<table>
<thead>
<tr>
<th>論文・著書情報</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>論題（和文）</strong></td>
</tr>
<tr>
<td><strong>著者（和文）</strong></td>
</tr>
<tr>
<td><strong>発行日</strong></td>
</tr>
</tbody>
</table>
Isolated word recognition using strings of phoneme-like templates (SPLIT)

Noboru Sugamura and Sadaoki Furui

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

(Received 2 February 1984)

This paper reports on the results of the application of a vector quantization technique to an isolated word recognition system. The basic idea underlying this system is to represent a speech spectral sequence by several discrete spectral symbols. In this system, word templates are represented as sequences of discrete phoneme-like (pseudo-phoneme) templates which are automatically generated from a training set of word utterances by a clustering technique. The new word recognition system and its advantages are explained. This recognition system is especially effective for speaker-dependent large-vocabulary word recognition, as well as speaker-independent word recognition using multiple word templates.

PACS number: 43. 70. Sc

1. INTRODUCTION

Most isolated word recognition systems use pattern matching techniques based on dynamic time warping.1,2 In these systems, it is assumed that the whole vocabulary has been uttered in advance to provide recognition standard patterns. Feature parameters, such as band pass filter outputs or LPC parameters, are extracted from the training utterances and stored frame by frame as word templates. In the recognition stage, the time sequence of feature parameters extracted from input speech and word templates are directly compared. We will be calling this method “direct-matching” in this paper. In the direct-matching method, since spectral information for every frame is precisely described using the extracted parameters themselves, high recognition accuracy can be obtained. However, when attempts are made to extend the word recognition vocabulary to a very large scale, the spectral distance calculation volume for dynamic time warping and memory size for word templates become very large. This is because the spectral calculation must be executed every frame for every template. Another recognition system has been reported for avoidance of these problems.3,4 This system employs phoneme spectral templates and a word dictionary represented by phoneme sequences, which contain the duration time for each phoneme. However, determination of the phoneme spectral templates and the word dictionary is sometimes difficult.

Separate from the word recognition field, narrow band transmission has also been studied that is based on vector quantization techniques.5,6 Using such examination techniques, it has ensured that speech is intelligible at an extremely low bit rate of under 1,000 bps. As a result of these experiments, it has come to be thought that word templates in speech recognition can be represented by sequences of discrete spectral patterns obtained by vector quantization. Based on this idea, we have invented a new word recognition system using Strings of Phoneme-Like Templates, which we call SPLIT method. The name SPLIT is derived also from the concept of SPLITting the multidimensional spectral space to produce the phoneme-like templates.
2. A NEW ISOLATED WORD RECOGNITION SYSTEM, SPLIT

2.1. Generating Phoneme-like Templates

A block diagram of the SPLIT system is shown in Fig. 1. In the first stage, phoneme-like templates and word templates are automatically generated, using training speech utterances. Phoneme-like templates do not exactly correspond to real phonemes. They are used as symbols representing discrete speech spectra. In this automatic generating process, there is no ambiguous process guided by human decision. More concretely the process for selecting discrete spectral patterns is as follows.

Training speech utterances are analyzed, and 1st through 13th-order autocorrelation coefficients and LPC cepstrum coefficients are extracted every 16 ms. A clustering technique is applied to the spectral vector set to generate the phoneme-like templates.

[Vector Quantization Algorithm]

$E_0 = (g_1, g_2, \ldots, g_N)$ is a training spectral vector set, consisting of a few thousand frames which are arbitrarily selected from training utterances. Phoneme-like templates set $F = (f_1, f_2, \ldots, f_S)$ are generated from $E_0$ as follows, where $d(g_i, g_j)$ is spectral distance between $g_i$ and $g_j$.

1. Initialization: Set $\theta$ (distance threshold for regarding vectors as in the same cluster) and $K$ (the number of phoneme-like templates to be generated).

2. Compute the spectral distance $d(g_i, g_j)$ for every combination of $i$ and $j$.

3. Extract vector $g_j$ which satisfies $d(g_i, g_j) < \theta$ for each $i$. This vector set is represented as $B(g_i, \theta)$ and the number of vectors is represented as $N(i)$.

4. Find vector $g_j$ which has the maximum $N(i)$.

5. Generate speech spectral patterns: The mean vector is calculated by averaging the autocorrelation coefficients of all vectors which belong to $B(g_i, \theta)$. This mean vector is then stored as one of the phoneme-like templates.

6. Eliminate the vector set $B(g_i, \theta)$ from $E_k$; $E_{k+1} = E_k - B(g_i, \theta)$

7. If $E_{k+1} = \emptyset$ (empty set) or $k = K$, stop; otherwise replace $k$ by $k + 1$ and go to step 3.

This algorithm is so simple that it needs only $K$ and $\theta$ to be set. In this method, a set of spectral distances between each pair of vectors is calculated only once. After step 3, only the number which satisfies $d(g_i, g_j) < \theta$ is calculated, using the vectors left at step 6.

The clustering method proposed here is based on the minimax criterion. Spectral patterns in high density space are extracted earlier than those in low density space. This method can suppress the maximum spectral distortion at a given threshold. Several other clustering techniques have been reported in other papers. Most of these algorithms are based

Fig. 1 Isolated word recognition system using phoneme-like templates (SPLIT system).
on the criterion for decreasing mean spectral distortion. In these methods, however, the mean spectral pattern is influenced by isolated patterns while mean spectral distortion is being decreased. The algorithm proposed here, though, can avoid the influence of these isolated patterns while the mean spectra are being generated.

Sugiyama et al. have tried to apply another vector quantization technique proposed by Linde to generation of phoneme-like templates in the SPLIT method. The differences between two clustering methods were examined from the viewpoint of spectral distortion and recognition accuracy. Results are reported in a paper of 11).

2.2 Generating Word Templates

Word templates in the present method are represented as sequences of phoneme-like templates, not through using exact spectral parameters. Each training word utterance is divided into a 16 ms duration succession, and the spectral distance between each frame and each phoneme-like template is computed. The symbol for the phoneme-like template, which leads to a minimization of the spectral distance, is stored in every frame.

2.3 Word Recognition

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 word utterance is analyzed every 16 ms, and the autocorrelation coefficients and LPC cepstrum coefficients are extracted at (e) in Fig. 1. The spectral distance between the individual input word frame and individual phoneme-like template is stored as an element of distance matrix (f) in Fig. 1. Two kinds of spectral distance measures were experimentally investigated. One is a cepstrum distance measure (CEP) and the other is a Weighted Likelihood Ratio (WLR) proposed by Sugiyama and Shikano. These spectral distance measures can be defined as:

\[
CEP = 2 \sum_{i=1}^{p} (C^f_i - C^g_i) \quad (1)
\]

\[
WLR = 2 \sum_{i=1}^{p} (C^f_i - C^g_i)(V^f_i - V^g_i) \quad (2)
\]

where \(C^f_i\) and \(C^g_i\) are LPC cepstrum coefficients, and \(V^f_i\) and \(V^g_i\) are autocorrelation coefficients. Suffixes (f) and (g) correspond to input word utterances and phoneme-like template, respectively. The number of parameters, \(p\), used in the experiments is 16.

2.4 Dynamic Time Warping

The total spectral distance between an input utterance and each word template is calculated by summing up the elements of the spectral distance matrix, referring to the sequence of the word template. In isolated word recognition systems using dynamic time warping, the recognition accuracy is influenced by the efficiency of the dynamic time warping algorithm. Our recognition system uses an efficient slope-constrained, unconstrained-endpoint dynamic time warping algorithm proposed by Shikano. The dynamic time warping algorithm is shown in Fig. 2.

Assume that B is an input word, A is a word template, and \(D(i, j)\) is defined as the spectral distance between the \(i\)-th frame of B and the \(j\)-th frame of A. \(G(i, j)\) denotes the cost function at point \((i, j)\). In this algorithm, paths in only two directions are permitted at the starting point, which is controlled by the cost function. Initial conditions and iterations are as follows.

(1) Initialization: \(G(1, 1) = 2D(1, 1)\)

\[
G(i, j) = G(i, j - 1) + D(i, j) \quad 2 \leq j \leq r/2
\]

\[
G(i, 1) = G(i - 1, 1) + D(i, 1) \quad 2 \leq i \leq r/2
\]

\[
G(1, j) = \infty \quad r/2 < j \leq r
\]

\[
G(i, 1) = \infty \quad r/2 < i \leq r
\]

\[
r = \min ((i \text{end}, j \text{end}))/4 + 3
\]

Here \(i \text{end}\) and \(j \text{end}\) are the respective number of
frames for word templates and input speech. \( r/2 \) is the permissible width for the starting region.

(2) Iterations

\[
G(i, j) = \min \left\{ G(i-1, j-2) + 2D(i, j-1), G(i-1, j-1) + D(i, j), G(i-2, j-1) + 2D(i-1, j), D(i, j) \right\}
\]

\[
2 \leq i \leq i_{\text{end}}, \ 2 \leq j \leq j_{\text{end}}
\]  

(4)

(3) Decision: The total cost function, \( S(A: B) \), is

\[
S(A: B) = \min \left( G(i, j) / (i + j) \right) \text{ (i, j) } \epsilon \text{ end region.}
\]  

(5)

With this dynamic time warping, \((i_{\text{end}} - r/2 \sim i_{\text{end}}, j_{\text{end}}) \) and \((i_{\text{end}}, j_{\text{end}} - r/2 \sim j_{\text{end}}) \) are end point regions. Hence, even if the last frame \((i_{\text{end}}, j_{\text{end}})\) is out of the adjustment window, \( S(A: B) \) can be calculated under conditions of \(|i_{\text{end}} - j_{\text{end}}| < r + r/2 \). Otherwise, the input word is rejected.

2.5 Several Advantages of the SPLIT Method

The SPLIT method has the following significant advantages over the conventional direct-matching method.

2.5.1 Word template memory savings

In the SPLIT method, word templates are represented by sequences of phoneme-like templates. Thus, drastic memory savings can be achieved when a comparison is made with the direct-matching method. The savings ratio is calculated approximately as a function of the number of word templates.

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

\[
R_d = \sum_{i=1}^{L} M_i \cdot N \cdot N_s
\]  

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

Here, \( n_b = \log_2 N_s \)

The reduction ratio for SPLIT to direct-matching is

\[
R_s(L) = \frac{N_a \cdot N \cdot N_s + \left( \sum_{i=1}^{L} M_i \right) \cdot n_b}{\left( \sum_{i=1}^{L} M_i \right) \cdot N \cdot N_s}
\]  

(8)

2.5.2 Distance calculation savings for the dynamic time warping

In the SPLIT method, the spectral calculation volume depends only on the number of phoneme-like templates. The savings ratio can then be calculated as follows.

The frame number for input speech is \( M_i \). The window length for dynamic time warping is assumed to be \( N_w \). Distance calculation amounts are:

\[
C_i = M_i \cdot N_a \cdot L
\]  

(9)

\[
C_s = M_s \cdot N_s
\]  

(10)

Here, \( C_s \) is the distance calculation amount for the direct-matching method, and \( C_s \) is the calculation amount with the SPLIT method.

The reduction ratio is

\[
R_s(L) = \frac{M_i \cdot N_a \cdot L}{M_s \cdot N_s \cdot L} = \frac{N_s}{N_s \cdot L}
\]  

(11)

The reduction ratio for spectral distance calculations and memory size with word templates, is shown in Fig. 3 as a function of vocabulary size, that is number of words to be recognized. A comparison is made with the direct-matching method. In this figure, parameter values are assumed to be

\[
M_i = 50 \text{ frames (for all } i, \text{ for simplicity)}
\]
N. SUGAMURA and S. FURUI: ISOLATED WORD RECOGNITION USING SPLIT METHOD

\[ N = 16 \text{ parameters} \]
\[ N_t = 16 \text{ bits} \]
\[ N_r = 256 \text{ templates} \]
\[ N_f = 15 \text{ frames} \]
\[ m_r = 8 \text{ bits} \]

The figure shows that the reduction ratios for both spectral distance calculation and word template memory size are drastic. That is, the SPLIT method takes up only a few percent of what the direct-matching method requires for 1,000 word recognition.

3. LARGE- VOCABULARY SPEAKER- DEPENDENT WORD RECOGNITION

3.1 Recognition Using 256 Phoneme-like Templates

In order to verify the effectiveness of the SPLIT method in large vocabulary word recognition, 641 city-name utterances by four male speakers were used. Every speaker uttered each word twice over a two-weekly interval. Experimental conditions were as follows.

Input speech was band-limited to 4 kHz, sampled at 8 kHz and then converted into digital signals by a 12 bit AD converter. After passing through the 32 ms Hamming window, thirteen auto-correlation coefficients were derived every 16 ms, and 10th order LPC analysis was executed. The first utterance set for each speaker was used in generating phoneme-like templates and word templates. In these processes, cepstrum distance measure (CEP) was used. The second utterance set was used for evaluation in the first experiment. In the second experiment, the combination of test and reference utterances was inverted. In these experiments, 256 phoneme-like templates were generated from 2,048 frames for every speaker. The following items were checked in the experiments.

(1) A comparison was made of recognition accuracies between the SPLIT method and the direct-matching method.

(2) A comparison was made of the spectral measure efficiency between the cepstrum distance and WLR distance.

Experimental results are listed in Table 1. This table shows the accumulated recognition accuracy. Based on these results, the following conclusions were reached.

(a) 96.3% recognition accuracy can be obtained for 641-word recognition on average for four speakers using the SPLIT method with a WLR spectral distance measure. Degradation from the direct-matching method was only 0.4%. These results mean that spectral information can be roughly quantized in each frame without resulting in a large decrease of recognition accuracy.

(b) 99.1% recognition accuracy can be obtained for the top five choices.

(c) Recognition accuracy with WLR was much higher than when cepstrum distance is employed in both methods. This means that the WLR spectral distance measure is effective for large vocabulary word recognition.

3.2 Classification of Unrecognized Words

Words which were not recognized correctly in the experiments were classified into the following categories.

(A) Endpoint detection error
(ex. SAKU->KUSATSU)

(B) Vowel confusion
(ex. UOZU->OOZU)

(C) Beginning consonant error
(ex. SAGA->KAGA)

(D) Middle consonant error
(ex. OOGAKI->OOMACHI)

(E) Others
(ex. MINOO->MINO)

Classification results with four speakers are shown in Table 2. The following conclusions can be obtained from these results.

(1) Endpoint detection error is only 1% of all
Table 2 Absolute number of recognition error occurrences in recognition of 641 city names. Each recognition error is classified into the one of the following five categories.

<table>
<thead>
<tr>
<th>Spectral distance measure</th>
<th>CEP</th>
<th>WLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition method</td>
<td>Direct-matching SPLIT</td>
<td>Direct-matching SPLIT</td>
</tr>
<tr>
<td>Error category</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>B</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>C</td>
<td>57</td>
<td>61</td>
</tr>
<tr>
<td>D</td>
<td>61</td>
<td>67</td>
</tr>
<tr>
<td>E</td>
<td>42</td>
<td>38</td>
</tr>
</tbody>
</table>

A: Endpoint detection error, B: Vowel confusion, C: Beginning constant confusion, D: Middle consonant confusion, E: Others

(2) The degree of vowel confusion is very small.
(3) WLR spectral measure decreases consonant confusion.

3.3 Recognition Using Fewer than 256 Phoneme-like Templates

The relation between the number of phoneme-like templates and spectral distortion when generating word templates, was examined, and the results are shown in Fig. 4. The arrows in the figure indicate the point of maximum recognition accuracy for each threshold condition. In this figures, a strong correlation is observed between spectral distortion when generating word templates and recognition accuracy.

Recognition accuracy was investigated as the next step, examining it as a function of the number of phoneme-like templates using utterances by one male speaker. Five kinds of phoneme-like template sets were generated, each of which consisted of 16, 32, 64, 128, or 256 templates. This was done by varying the threshold value, θ, for clusterings from 0.05 to 0.30 with a step of 0.05. Using these six phoneme-like template sets, recognition experiments were then executed. The relationship between recognition accuracy and the number of phoneme-like templates is shown in Fig. 5. The results show that a high recognition accuracy can be obtained even when the number of phoneme-like templates is less than 64.

If the threshold, θ, is set appropriately, a 94.0% recognition accuracy can be obtained using only 16 phoneme-like templates. This number is much less than the number of Japanese phonemes. This means that, even if the spectral pattern is roughly quantized,
Table 3  Recognition accuracy for each speaker in relation to number of phoneme-like templates.  
\( N \) is number of phoneme-like templates and R.R. is recognition accuracy.  \( \theta \) is threshold value in generating phoneme-like templates.

<table>
<thead>
<tr>
<th>( N )</th>
<th>R.R.</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td>16</td>
<td>R.R.</td>
<td>92.5</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>0.20</td>
</tr>
<tr>
<td>32</td>
<td>R.R.</td>
<td>94.2</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>0.20</td>
</tr>
<tr>
<td>64</td>
<td>R.R.</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>0.15</td>
</tr>
<tr>
<td>128</td>
<td>R.R.</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>0.10</td>
</tr>
<tr>
<td>256</td>
<td>R.R.</td>
<td>96.6</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Direct-matching</td>
<td>97.0</td>
</tr>
</tbody>
</table>

Results show that the recognition accuracy does not decrease rapidly with a decrease in the number of phoneme-like templates; that is, 92.9% recognition accuracy can be obtained even when the number of phoneme-like templates is 16.  When there are 32 phoneme-like templates, degradation from the direct-matching method is 2.2%, where the distance calculation volume and memory size for word templates are 0.3% and 2% of the direct-matching method, respectively.  The smaller number of phoneme-like templates can be used in several fields of application where memory and calculation savings are urgent requirements.

3.4 Relationship between Phoneme-like Templates and Real Phonemes

The relation between 32 phoneme-like templates used in the recognition test and real phonemes was also investigated.  Each frame in the spoken words was manually labeled by phonemes, and correspondence between phoneme-like templates assigned to each frame and the labeled phoneme name was investigated.  Figure 7 shows the 32 phoneme-like templates mapped onto a two dimensional plane generated by multiple dimensional analysis and corresponding phonemes.  The results show that five or six phoneme-like templates belong to the same vowel cluster, and that several templates correspond to such consonants as /sl/, /m, n, /p, t, k/.
4. SPEAKER-INDEPENDENT WORD RECOGNITION BASED ON MULTIPLE WORD TEMPLATES

The SPLIT method as tested by large-vocabulary speaker-dependent word recognition can also be applied to speaker-independent word recognition. One of the most effective speaker-independent word recognition methods is a method based on multiple templates which cover the variability in spoken utterances among speakers. However, this system has several problems, one of which is the increase in calculation volume in proportion to the number of word templates.

Since the spectral distance calculation volume is independent of the number of word templates in the SPLIT method, it would seem that the best use of this method can be made in speaker-independent word recognition based on multiple word templates. To make it possible to apply the SPLIT method to such speaker-independent word recognition based on multiple templates, several important study items were experimentally investigated. These items included:

(1) How to make multiple word templates; and
(2) How to make phoneme-like templates which are used commonly all speakers.

As for the first item, nearly the same algorithm as that used in generating phoneme-like templates was applied to selection of multiple templates. In the process of selecting these multiple templates, the spectral distance between two words, as obtained through dynamic time warping was calculated instead of the spectral distance between two vectors. Moreover, typical word templates were selected instead of computing the average templates. The word templates were selected on the basis of the distribution of word utterances. Namely, word utterances which had many neighboring of word utterances were picked as word templates.

In look at the second item, three phoneme-like template sets were evaluated by a recognition experiment. The vocabulary consisted of 31 control words for a laboratory automation service. The first phoneme-like template set was generated using utterances by one male speaker, the second one was generated using utterances by one female speaker and the third one was generated using utterances by one male and one female speaker. These three sets were then tested by confirming word recognition with 39 male and 34 female speakers. The results of these experiments are shown in Table 4. It can be seen that the third set is most effective for both male and female speakers.

The next recognition experiment dealt with 8,184 words (31 words/speaker) uttered by 264 speakers through telephone lines and exchanges. Two kinds of tests were executed to verify the effectiveness of clustering the multiple templates.

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

(2) A recognition test where clustering was not employed was also carried out. 264 speakers were arbitrarily divided into groups of about 31 speakers.
N. SUGAMURA and S. FURUI: ISOLATED WORD RECOGNITION USING SPLIT METHOD

<table>
<thead>
<tr>
<th>Condition</th>
<th>With clustering (18 templates/word)</th>
<th>Without clustering (30 templates/word)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition accuracy</td>
<td>98.0%</td>
<td>97.2%</td>
</tr>
</tbody>
</table>

Word utterances by each speaker were recognized using utterances by 30 other speakers as multiple templates. Rotating this 31 times, the average recognition accuracy was then calculated for each group.

Experimental results are shown in Table 5. These results show that multiple templates, which were generated by the clustering technique, work quite well. High recognition accuracy can be obtained with the SPLIT method for speaker-independent word recognition, as well as for speaker-dependent word recognition.

5. CONCLUSIONS

This paper has been used for the proposal of a new word recognition method, named SPLIT, and several advantages of this method have been described. The efficiency of this method in speaker-dependent large-vocabulary word recognition and speaker-independent isolated word recognition based on multiple templates was clarified through several experiments. The experimental results show that spectral information used in isolated word recognition based on dynamic time warping can be roughly quantized in each frame without there being a large decrease in the recognition accuracy. Looking in more detail, the following results were obtained.

1. A 96.3% recognition accuracy was achieved for recognition of 641 city names spoken by four male speakers using 256 phoneme-like templates.

2. A 94.5% recognition accuracy was obtained using 32 phoneme-like templates. In this case, the spectral distance calculation volume and memory size for word templates are respectively 0.3% and 2% that for the conventional direct-matching method.

3. The effectiveness of a simple clustering technique in generating word templates for speaker-independent word recognition was ascertained, and 98.0% recognition accuracy could be achieved in a 31 word recognition test.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. F. Itakura for his encouragement and suggestions. They would also like to thank Dr. K. Shikano for his invaluable discussions and for letting us use his dynamic time warping algorithm. The authors are indebted to Mr. K. Hakoda for discussions and his offer of the speech database used in the early stages of this work.

REFERENCES

13) H. Sakoe and S. Chiba, "Dynamic programming

