

論文 / 著書情報
Article / Book Information

Title	Speaker-Dependent Large Vocabulary Word Recognition Using the SPLIT Method
著者	古井 貞熙
Author	Noboru Sugamura, Sadaoki Furui
掲載誌/書名	Review of the Electrical Communication Laboratories, Vol. 34, No. 3, pp. 327-333
Journal/Book name	, Vol. 34, No. 3, pp. 327-333
発行日 / Issue date	1986,

Speaker-Dependent Large Vocabulary Word Recognition Using the SPLIT Method

Noboru SUGAMURA and Sadaoki FURUI

This paper describes the results of applying a vector quantization technique to an isolated word recognition system. The basic idea underlying this system is to represent a speech spectral sequence using several discrete spectral symbols. In this system, word templates are represented as sequences of discrete phoneme-like templates which are automatically generated from a training set of word utterances by a clustering technique. The new word recognition system and its advantages in speaker-dependent large vocabulary word recognition are explained through reference to several experiments.

1 Introduction

Isolated word recognition systems are used on limited conditions: they are speaker-dependent and involve only a few hundreds of words etc. Isolated word recognition methods are divided into two categories: template matching using dynamic programming (DP)⁽¹⁾⁽²⁾, and phoneme recognition using feature parameters extracted from speech⁽³⁾. Pattern matching techniques based on DP are used in most isolated word recognition equipment, since DP matching is very effective for speaker-dependent word recognition. With these methods, the entire vocabulary is usually uttered in advance to provide word templates. Feature parameters, such as bandpass filter output or LPC parameters, are extracted from the training utterances and stored frame-by-frame as word templates. However, the amount of spectral distance calculation in executing DP matching is nearly proportional to the number of word templates, and thus becomes very large in large-vocabulary word recognition.

This paper proposes a new word recognition method where the amount of spectral distance calculation is independent of the number of word templates. Furthermore, an automatic word template generation method is described which reduces training utterances for generating word templates.

2 New Isolated Word Recognition Method, SPLIT

Isolated word recognition methods using DP matching are divided into two types according to how the word templates are represented. These two methods are shown in Fig. 1(a) and (b). In the first method (a), word templates are represented using the extracted parameters themselves.⁽¹⁾ In the recognition stage, the time sequences of feature parameters extracted from

input speech are directly compared with every word template. This method is called "direct-matching" in this paper. In the direct-matching method, since spectral information for every frame is precisely represented using the extracted parameters themselves, high recognition accuracy can be obtained. However, when the vocabulary increases 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 at every frame for every template.

Another recognition method⁽²⁾ has been reported to avoid these problems. This method employs phoneme spectral templates and word templates represented by phoneme sequences, which contain the duration time for each phoneme as shown in Fig. 1(b). This method

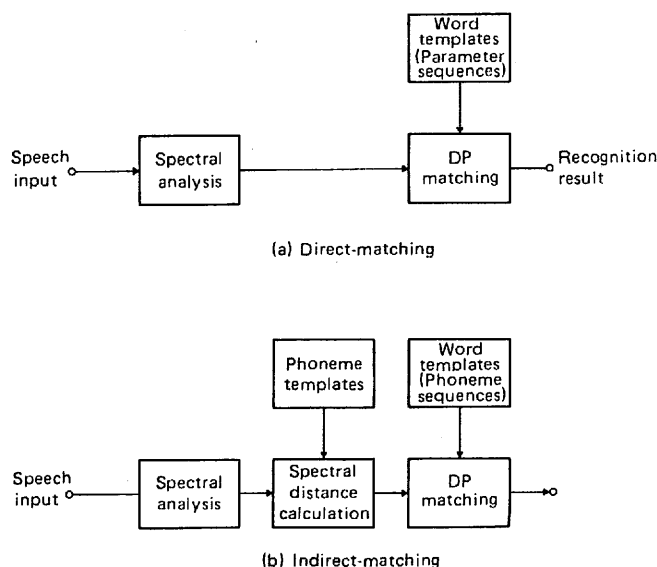


Fig. 1—Recognition methods based on DP matching.

is called "indirect-matching." With this method, however, generation of the phoneme spectral templates and representation of word templates are rather difficult. Furthermore, high recognition accuracy cannot be expected, since the spectral dynamics cannot be represented precisely.

This paper presents a new word recognition method located between these two methods. With this method, every word template is represented using a few hundred discrete spectral patterns. These patterns are called phoneme-like templates, since they do not correspond precisely to phonemes. These templates are generated automatically by a vector quantization technique⁽⁴⁾⁽⁵⁾. The method is called SPLIT, since word templates are represented using Strings of Phoneme-Like Templates⁽⁶⁾. The name SPLIT is also derived from the concept of SPLITting the multidimensional spectral space to produce the phoneme-like templates. Using these templates, the word templates can be represented more precisely than in the case of using phoneme templates. Therefore, the new method has the advantages of "indirect-matching" while also making possible high recognition accuracy.

2.1 Word Recognition Procedure in SPLIT

The word recognition procedure using the SPLIT method is outlined in Fig. 2. The training and recognition processes are:

(1) Generation of phoneme-like templates,

- (2) Generation of word templates and
- (3) Word recognition.

2.1.1 Generating Phoneme-like Templates

In the first stage, phoneme-like templates are automatically generated using training utterances. Phoneme-like templates do not exactly correspond to real phonemes. They are only used as symbols representing discrete speech spectra. The procedure for selecting discrete spectral patterns is as follows.

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

[Vector Quantization Algorithm]⁽⁵⁾

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

(1) Initialization: Set θ (distance threshold for regarding vectors as in the same cluster) and K (the number of phoneme-like templates to be generated). Set $k = 0$.

(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

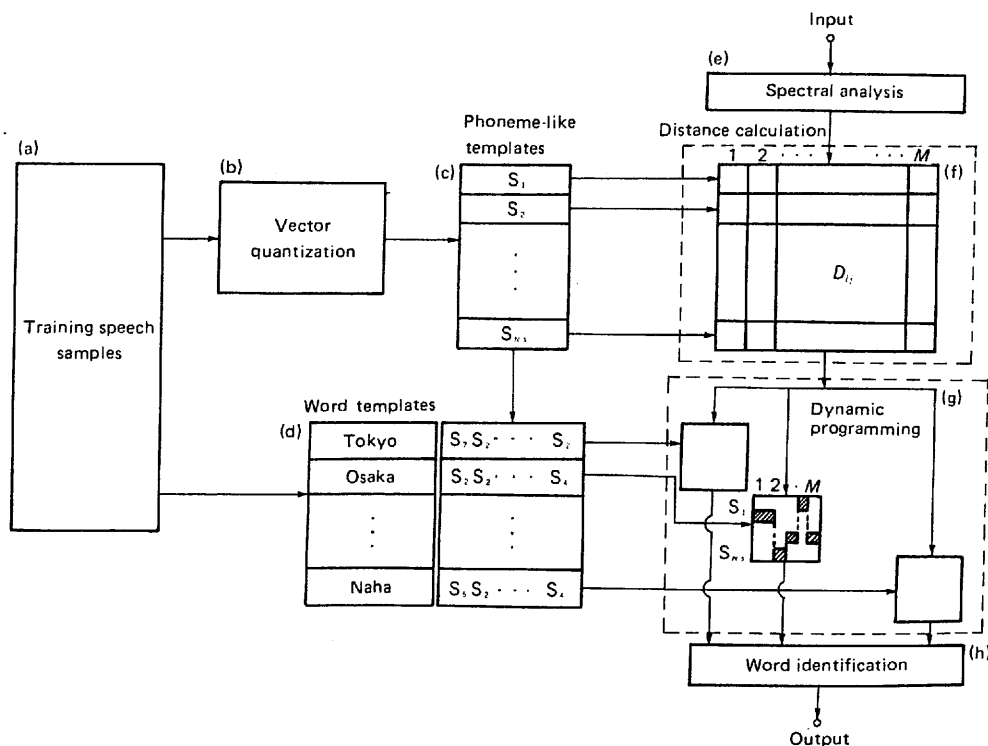


Fig. 2—Isolated word recognition system using phoneme-like templates (SPLIT).

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_i 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 stored as one of the phoneme-like templates.

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

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

Only K and θ must be given in this algorithm. 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.

2.1.2 Generating Word Templates

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

2.1.3 Word Recognition

Phoneme-like templates and word templates are stored at (c) and (d), respectively, in Fig. 2. Word recognition is carried out as follows.

The input word utterance is analyzed every 16 ms, and the autocorrelation coefficients and LPC cepstrum parameters are extracted at (e) in Fig. 2. 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. 2. The total spectral distance between an input word and each word template is calculated by summing up the elements of the spectral distance matrix, referring to the sequence of the word template. At the final stage, the minimum matching distance word is selected as the recognition result.

2.2 Several Advantages to SPLIT

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

2.2.1 Word Template Memory Savings

In the SPLIT method, word templates are represented by sequences of phoneme-like templates. Thus, drastic memory savings for word templates can be achieved when compared 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 and the i -th word has M_i frames. Each frame has N dimensional feature vectors, 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

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

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

Here, $n_b = \log_2 N_s$.

The reduction ratio for SPLIT to direct-matching is

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

2.2.2 Spectral Distance Calculation Savings for DP Matching

Using SPLIT, the spectral calculation amount depends only on the number of phoneme-like templates. The savings ratio can be calculated as follows.

The frame number for input speech is M_1 . The window length for DP matching is assumed to be N_w . Spectral distance calculation amounts are

$$C_d = M_1 \cdot N_w \cdot L \quad (4)$$

$$C_s = M_1 \cdot N_s \quad (5)$$

Here, C_d is the spectral distance calculation amount for the direct-matching method, and C_s is the calculation amount with the SPLIT method.

The reduction ratio is

$$R_c(L) = \frac{M_1 \cdot N_s}{M_1 \cdot N_w \cdot L} = \frac{N_s}{N_w \cdot L} \quad (6)$$

The reduction ratios for spectral distance calculations and memory size for word templates are shown in Fig. 3 as a function of vocabulary size, and a comparison is made with the direct-matching method. In Fig. 3, parameter values are assumed to be:

$M_i = 50$ frames (for all i , for simplicity),

$N = 16$ parameters,

$N_a = 16$ bits,
 $N_s = 256$ templates and
 $N_w = 8$ bits.

The figure shows that the reduction ratios for both spectral calculation and word template memory size are drastic. That is, the SPLIT method involves only a few percent of what the direct-matching method requires for 1000-word recognition.

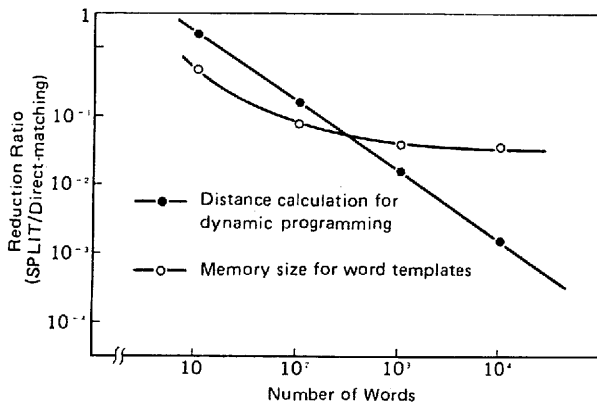


Fig. 3—Relationship between number of words and reduction ratio, from the viewpoint of memory size and amount of distance calculation for dynamic programming.

3 Large Vocabulary Speaker-dependent Word Recognition

3.1 Recognition Conditions

In order to verify the effectiveness of the SPLIT method in large vocabulary word recognition, 641 Japanese city-name utterances by four male speakers were used. Every speaker uttered each word twice over a two-week period. 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. 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 2048 frames for every speaker. In the experiment, Weighted Likelihood Ratio (WLR)⁽⁷⁾ was used as spectral distance measure. WLR can be defined as

$$\text{WLR} = 2 \sum_{i=1}^P (C_i^{(f)} - C_i^{(g)}) (V_i^{(f)} - V_i^{(g)}). \quad (7)$$

Here, $C_i^{(f)}$ and $C_i^{(g)}$ are LPC cepstrum parameters, and $V_i^{(f)}$ and $V_i^{(g)}$ 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 was 16.

Comparison of recognition accuracy was performed between the SPLIT method and direct-matching method.

3.2 Recognition Results

Experimental results are shown in Fig. 4. Based on these results, the following conclusions were arrived at.

(a) An average 96.3% recognition accuracy for four speakers can be obtained for 641-word recognition using the SPLIT method. 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 in recognition accuracy.

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

(c) Words that were not recognized correctly in the experiments can be classified into five categories: (1) Endpoint detection error (ex. SAKU → KUSATSU), (2) Vowel confusion (ex. UOZU → OZU), (3) Beginning consonant error (ex. SAGA → KAGA), (4) Middle consonant error (ex. OOGAKI → OOMACHII) and (5) others (ex. MINOO → MINO).

Endpoint detection error was only 1% of all data and vowel confusion was also very small.

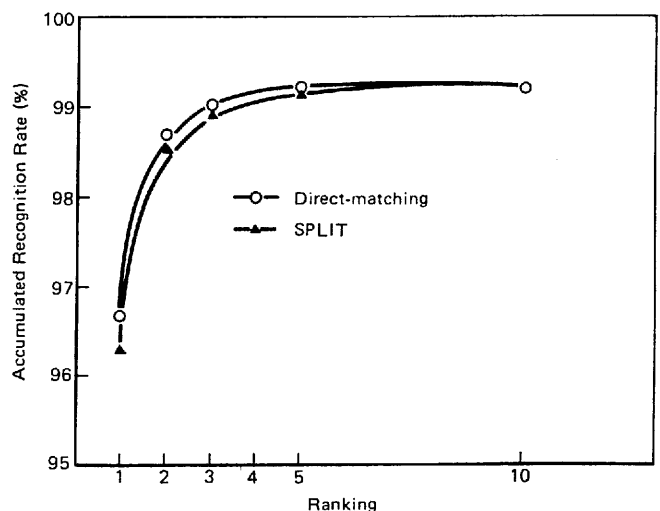


Fig. 4—Experimental results in speaker-dependent word recognition for 641 city names using WLR spectral distance measure.

4.2 Features of Proposed Method

The automatic word template generation method proposed here has the following features.

Spectral change in CV syllables is conserved, because each CV syllable is segmented from real speech data. Fewer spectral distance calculations are required, since the method is based on the SPLIT. Furthermore, word templates can be represented as sequences of Kana characters. Thus, addition of new words for recognition is very easy. On the other hand, DP matching cannot be executed until the endpoint is detected owing to pre-selection and adjustment of the duration time. Moreover, training speech data are needed for generating CV syllable templates.

4.3 Experimental Conditions and Results

128 CV syllable templates and phoneme-like templates were generated using 120 words from the first utterance set. Furthermore, correspondence between several phoneme-like templates and vowels was determined. Recognition experiments were executed using the following speech data:

- (1) 641 words in the first and second utterances,
- (2) Six sets of 100 words from the 641 words.

Experimental results are shown in Table 1. About 90% and 97% recognition accuracies were obtained for 641 and 100 words respectively. Degradation from using word templates generated by speech data was 6% in 641-word recognition. Most errors were caused by confusion of consonants.

Table 1 RECOGNITION RESULTS USING AUTOMATICALLY GENERATED WORD TEMPLATES.

	100 Words	641 Words
1st Utterances	98.0% (average)	91.9%
2nd Utterances	96.7% (average)	89.2%

5 Conclusion

This paper has proposed a new word recognition method, SPLIT. The efficiency of this method in speaker-dependent large vocabulary word recognition was shown through several experiments. The experimental results show that spectral information used in isolated word recognition based on DP matching can be roughly quantized in each frame without a large decrease in recognition accuracy.

A 96.3% recognition accuracy was achieved for recognition of 641 city names spoken by four male speakers using 256 phoneme-like templates. In this case, the spectral distance calculation volume and memory size for word templates were respectively 0.6% and 3% that for the conventional direct-matching meth-

od. In addition, an automatic word template generation method has been proposed for providing initial word templates, to reduce training utterances. A 90% recognition accuracy for 641 words was obtained.

Acknowledgements

The authors wish to thank Dr. Fumitada Itakura for his encouragement and suggestions. They would also like to thank Dr. Kiyohiro Shikano for his invaluable discussions and for use of his dynamic programming tools. The authors are also indebted to Kazuo Hakoda for discussions and his offer of the speech database used in the early stages of this work, and to Dr. Yoshinori Sagisaka for use of his programs on speech synthesis by rules.

References

- (1) H. Sakoe and S. Chiba: Recognition of Continuously Spoken Words Based on Time-Normalization by Dynamic Programming (In Japanese), *Trans. of ASJ*, 27, 9, p. 483, 1971.
- (2) M. Kohda, S. Hashimoto and S. Saito: Spoken Digit Mechanical Recognition System (In Japanese), *Trans. of IECE of Japan*, J55-D, 3, p. 186, 1972.
- (3) S. Makino, S. Honma and K. Kido: Spoken Word Recognition Based on Phoneme Recognition (In Japanese), *Trans. of the Committee on Speech Research*, S83-18, p. 134, June, 1983.
- (4) Y. Linde, A. Buzo and R. M. Gray: An Algorithm for Vector Quantizer Design, *IEEE Trans. COM.*, 28, 1, p. 84, 1980.
- (5) N. Sugamura and F. Itakura: Speech Data Compression by Spectral Pattern Matching (In Japanese), *Trans. of IECE of Japan*, J65-A, 8, p. 834, Aug., 1982.
- (6) N. Sugamura and S. Furui: Large Vocabulary Word Recognition Using Phoneme-Like Templates (In Japanese), *Trans. of IECE of Japan*, J65-D, 8, p. 1041, Aug. 1982.
- (7) M. Sugiyama and K. Shikano: LPC Peak Weighted Spectral Matching Measure (In Japanese), *IECE Trans.*, J64-A, p. 409, May, 1981.
- (8) A. E. Rosenberg, L. R. Rabiner, J. G. Wilpon and D. Kahn: Demisyllable-Based Isolated Word Recognition System, *IEEE Trans.*, ASSP 31, 3, p. 713, 1983.
- (9) K. Hakoda: Mechanical Production of Word Dictionary in Spoken Word Recognition (In Japanese), *Proceedings of ASJ*, p. 395, Oct. 1980.
- (10) Y. Sagisaka and Y. Tohkura: Rule of Segmental Durations Using Statistical Features of Segment (In Japanese), *Technical Reports of IECE of Japan*, EA80-66, p. 51, 1980.

The Authors



Noboru Sugamura

Research Engineer in the Intelligent Communication Terminal Section, Customer Equipment Department, NTT Integrated Communications Laboratories. He is currently working on research and development pertaining to speech recognition. After joining the ECL system in 1976, he studied speech analysis and synthesis of narrowband speech coding, especially with respect to the LSP method. For the past five years, he has studied

speech recognition using vector quantization (SPLIT).

Dr. Sugamura received his B.S., M.D. and Doctor degrees in electrical engineering from Osaka University in 1974, 1976 and 1985. He is a member of the Acoustical Society of Japan, the Institute of Electronics and Communication Engineers of Japan, and the Information Processing Society of Japan. He received the Maritorious Research Award for research on LSP speech analysis and synthesis, and the Best Paper Award for his paper "Speech Data Compression by LSP Analysis-Synthesis Technique" from IECEJ in 1982. He also received the Sato Paper Award for the paper "Isolated Word Recognition Using Strings of Phoneme-like Templates (SPLIT)" from the ASJ in 1985.



Sadaoki Furui

Head in the Fourth Section, Information Science Department, NTT Basic Research Laboratories. There, he is currently doing research on speech perception and speech recognition. After joining the ECL system in 1970, he went to work studying analysis of speaker characterization of information in speech waves, its application to speaker recognition as well as interspeaker normalization in speech recognition, and a vector-quantization-

based speech recognition algorithm.

Dr. Furui received his B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University in 1968, 1970, and 1978. He is a member of the Acoustical Society of Japan, the Institute of Electronics and Communication Engineers of Japan, and the Institute of Electrical and Electronics Engineers. He received the Yonezawa Prize for his paper "Analysis of Speaker Dependent Features in Spectral Transition in Speech Wave" from the IECEJ in 1975, and the Sato Paper Award for the paper "Isolated Word Recognition Using Strings of Phoneme-like Templates (SPLIT)" from the ASJ in 1985.

* * * *