuki SAKAI (坂井利之) : Professor, Department of Information Science, Kyoto University.

accuracy, however, it requires the registration of word–speech pattern (word–reference pattern) for each word by each speaker in advance. Therefore large vocabulary makes it almost impossible to perform the registration as well as real time recognition due to the large amount of processing data (the number of word–level DP–matching in two–level DP–matching is the product of the number of registered word–reference patterns and the length of the input speech pattern). To enable registration–free connected word recognition in real time, speaker independent approach is required based on compressed speech data.

For the speaker independent approach, the following techniques have been proposed to normalize the variation by speakers.

(1) Multiple word–reference patterns are prepared for each word based on statistical method.

(2) Speaker–adaptation is employed based on learning.

(3) Normalization algorithm is employed to absorb the difference in excitation and vocal tract.

(4) Speaker–independent phoneme or syllable reference patterns are used insead of speaker–dependent word–reference patterns to absorb the variations by speakers. [2]

We employ the method (4), or phoneme–unit–based approach for speaker–independense, because the method has the effect of speech data compression which enables the high speed processing and decreases the required memories. On the phoneme string obtained from the input speech pattern, connected word recognition my be carried out by the two–level DP–matching algorithm for speaker–independent recognition.

Phoneme string is produced by phoneme recognition through segmentation and phonemic labeling which divide the input speech pattern into segments and then assign the phonemic symbols to them. In the segmentation, insensitivity is desirable to speaking rate or structural difference of speech production organs. To achieve such insensitivity, we propose an optimal segmentation technique by the variance–minimization which globally divides the input speech pattern into segments by minimizing the sum of the variance of each segment, instead of local and sequential division. In this optimal segmentation technique, Dynamic Programming is used to seek the optimum number of segments and their boundaries. [3] A threshold value used in the segmentation is also designed to be insensitive to the differences by speakers.

2. Overview of the Phoneme Recognition

2.1 System Organization

Fig. 2–1 shows the block diagram of our phoneme recognition system for speaker–independent connected word recognition. At present the task is limited to the recognition of the connectedly spoken numerals by any speakers up to four digits. As shown in Fig. 2–1, the system mainly consists of two blocks. They are acoustic analysis and phoneme recognition.

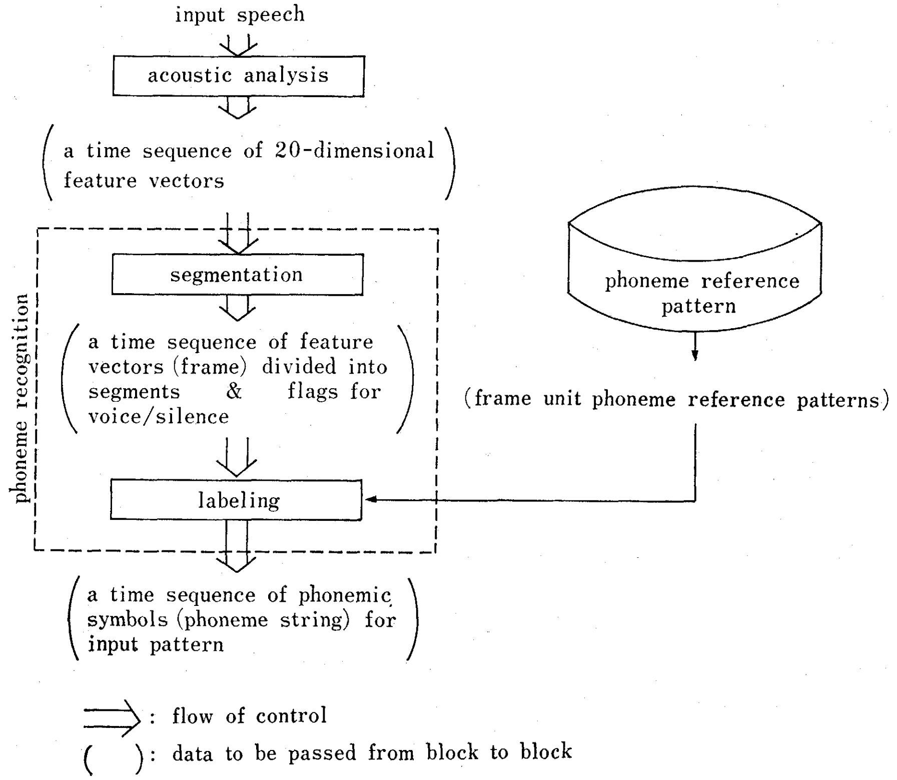

input speech
acoustic analysis
a time sequence of 20-dimensional feature vectors
segmentation
a time sequence of feature vectors (frame) divided into segments & flags for voice/silence
labeling
phoneme recognition
phoneme reference pattern
(frame unit phoneme reference patterns)
a time sequence of phonemic symbols (phoneme string) for input pattern
: flow of control
( ): data to be passed from block to block

Fig. 2–1. Block diagram of Phoneme recognition.

2.2 Outline of the Processing

2.2.1 Acoustic Analysis

A speech signal is first passed into a pre–emphasis circuit with a slope of 6–dB per octave below 1600 Hz for improving the signal–to–noise ratio at high frequencies, and then fed into the 20–channel filter–bank. After they are full–wave–rectified and smoothed by the low–pass filter (cut–off frequency: 40 Hz), the output waves are sampled at every 10 ms interval (frame interval) and digitized with an accuracy of 10 bits. The center frequencies of the 20 channels increase in order

Table 2–1. Center frequency (Hz) of bandpass filter.

channel No.	center freq.	channel No.	center freq.	channel No.	center freq.	channel No.	center freq.
1	210	6	500	11	1,190	16	2,830
2	250	7	595	12	1,410	17	3,360
3	297	8	707	13	1,680	18	4,000
4	354	9	841	14	2,000	19	4,760
5	420	10	1,000	15	2,380	20	5,660

by a factor $2^{1/4}$. These frequencies are shown in Table 2–1. As the result of this acoustic analysis, a time sequence of 20–dimensional feature vectors are obtained from the input speech.

In the following segmentation, we calculate from these feature vectors the frame–power which corresponds to speech energy. The frame–power is defined as the norm of feature vectors at each frame and expressed as follows;

$$FP_i=\sqrt{\sum_{q=1}^{20} X_{iq}^2} \qquad (2.1)$$

where X_{iq} is the element of the feature vector X_i at the i–th frame.

2.2.2 Phoneme Recognition

The phoneme recognition process is further divided into two sub–blocks as shown in Fig. 2–1. They are the segmentation block and labeling block. In the segmentation block, an input speech pattern is at first divided into voice segments and silence segments. The voice segments are further divided into smaller segments corresponding to phonemes using the optimal segmentation technique by the variance–minimization. Accordingly, the data to be passed from the segmentation block to the following labeling block is a time sequence of feature vectors divided into segments and flags to identify the voice or silence segments.

In the labeling block, the phonemic symbols are assigned to each segment using phoneme–reference patterns each of which is composed of one frame. Up to this stage the input speech pattern is converted to a time sequence of phonemic symbols (phoneme string).

The details of these blocks are to be described in the following section.

3. OPTIMAL SEGMENTATION TECHNIQUE BY THE VARIANCE–MINIMIZATION

3.1 Sequential Segmentation

The segmentation is the process to divide the input speech pattern, which is continuously changing under the physical constraints of the articulation organs, into the segments corresponding to phonemes which are discrete symbols. The problems inherent to the segmentation are caused by this conversion from continuity to discreteness.

So far, two approaches are proposed for the segmentation. One is the sequential segmentation by detecting the local changes, maximum or minimum. The other is the optimal segmentation performed by solving the minimization problem. In the sequential segmentation, the problems caused by the conversion from continuity to discreteness are clarified as follows.

(1) The detection errors of the segment boundaries are propagated to the following process.

(2) The large amount of information and processing time are required for the detection of the ambiguous segment boundaries.

(3) The threshold value to detect the local changes are sensitive to speakers, speech rate and the input speech evel.

To solve these problems, especially (3) for speaker-independence, we incorporate the optimal segmentation.

3.2 General Principle of the Optimal Segmentation

At segment boundaries the voice is continuously changing so that it is difficult to decide the segment boundaries definitely. Accordingly, in the optimal segmentation, the uniformity of the features in each segment is intensified in stead of detecting the changes. As the mesure of the uniformity, we employ the variance of the feature vectors.

The number of segments and segment boundaries of the input speech pattern are calculated by minimizing the total sum of these variances. This optimal segmentation technique by the variance-minimization has the following advantages.

(1) Error propagation does not occur.

(2) There is no threshold so that speaker-independence is achievable.

(3) Segmentation is performed by solving the minimization problem so that the process is simplified.

(4) The high ability of segmentation is achieved by small amount of information.

3.3 Formalization of the Optimal Segmentation

Let A denote the input speech pattern as follows.

$$A = a_1 a_2 \cdots\cdots a_i \cdots\cdots a_N \tag{3.1}$$

where a_i is a m-dimensional feature vector like $a_i = (a_{i1}, a_{i2}, \cdots\cdots, a_{im})$

Our purpose is to divide the input speech pattern into the $K(1 \leq K \leq N)$ segments. Let j_k be the frame number located on the boundary between the k-th segment and the k+1-th segment as depicted in Fig. 3-1. The variance v_k of the feature vector within the k-th segment is given as the following expression.

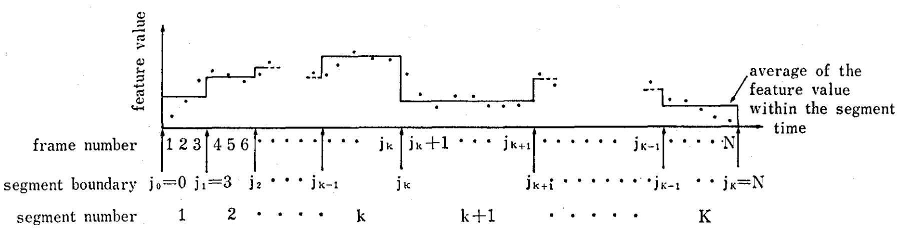


Fig. 3-1. Conception of the optimal segmentation technique by the variance-minimization. (one-dimensional feature vector)

$$v_k = \sum_{p=1}^{m} \sum_{i=j_{k-1}+1}^{j_k} (a_{ip} - \bar{a}_p^k)^2 / (j_k - j_{k-1}) \quad (k=1 \sim K) \tag{3.2}$$

where $\bar{a}_p^k$ is the average of the p-th element of the feature vector within the k-th segment as follows.

$$\bar{a}_p^k = \sum_{i=j_{k-1}+1}^{j_k} a_{ip}/(j_k - j_{k-1}) \quad (p=1\sim m,\ k=1\sim K) \tag{3.3}$$

The number of the segments K and the segment boundaries $\{j_k\}$ are computed by minimizing the total sum V_K of the variances of all segments all over the input speech as follows.

$$\min_{K,\{j_k\}} V_k = \min_K[\min_{\{j_k\}}\{\sum_{k=1}^{K} v_k\}]$$

$$= \min_K[\min_{\{j_k\}}\{\sum_{k=1}^{K}\sum_{p=1}^{m}\sum_{i=j_{k-1}+1}^{j_k}(a_{ip}-\bar{a}_p^k)^2/(j_k-j_{k-1})\}] \tag{3.4}$$

The initial condition is $j_0=0$, $j_K=N$. The expression (3.4) is the minimization problem of the weighted summation so that it can be solved by DP (Dynamic Programming). The DP-equation is as follows.

$$T(m) = \min_{0<l<m}\{T(l) + v(l+1,\ m)\} \quad 1 \leq m \leq N \tag{3.5}$$

where T(m) is the minimum partial sum of the variances up to the m-th frame with respect to l. The v(l+1, m) is the variance within the segment starting at the l+1-th frame and ending at the m-th frame.

To reduce the computational time, we introduce the reasonable constraints that the segment length $(j_k - j_{k-1})$ is bound as follows.

$$2 \leq \theta_{min} \leq j_k - j_{k-1} \leq \theta_{max} \leq N \tag{3.6}$$

The threshold to these parameters θ_{min} and θ_{max} do not depend on speakers as shown by the later experiment. The segmentation results are passed to the labeling block with the information about the length, starting and ending frame number and the flag to indicate the voice/silence of each segment.

3.4 Experimental Result of the Optimum Segmentation

We have evaluated the abilites of the optimal segmentation in the following points.

(1) The feature parameter sensitivity.

(2) Speaker and the speech rate dependency.

(3) The segmentation error rate with respect to each phoneme.

3.4.1 Feature Parameter Sensitivities

The following nine feature parameters are compared each other to obtain the most effective feature parameters for the optimal segmentation.

(a) Frame power

$$P_{1i} = FP_i = \sqrt{\sum_{q=1}^{20} X_{iq}^2} \quad \text{(one-dimension)}$$

where X_{iq} is the q-th element of the feature vector X_i at the i-th frame. Here X_i is the output level from the 20-channel filter-bank.

(b) Euclid distance between consecutive two frames.

$$P_{2i} = \sqrt{\sum_{q=1}^{20}(X_{iq} - X_{i-1q})^2} \quad \text{(one-dimension)}$$

(c) Power ratio of the lower three channels to the frame power

$$P_{3i}=\sqrt{\sum_{q=1}^{3} X_{iq}^2/FP_i} \quad \text{(one-dimension)}$$

(d) Power ratio of the middle 14 channels to the frame power

$$P_{4i}=\sqrt{\sum_{q=4}^{17} X_{iq}^2/FP_i} \quad \text{(one-dimension)}$$

(e) Power ratio of the higher three channels to the frame power

$$P_{5i}=\sqrt{\sum_{q=18}^{20} X_{iq}^2/FP_i} \quad \text{(one-dimension)}$$

(f) Minimum distance between each of the five vowels (/a, /i/, /u/, /e/, /o/) and the frame.

$$P_{6i}= \min_{v=\{/a/,/i/,/u/,/e/,/o/\}} \sqrt{\sum_{q=1}^{20}(X_{iq}-r_{vq})^2} \quad \text{(one-dimension)}$$

where, $r_v=(r_{v1}, r_{v2}, \cdots\cdots, r_{v20})$ is the reference pattern of the five vowels.

(g) Feature vector

$$P_{7i}=(P_{3i}/P_{4i},\ P_{5i}/P_{4i}) \quad \text{(two-dimension)}$$

(h) Feature vector

$$P_{8i}=(P_{3i},\ P_{4i},\ P_{5i}) \quad \text{(three-dimension)}$$

(i) Feature vector from the 20-channel filter-bank

$$P_{9i}=X_i=(X_{i1},X_{i2},\ \cdots\cdots,\ X_{i20}) \quad \text{(20-dimension)}$$

The result is shown in Table 3-1. The experimental conditions are as follows.

Table 3-1. Experimental result of the segmentation by the different nine feature parameters. (unit %)

feature parameter	P_{1i}	P_{2i}	P_{3i}	P_{4i}	P_{5i}	P_{6i}	P_{7i}	P_{8i}	P_{9i}
correct	63.9	42.2	56.5	51.1	46.0	53.2	42.6	49.4	55.3
split	30.9	25.3	13.9	16.5	19.0	24.1	14.8	18.6	36.7
merge	2.1	16.0	14.3	16.0	17.3	11.0	20.7	16.0	3.4
omission	3.0	16.5	15.2	16.5	17.7	11.8	21.9	16.0	4.6
dimension	1	1	1	1	1	1	2	3	20

threshold : $\theta_{\min}=3$ (30 msec), $\theta_{\max}=15$ (150 msec)
input speech : 10 sentences containing 233 phonemes
speaker : one male
speech rate : 8.1 phonemes per second on an average.

In the Table 3-1, "split" indicates that the feature vector sequence corresponding to one phoneme is further divided into small segments. "Merge" and "omission" are the phenomena in the case where the feature vector sequence corresponding to several phonemes is regarded as one segment; the longest sequence of the feature vectors corresponding to one phoneme contained in the segment is called the merged segment and the others are called the omitted segments. It should be noted that the input speech pattern is controlled to be excessively divided into segments due to the following two reasons by setting the threshold for $\theta_{\min}$ and

θ_{max} to be the small value.

(1) "Merge" and "omission" cause the fatal errors in the connected word recognition.

(2) "Split" can be absorbed in the connected word recognition.

From the experimental results, it is clear that the feature parameter P_{1i} (frame power) and P_{9i} (output from the 20-channel filter-bank) are effective in the sense that their rate of "merge" and "omission" is lower than others. Taking into account that the amount of computation is proportional to the dimension of the feature parameter (vector), we employ the frame power P_{1i} as the feature parameter for the optimal segmentation.

It is interesting to consider why the feature parameters P_{1i} and P_{9i} show the high ability in the segmentation. In Japanese language, a vowel follows a consonant so that the feature parameters whose values change relatively large between vowels and consonants show the high ability in the segmentation.

3. 4. 2 Speaker and Speech Rate Dependency

The experimental conditions are as follows.

threshold : $\theta_{min}=3$, $\theta_{max}=15$
feature parameter : frame power
input speech : 10 sentences containing 198 phonemes
speaker : five males. Two males speak at the different speech rate.

The experimental result is shown in Table 3-2. The result shows that the error rate ("merge", "omission") depends on the speech rate but not on the speakers. This dependency on speech rate resutls from the effect of the coarticulation in high-speed input rate.

Table 3-2. Experimental result of the segmentation by the different speakers and speech rate. (unit %)

speaker	A-1	A-2	B-1	B-2	C	D	E
correct	68.7	68.7	63.1	60.6	63.9	70.7	70.7
split	16.2	22.2	23.4	21.7	30.9	22.2	14.6
merge	6.1	4.0	6.3	8.6	2.1	3.0	4.5
omission	9.1	5.1	7.2	9.1	3.0	4.0	10.1
phoneme/second	16.5	12.4	13.4	14.9	8.9	12.5	13.7

3. 4. 3 Segmentation Error Rate with Respect to Each Phoneme

The experiment has been conducted under the same conditions of the experiment for speaker and speech rate dependency as described in (3. 4. 2). The experimental result is shown in Table 3-3. From the Table, the followings can be concluded.

(1) Vowels and unvoiced fricatives /s/, /c/ tend to be split due to the small value of θ_{max} for their long stability.

Table 3-3. Experimental result of the segmentation by different speakers with respect to each phoneme. (unit %)

phoneme	/a/	/i/	/u/	/e/	/o/	/k/	/t/	/p/	/s/	/c/	/h/
correct	40.3	40.9	54.5	60.2	64.0	97.7	93.6	100.0	86.1	40.0	97.7
split	45.2	48.3	27.3	33.3	29.7				12.5	60.0	
merge	14.4	9.4	13.6	4.9	5.7	0.8	2.6		1.4		2.3
omission		1.3	4.5	1.6	0.6	1.6	3.8				

phoneme	/n/	/m/	/r/	/j/	/w/	/g/	/d/	/b/	/z/	/N/	total
correct	75.6	78.8	68.1	36.0	36.8	100.0	93.6	94.6	66.7	60.0	66.0
split	4.9	9.1		44.0					16.7	22.0	22.4
merge	7.3	9.1	4.3	4.0					16.7	12.0	6.5
omission	12.2	3.0	27.5	16.0	63.2		6.4	5.4		6.0	5.0

(2) Voiced sound except for plosives tend to be omitted due to the small change between vowels and voiced sound.

As the experimental results, using the value of the power at each frame as a feature parameter, the error rate of the "omission" and "merge" in the segmentation was 5% and 6.5% respectively.

4. Speaker-Independent Phonemic Labeling

4.1 Phonemic Symbols and Phonemic Labels

At present, the task for the system is the connectedly spoken Japanese numerals. Table 4-1 shows the representation of Japanese numerals by phonemic symbols. The number of the required phonemic symbols is 14 as shown in Table 4-1, however, we have used 18 phonemic labels, which were expanded from phonemic symbols, as shown in Table 4-2. The additional four phonemic labels are silence label and the three labels derived from /i/, /k/ and /j/. I2 is the nazalized /i/ in the word /ni/. K6 and K9 are the plosive /k/ in /roku/ and the fricative /k/ in /kju/ respectively. Y4 and Y9 are phonetically different in Japanese.

4.2 Labeling Technique

4.2.1 Classification of Labeling Technique

There are following three techniques to assign the phonemic labels to each segment.

(1) Template matching based on the similarity between the segment and phonemic reference patterns.

(2) Statistical decision like a discriminant function which discriminates the phonemic class using several feature parameters.

(3) Logical decision like a decision tree which classifies the unknown segment into the certain phonemic class by using the feature parameters sequentially.

The problem of the pattern matching (1) is time consumption because the amount of computation is proportional to the number of reference patterns. In order to

Table 4-1. Representation of Japanese numerals by phonemic symbols.

numeral	phonemic symbol
1	/ici/
2	/ni/
3	/SaN/
4	/joN/
5	/go/
6	/roku/
7	/nana/
8	/haci/
9	/kju/
0	/re(i)/

Table 4-2. phonemic symbols and phonemic labels used in the system.

phonemic symbol	phonemic Label
/a/	A
/i/	I1, I2*1
/u/	U
/e/	E
/o/	O
/k/	K6, K9*2
/s/	S
/n/	N
/g/	G
/r/	R
/c/	C
/h/	H
/j(y)/	Y4, Y9*3
/N/	X
silence	•

*1 I2 indicates the nasalized/i/in/ni/.

*2 K6, K9 indicate the plosive/k/in/roku/and fricative/k/ in/kyu/respectively.

*3 Y4, Y9 indicate the semivowel/j/in/jon/and yoōn/j/in/kju/respectively.

reduce the computational time, pre-selection of the reference patterns for matching is required. The problem of statistical decision (2) is the processing redundancy that all the feature parameters are used even in the case where a few parameter are sufficient. The problem of logical decision (3) is the low matching ability due to the lack of the error recovery.

We employ the mixed technique (1) and (3) to reduce the computational time by using the technique (3), to some extent, for pre-selection of the reference patterns to be matched.

4.2.2 Advantages of Labeling after Segmentation

The phonemic labeling is performed on the divided segments of the input speech. The advantages of this kind of labeling are as follows.

(1) Contextual (inter-segments) information is available so that plosives can be detected by checking the previous segment to be silent or not.

(2) Intra-segment information is available so that the candidates for corresponding phonemic labels can be reduced by using the peak and valley in the spectrum.

(3) Coarticulation effect can be avoided by selecting the middle frames and discarding the frames around the segment boundaries. This process has the another advantage that the processing time is reduced due to the reduction of the data.

Taking these advantages into consideration, we employ the segment labeling technique based on the middle three frame labeling within the segments.

4.2.3 Flow in the Labeling Process

Fig. 4–1 shows the flow in the labeling process. The process is divided into three blocks. The first is the frame selection block to select the frames to be matched with the phonemic reference patterns within the segment. The second is the frame–based–labeling block to match the selected three frames with the phonemic reference patterns. The last one is the segment label decision block to decide the segment label based on the majority decision about the phonemic labels of the three frames.

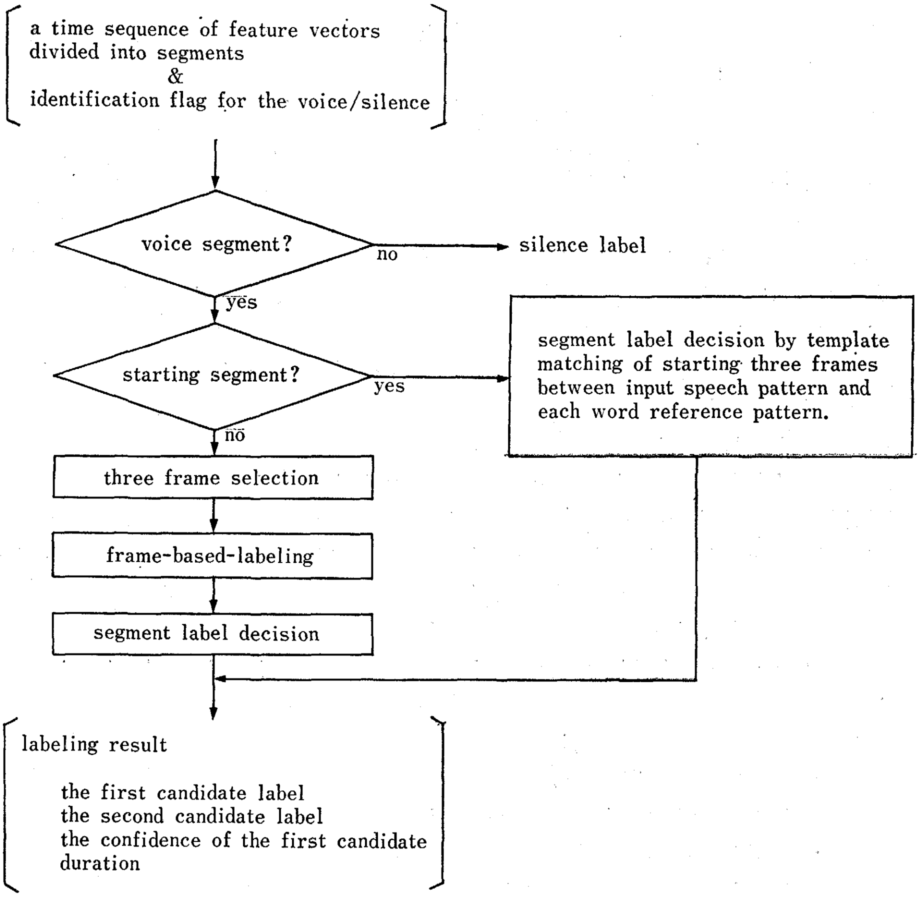

a time sequence of feature vectors divided into segments & identification flag for the voice/silence
voice segment?
no
silence label
yes
starting segment?
yes
segment label decision by template matching of starting three frames between input speech pattern and each word reference pattern.
no
three frame selection
frame-based-labeling
segment label decision
labeling result
the first candidate label
the second candidate label
the confidence of the first candidate
duration

Fig. 4–1. Flow in the labeling process.

(1) Frame selection block

The middle three frames are selected from the segments. For the plosive segments which can be detected by checking the existence of silence before the segment (contextual information), the first three segments are selected because they shows their effective features. For the segments before silence, the first three segments are also selected because of the decrease of the power at the middle frames within the segment.

(2) Frame based-labeling

Phonemic labels are assigned to the selected three frames in each segment. Before template matching between the three frames and the phonemic reference patterns, the number of the phonemic label candidates to be assigned to each segment is reduced by a rough decision tree (logical decision for the labeling). The rough decision tree classifies the phonemic labels into four groups in the following manner.

(a) Plosive consonants are decided by the existence of silence before the segment.

(b) Consonants are decided by the existence of the valley of the power within the segment.

(c) Vowels and semivowels are decided by the existence of the peak of the power within the segment.

(d) Another phonemic labels are grouped which are not included in (a) to (c).

Among the reduced phonemic label candidates for the frames, phonemic labels are determined by using the concentration of the power and template matching with phonemic reference patterns.

The phonemic reference patterns are created by averaging the phonemic patterns of ten males.

(3) Segment label decision

Segment labels are decided based on the majority decision about the phonemic labels for three frames. For the starting segment of the input speech, to cope with the distortion, the phonemic labels are decided by selecting the first three frames from the segment and matching them with the phonemic reference patterns for the heading part of each word.

The output from this block for the i-th segment contains the following information.

$$(I_1,\ I_2,\ P_{in},\ T)_i \quad (i=1\sim l_{in})$$

I_1: the first candidate phoneme

I_2: the second candidate phoneme if necessary

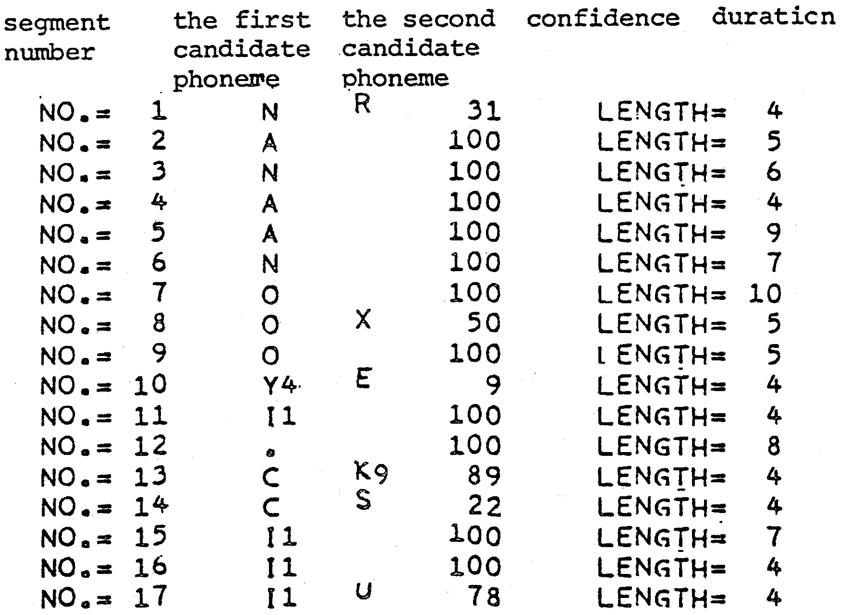
P_{in}: confidence of the first candidate phoneme

T : duration of the segment in terms of the number of frames

where l_{in} is the number of segments contained in the input speech. We call this output the input pattern label.

As the result of this labeling for the input speech, a sequence of input pattern labels are obtained. Fig. 4–2 shows the example of the labeling and a sequence of the input pattern labels for the input speech 「751」.

4. 3 Experimental Result



```
segment     the first   the second   confidence   duraticn
number      candidate   candidate
            phoneme     phoneme
 NO.=  1        N        R      31      LENGTH=  4
 NO.=  2        A              100      LENGTH=  5
 NO.=  3        N              100      LENGTH=  6
 NO.=  4        A              100      LENGTH=  4
 NO.=  5        A              100      LENGTH=  9
 NO.=  6        N              100      LENGTH=  7
 NO.=  7        O              100      LENGTH= 10
 NO.=  8        O        X      50      LENGTH=  5
 NO.=  9        O              100      LENGTH=  5
 NO.= 10        Y4       E       9      LENGTH=  4
 NO.= 11        I1             100      LENGTH=  4
 NO.= 12        .              100      LENGTH=  8
 NO.= 13        C        K9     89      LENGTH=  4
 NO.= 14        C        S      22      LENGTH=  4
 NO.= 15        I1             100      LENGTH=  7
 NO.= 16        I1             100      LENGTH=  4
 NO.= 17        I1       U      78      LENGTH=  4
```

Fig. 4-2. Example of the labeling ([751]/nanagoici/)

Table 4-3. Experimental result of labeling. (unit %)

out \ in	A	I1	I2	U	O	C	K6	K9	C	S	H	N	G	R	Y4	Y9	X	.	omission	total
A	92.8			0.7	1.7	1.7									1.1	0.1	1.6	0.2	0.2	1020
I1		45.1	14.5	10.6	2.7			2.5	0.1				0.3	0.6	9.7	4.3	9.4	0.1		678
I2		5.4	50.4	8.4	0.4	0.4		3.6			0.2	3.1			7.7		20.1	0.2		478
U	2.4			61.2	0.3	16.2						2.4	2.6		7.4	0.3	3.5	3.8		340
O	5.5			0.5	74.5	9.0									9.9	0.5				365
C	8.9			2.8	0.3	85.0									2.0		0.9		0.1	892
K6							90.4		0.8				8.0						0.8	125
K9		7.3						91.3					0.5	1.0						207
C		0.3						6.9	89.0	3.6				0.3						364
S								1.0	5.2	92.9		0.5	0.5							210
H											88.0	0.7		3.3					8.0	150
N		0.5		4.1	0.9	0.2					1.4	52.1	0.7	14.0	11.9	1.8	11.0		1.4	436
G				2.2			0.7				14.2	31.3	38.1	3.7			0.7		9.0	134
R	0.9			0.9							1.3	16.6	0.4	49.8	5.7	0.4			24.0	229
Y4	4.5	4.5	5.1	1.1	7.9	0.6						2.8		8.4	56.7	4.5	2.2		1.7	178
Y9		2.5	0.3	0.6	6.7						0.3	1.3			57.3	30.9				314
X	1.7	0.5		17.9	0.7	2.1					1.2	11.7			9.8	1.2	52.0	0.5	0.7	419

Table 4–3 shows the confusion matrix obtained by the phonemic labeling. The experimental conditions are as follows

speaker: three males

input speech: 10 one–digit–numerals

100 two–digit–numerals
50 three–digit–numerals
20 four–digit–numerals

From the Table 4–3, the followings are concluded.

(1) (S, C, H, K9) shows the high recognition accuracy due to their stability.

(2) K6 also shows the high recognition accuracy due to the pre–decision as plosive before labeling.

(3) Voiced consonants show the low recognition accuracy due to the large spectrum variations within the segment.

(4) Recognition errors to X or U occur frequently because high frequency components decrease at the end of the input speech so that the recognition error to X or U occurs which has relatively strong low frequency components.

(5) Y4 shows the low recognition accuracy due to its transient property.

6. CONCLUSION

In this paper, we described the phoneme recognition for speaker–independent connected word recognition.

To normalize the speaker variation, we employed the phoneme recognition by the segmentation and labeling in the segmentation, the optimal segmentation technique by the variance–minimization is proposed and showed the effectiveness for speaker variation.

In the phoneme labeling, segment labels are determined on the basis of the frame labels which are determined by template matching between phonemic reference patterns and the frames. It is devised that the phonemic reference patterns to be matched with the frames are reduced using the inter–segment and intra–segment information.

The remaining works will be the development of the connected word recognition on these labeled segment and the tuning up of the labeling abilities.

REFERENCES

[1] H. Sakoe: Two-level DP-matching-A dynamic programming based pattern matching algorithm for connected word recognition, IEEE Trans. Acoust., Speech, Signal Processing, ASSP-27, 6, pp.588-595, 1979.

[2] S. Nakagawa: A machine understanding system for spoken Japanese sentences, Doctoral thesis, Kyoto univ., 1976.

[3] H. Sakoe and S. Chiba: Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust., Speech, Signal Processing, ASSP–26, 1, pp.43–49, 1978.

(Aug. 31, 1983, received)