ISOLATED WORD RECOGNITION USING PHONEME-LIKE TEMPLATES

Noboru SUGAMURA,Kiyohiro SHIKANO and Sadaoki FURUI

Musashino Electrical Communication Laboratory
Nippon Telegraph and Telephone Public Corporation
Musashino,Tokyo 180,Japan

ABSTRACT

This paper describes new techniques for use in a word recognition system. This recognition system is especially effective in speaker-dependent large vocabulary word recognition or speaker-independent word recognition based on multiple reference templates. In this system, word templates are represented as sequences of descrete phoneme-like (pseudo-phoneme) templates which are automatically determined from a training set of word utterances by a clustering technique. In speaker-dependent 641 city names word recognition experiments, 96.3% recognition accuracy was obtained using 256 phoneme-like templates.

1. INTRODUCTION

In most of isolated word recognition systems, pattern matching techniques based on dynamic time warping are used and whole vocabulary words are assumed to be uttered in advance for training. Feature parameters, such as band-pass filter outputs or LPC parameters, are extracted from those training utterances and stored frame by frame as word templates. In the recognition stage a time sequence of feature parameters extracted from input speech and word templates are directly compared. We call this method direct-matching in this paper. Since spectral information in the direct-matching method is directly described using extracted parameters, high recognition accuracy can be obtained. However, in the large vocabulary word recognition, the amount of spectral distance calculation for dynamic time warping and the memory size for the word templates become very large. Another recognition system was also reported to avoid these problems (1),(2), in which spectral information is described as a sequence of phoneme spectral patterns and their durations.
However, it is sometimes difficult to determine phoneme templates and word templates automatically.

Apart from the word recognition field, narrow band transmission has been studied based on vector quantization techniques (3),(4). The speech has been ensured to be understandable in extremely low bit rate under 800 bps. After these experiments, it was conceived to represent word templates in speech recognition as sequencies of descrete spectral patterns which are vector quantized into a small number of spectral patterns. This new system has been established through several experiments (5).

This paper introduces the new word recognition system using Strings of Phoneme-LIke Templates, called SPLIT method, and its features. Its efficiency in large vocabulary word recognition and speaker-independent word recognition using multiple references is described.

2. NEW ISOLATED WORD RECOGNITION SYSTEM,SPLIT

A SPLIT system blockdiagram is shown in Fig.1. In the first stage of this system, phoneme-like templates and word templates are generated using training speech samples.

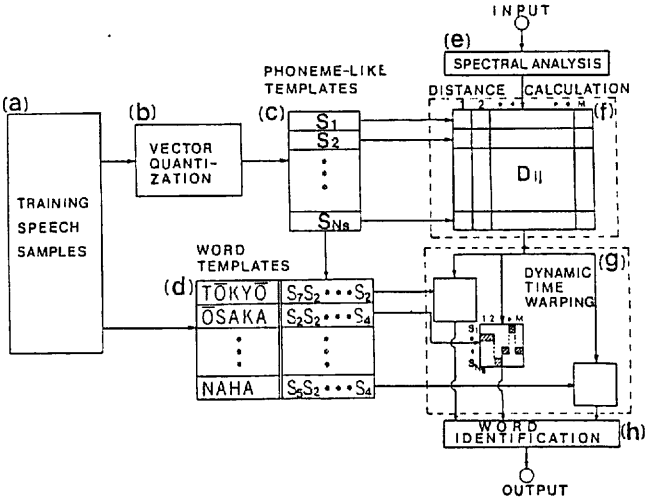

Fig.1 Word recognition system blockdiagram using Strings of Phoneme-like Templates, SPLIT.

16.3

CH1841-6/83/0000-0723 $1.00 © 1983 IEEE

2.1 Generating Phoneme-like Templates

The phoneme-like templates are generated automatically by a clustering technique using training utterances. The clustering algorithm is as follows, $E_o = (g_1, g_2, \text{—}, g_n)$ is a training spectral vector set, which consists of a few thousand frames arbitrarily selected from training utterances. From E_o, the K phoneme-like templates set $F=(f_1, f_2, \text{—}, f_K)$ is generated. $C(E_i)$ is defined as the mean vector of E_i and symbol $d(g_i, g_j)$ is defined as the spectral distance between g_i and g_j.

(1) Initialization:Set δ (the distance threshold for vectors which are regarded as the same cluster) and K (the number of phoneme-like templates desired). Set k=1.

(2) Compute the spectral distance $d(g_i, g_j)$.

(3) Count the number of g_j which satisfies $d(g_i, g_j) < \delta$ for all i. This vector set is represented as $B(g_i, \delta)$ and the number is represented as N(i).

(4) Find the vector g_i which has the maximum number of N(i).

(5) Generate the mean vector by calculating mean value of every auto-correlation coefficient from all the vectors belonging to $B(g_i, \delta)$ and stored as f_i, one of the phoneme-like templates.

(6) $E_k = E_{k-1} - B(g_i, \delta)$

(7) If $E_k = 0$ (empty set) or k=K, then stop, otherwise replace k by k+1 and go to step (3).

This algorithm is very simple, because it is only necessary to give K and δ. In this method, the spectral distance between two vectors is calculated only one time. After step (3), only the number which satisfies $d(g_i, g_j) < \delta$ is calculated, considering the vectors eliminated at step (6).

2.2 Generating Word Templates

Word templates are represented as sequences of phoneme-like templates. Namely, each training word is divided into a 16 msec duration succession and spectral distance between each segment and each phoneme-like template is calculated. The symbol of phoneme-like template which minimizes the spectral distance is stored in every frame. Word templates are stored as sequences of phoneme-like templates instead of using the exact spectral parameters.

2.3 Word Recognition

The phoneme-like templates and word templates are stored at (c) and (d), respectively, in Fig.1. Word recognition is carried out as follows. The input utterance is analyzed every 16 msec and auto-correlation coefficients and LPC cepstrum coefficients are extracted at (e) in Fig.1. The spectral distance between each input word frame and each phoneme-like template is stored as an element of a distance matrix of (f). In this system, WLR(Weighted Likelihood Ratio) is used as a distance measure (6). The total spectral distance between input word and each word template is calculated by summing up elements of the spectral distance matrix, referring to the sequence of the word template. An efficient time warping algorithm is used to minimize the total spectral distance (7),(8).

2.4 Some Features of the SPLIT System

The SPLIT system has the following significant features over the direct-matching.

(1) Drastic memory saving for word templates

In the SPLIT system, word templates are represented by sequences of phoneme-like templates. Thus drastic memory saving is achieved in comparison with the direct-matching. The saving ratio is calculated approximately as a function of the word templates number.

Assume the vocabulary size is L words. The i-th word has M_i frames. Each frame is N dimentional feature vector and its accuracy is N_a bits. Denote the number of phoneme-like templates as N_s. Using these notations, the memory amounts for word templates in the direct-matching (R_d) and SPLIT method (R_s) are given as

$$R_d = \left(\sum_{i=1}^{L} M_i\right) \cdot N \cdot N_a \qquad (1)$$

$$R_s = N_s \cdot N \cdot N_a + \left(\sum_{i=1}^{L} M_i\right) \cdot n_b \qquad (2)$$

where $n_b = \log_2 N_s$

The reduction ratio for the SPLIT method to the direct-matching is

$$R_{WD}(L) = \frac{N_s \cdot N \cdot N_a + \left(\sum_{i=1}^{L} M_i\right) \cdot n_b}{\left(\sum_{i=1}^{L} M_i\right) \cdot N \cdot N_a} \qquad (3)$$

(2) Drastic distance calculation saving for a dynamic time warping

In the SPLIT system, spectral distance calculation amount depends only on the number of templates. The saving ratio is calculated as follows.

The number of input speech frames is M_I. The length of window in dynamic time warping is assumed to be N_w. Then, calculation amounts are represented respectively.

$$C_d = M_I \cdot N_w \cdot L \qquad (4)$$

$$C_s = M_I \cdot N_s \qquad (5)$$

where C_d is the distance calculation amount in the direct-matching and C_s is in the SPLIT.

The reduction ratio is

$$R_{CAL}(L) = \frac{M_I \cdot N_s}{M_I \cdot N_w \cdot L} = \frac{N_s}{N_w \cdot L} \qquad (6)$$

The reduction ratio of calculation and memory amount for word templates compared with the direct-matching is shown in Fig.2 as a function of the number of words to be recognized.

In this figure, parameters are assumed to have following values.

M_i =50 frames (for all i,for simplicity), N=16 parameters, N_a =16 bits, N_s =256, n_b =8 bits and N_w=15 frames.

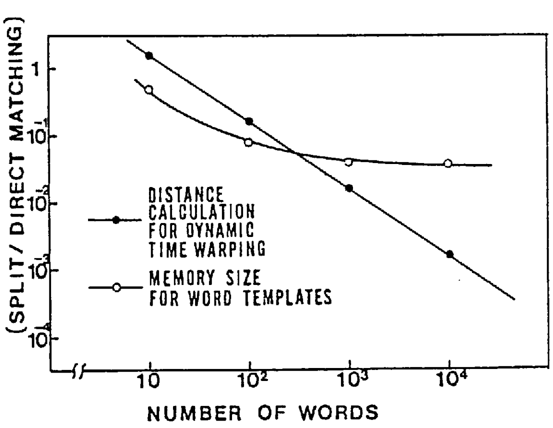

Fig.2 Relation between number of words and reduction ratio of the SPLIT method to the direct-matching from the viewpoint of memory size and the amount of distance calculation for dynamic time warping.

3. EXPERIMENTAL RESULTS ON LARGE VOCABULARY SIZE SPEAKER-DEPENDENT WORD RECOGNITION

3.1 In the Case of 256 Phoneme-like Templates

In order to verify the SPLIT system ability in large vocabulary word recognition, 641 city name utterances by four male speakers were used. Every speaker uttered each word twice at two week intervals. Experimental conditions are as follows.

Input speech is band limited to 4 kHz, sampled at 8 kHz and is converted into digital signal by a 12 bit AD converter. After passing the 32 msec Hamming window, auto-correlation coefficients are derived every 16 msec. After these processes, 10th order LPC analysis is executed. The first utterance set is used in generating phoneme-like templates and word templates. The second utterance set is used for a test at the first experiment. At the second experiment, such conditions are inverted. In the experiments, 2,048 frames were used for each speaker to generate 256 phoneme-like templates.

96.3 % recognition accuracy was obtained on average for four speakers by the SPLIT method. Degradation from the direct-matching was only 0.4 %, which means that spectral information can be roughly quantized in each frame without largely decreasing the recognition accuracy.

3.2 In the Case of Phoneme-like Templates Fewer than 256

Next stage, the relation between the number of phoneme-like templates and the recognition accuracy was investigated. Five kinds of phoneme-like template sets were generated, each of which consist of 16, 32, 64, 128 or 256 templates, by changing the threshold value δ for the clustering.

A preliminary experiment showed the strong correlation between the spectral distortion in generating word templates and the recognition accuracy in speaker-dependent word recognition. Based on this experiment, the phoneme-like templates were generated, so that total spectral distortion of the word templates represented by the sequences of the phoneme-like templates became the minimum. Using these optimum template sets, recognition experiments were carried out by the utterances of the four speakers. The relation between the averaged recognition accuracy and the number of phoneme-like templates is shown in Fig.3. This figure shows that the recognition accuracy does not decrease rapidly by the decreasing of the number of phoneme-like templates, and 92.9 % recognition accuracy is obtained even when the number of phoneme-like templates is 16.

A method of minimizing the spectral distortion was reported in (9). In order to compare the efficiency of this method with ours, this method was also tried to generate the phoneme-like templates. The experimental results show that, when 128 phoneme-like templates were used, the mean spectral distortion averaged over all training utterances by the four speakers was 0.068, which was slightly less than that of our method (0.080). However, the recognition accuracy was nealry equal for two methods (10). This result means that small spectral distortion difference hardly influences the accuracy of the recognition based on the dynamic time warping.

In the case of 32 phoneme-like templates, degradation from the direct-matching was 2.2 %, where the amount of distance calculation and word templates are 0.3 % and 2 % of the direct-matching, respectively. The small number of phoneme-like templates can be used in several application fields.

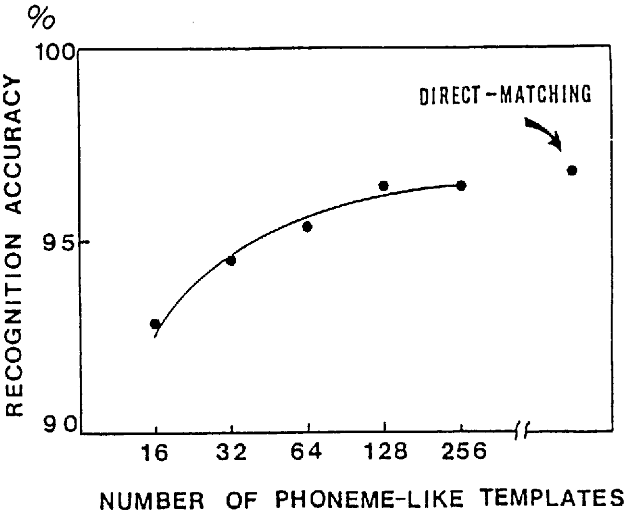

Fig.3 Relation between the averaged recognition accuracy and the number of phoneme-like templates

4. SPEAKER-INDEPENDENT ISOLATED WORD RECOGNITION

This section describes a speaker-independent word recognition system based on multiple templates using the SPLIT method. In the SPLIT system, as the spectral distance calculation amount is independent of the number of word templates, the best use can be made in speaker-independent word recognition. This system has the following important study items.

(1) How to make the multiple word templates (11).

(2) How to make the phoneme-like templates which are used commonly for all speakers.

The nearly same algorithm used in generating phoneme-like templates is applied for selecting multiple templates. The different points are to replace the spectral distance between two vectors by the spectral distance between two words using dynamic time warping and to select word templates instead of computing mean templates. The word templates are selected on the basis of the distribution of word utterances. Namely, if W_i has many words in its near neighbor, that word W_i is picked up as one of the word templates. for the second item, phoneme-like templates were generated using utterances by three male speakers and one female speaker.

In the recognition experiments, 8,184 words (31 words/speaker) uttered by 264 speakers through telephone lines were used. To check the effect of choosing multiple templates, the following experiment was executed.

(1) Using 264 utterances, the multiple templates were generated for each word by the clustering technique. Average number of word templates for each word was set to 18. Utterances by the all speakers, excepting the speaker whose utterance was used as one of the multiple templates, were used for recognition test.

(2) The recognition experiment without clustering was also carried out. 264 speakers were arbitrarily divided into groups of about 31 speakers. A speaker was selected from 31 speakers and word utterances by that speaker were recognized using utterances by 30 other speakers as multiple templates. Rotating this 31 times, the average recognition accuracy was calculated for each group.

Experimental results are shown in Table.1. Results show that multiple templates, which were generated by clustering, work quite well.

Table.1 Clustering effect in generating multiple word templates for speaker-independent word recognition.

Condition	With Clustering (18 templates/word)	Without Clustering (30 templates/word)
Recognition Accuracy	98.0 %	97.2 %

5. CONCLUSION

This paper proposes a new word recognition system, SPLIT, and describes some features of this system. The efficiency of this system in speaker-dependent large vocabulary word recognition and speaker-independent word recognition based on multiple templates was clarified through several experiments. Further studies include improving the method of choosing multiple templates for speaker-independent large vocabulary word recognition.

ACKNOWLEDGEMENT

The authors appreciate the guidance provided by Dr.Noriyoshi KUROYANAGI, Director of Research Division, Dr.Shin-ichiro HASHIMOTO, Deputy Director of Research Division and Dr.Fumitada ITAKURA, Chief of the Fourth Research Section. Special thanks go to many speakers and members of the Forth Research Section.

REFERENCES

(1) M.Kohda,S.Saito,"Speech recognition by Incomplete Learning Samples", IEEE Conf. on Speech Commun., Processing,Rep.H-10,1972.

(2) S.Furui,"A Training Procedure for Isolated Word Recognition Systems",IEEE Trans. on ASSP,vol. ASSP-28, pp.129-136, Apr. 1980.

(3) C.P.Smith,"Perception of Vocoder Speech Processed by Pattern Matching",JASA, vol.46, No.6, pp.1562-1571 July.1969.

(4) N.Sugamura,F.Itakura,"Speech Data Compression by Spectral Pattern Matching", IECE Trans., vol.J65-A, No.8, PP.834-841, Aug.1982.

(5) N.Sugamura,S.Furui, "A Large Vocabulary Word Recognition System Using Pseudo-Phoneme Templates", IECE Trans., vol.J65A-D, NO.8, pp.1041-1048, Aug. 1982.

(6) M.Sugiyama,K.Shikano, "LPC Peak Weighted Spectral Matching Measure", IECE Trans., vol.J64-A, No.5 pp.409-416, May.1981.

(7) H.Sakoe,S.Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition",IEEE Trans. on ASSP,vol.ASSP-26,pp.43-49,Feb.1978.

(8) K.Shikano,M.Sugiyama, "Evaluation of LPC Spectral Matching Measure for Spoken Word Recognition", IECE Trans., vol.J65-D, No.5, pp.535-541, May.1982.

(9) Y.Linde,A.Buzo,R.M.Gray, "An Algorithm for Vector Quantizer Design", IEEE Trans. on Com., vol. COM-28 pp.84-95,Jan.1980.

(10) M.Sugiyama,K.Shikano, "Centroid for LPC Spectral Measure", Trans. of the Committee on Speech Research, ASJ, S82-12, May.1982.

(11) L.R.Rabiner,"On Creating Reference Templates for Speaker Independent Recognition of Isolated Words", IEEE Trans. on ASSP, vol.ASSP-26, pp.34-45, Jan. 1978.

IECE : the Institute of Electrical and Communication Engineers of Japan.

ASJ : Acoustical Society of Japan