STUDIA PHONOLOGICA XXI (1987)

Synthesis of Speaker-Adaptive Word Templates by Concatenation of the Monosyllabic Sounds

Yasuhisa NIIMI and Yutaka KOBAYASHI

Summary

This paper describes a new method for synthesizing speaker-adaptive templates in an isolated word recognition system based on the time-warping algorithm. We prepare in advance a template for each word class by averaging utterances spoken by a few talkers, and use it as a guide to excerpt the subpattern of an isolated monosyllable that is best matched against a syllable embedded in the word. The excerpted subpatterns are concatenated and smoothed at their boundaries to create speaker-adaptive word templates. The method was evaluated through the recognition of 52 Japanese city names spoken by 20 male talkers. The synthesized templates for each talker gave the average recognition rates of 99.1%. This shows that the proposed procedure is promising.

1 Introduction

This paper describes a new method for synthesizing speaker-adaptive templates in an isolated word recognition system based upon the time-warping algorithm [1]. This algorithm has enjoyed widespread popularity and achieved a great success in speaker-dependent word recognition systems in which the reference templates are custom-made for a speaker. With a large vocabulary system, it is, however, inconvenient for a new speaker to register his own references for all the words in the vocabulary, because the time reauired can not be allowed practically. It is, therefore, desirable to develop a speaker-independent system, or a speaker-adaptive system.

In order to cope with interspeaker varieties of utterances speaker-independent systems have used, for a word class, multiple templates created using clustering techniques [2], or a speaker-independent template produced through averaging a number of utterances [3]. A recent progress in this area is to model an acoustic manifestation of a word in a hidden Markov process [4]. In this approach,

Yasuhia NIIMI (新美康永): Associate Professor, Department of Computer Science, Kyoto Institute of Technology.
Yutaka KOBAYASHI (小林 豊): Assistant, Department of Computer Science, Kyoto Institute of Technology.
The authors have done the research under the direction of Dr. Shuji Doshita, Professor of Information Science, Kyoto University.

a stochastic automaton is used as a reference (for a word class), and its state transition probabilities and emission probabilities are estimated through a training process. All of these attempts require a great number of training samples for a word class.

A goal of speaker-adaptive systems is to adapt word templates to a new speaker using his training utterances. In several attempts a word is described as a string of smaller linguistic units than a word—such as demisyllables [5] and phonemes—and a word template is created by concatenation of the templates for those units which are extracted from training utterances of a new speaker. In another approach a speaker-adaptive template is obtained by modification of a speaker-independent template with speaker-specific feature vectors [6].

In this paper we propose a method for synthesizing speaker-adaptive templates by concatenation of monosyllabic sounds. Given a word class w, which is assumed to be in the form of $C_1V_1C_2V_2$ (C's and V's mean consonants and vowels respectively), a speaker-adaptive template is created in the following four steps.

1) A template for the word class w is prepared in advance by averaging utterances spoken by a few speakers. It is called a 'guideline template'.

2) The guideline template for w is divided into the four small segments C_1V_1-V_1C_2-C_2V_2-$(C_2)V_2$. These segments are called 'CV-segments'.

3) The matching is performed between a CV-segment and the corresponding monosyllabic sound spoken in isolation. The segment in the form of VC is matched against the time-reversed pattern of the CV sound. The subpattern of the monosyllabic sound best matched against the CV-segment is excerpted as a unit to synthesize the word template.

4) The excerpted units are concatenated and smoothed at their boundaries. Then the concatenated pattern and corresponding guideline template are averaged to create the speaker-adaptive word template for the class w.

In the purpose of the time-normalization the guideline templates and monosyllabic sounds are resampled before the steps (2), (3) and (4) so that distances become equal between two successive new samples.

The method outlined above looks similar to the one proposed by L. R. Rabiner et al. [5]. In their method a set of source words which include demisyllables and boundary locations of demisyllables within a source word are given. A new talker spoke these source words and the reference patterns for his demisyllables are extracted based upon the non-linear time-warping technique. The number of source words which a talker must speak would increase with a larger vocabulary system. On the other hand, although our method requires a guideline template for each word in the vocabulary, a new talker has only to speak at most 100 monosyllables.

2 An Overview of the Word Recognition System

Figure 1 depicts an overview of the recognition system of isolated words in which the proposed procedure for the speaker adaptation works. An incoming

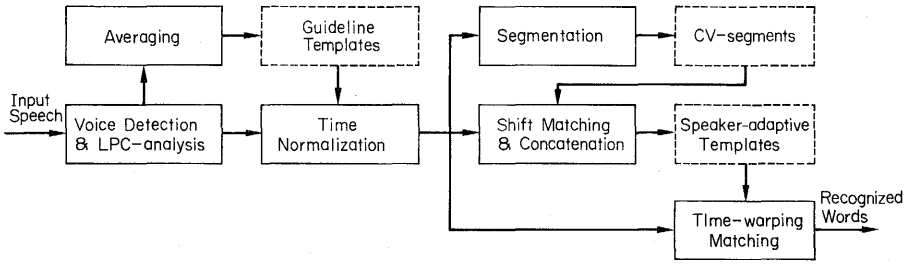

Fig. 1. An overview of the speaker-adaptive word recognition system

signal is pre-emphasized, low-pass filtered and digitized at 10 kHz with an accuracy of 12 bits. Then the energy and zero crossing rate of the signal are calculated every 10 ms for the automatic end-point detection. The interval of the speech signal is subjected to the 14-th order LPC-analysis with Hamming window of 25.6 ms to produce 20 cepstral coefficients every 10 ms.

A set of cepstral coefficients can be represented as a point in the 20 dimensional feature space. Thus the sequence of feature vectors can be viewed as a trace of sample points in the feature space. Stationary parts of the speech signal cause a high point density along this trace, while rapid spectral changes lead to sample points that are spaced far apart on this trace. For the purpose of the time-normalization for subsequent pattern matching, we resample the trace so that distances become equal between two successive new sample points. In the following we shall refer to this method as 'trace segmentation', a new sample point as a 'frame' and a distance between two successive frames as a 'distance quantization unit'.

The system illustrated in Fig. 1 works in three phases. The first phase prepares in advance a guideline template for each word class. The designing utterances collected from several talkers, after subjected to LPC-analysis, are fed into the averaging component to produce guideline templates. These templates are then subjected to the trace segmentation in which a distance quantization unit is determined so that the averaged number of sample points for all the words in the vocabulary does not change before and after the time-normalization. Using a set of rules as to where the segmentations would be made, we visually inspect spectral representation of the guideline templates and establish the locations of the CV-segments. The rules for segmentation will be explained in the section 4.

The second phase builds up a set of speaker-adaptive templates from the training utterances of a new speaker who intends to adjust the word recognition system to himself. Monosyllables spoken in isolation, after subjected to the LPC-analysis and the trace segmentation, are fed into the shift matching component.

In this component we scan a CV-segment across a time-normalized monosyllable, sum up distances between feature vectors of the CV-segment and those of the monosyllable, and find where the best match is attained, in other words, summation of the distances is minimized. The linear matching, instead of the time-

warping matching, is used in this component because a CV-segment and a monosyllable are both time-normalized through the trace segmentation. The rules as to which monosyllable would be associated with a CV-segment will be described in the section 4.

The subpattern of the monosyllable which has been best matched against the CV-segment is excerpted as a unit to synthesize a speaker-adaptive template. These excerpted monosyllables, referred to as 'monosyllabic CV-segments', are then concatenated according to a given lexical specification of each word in terms of CV-segments. Finally the concatenated word pattern and corresponding guideline template are averaged with a ratio of μ: $(1-\mu)$ $(0 \leqq \mu \leqq 1)$ to create the speaker-adaptive word template.

The third phase recognizes unknown words of a new speaker who has adapted the system to himself. An incoming speech signal is input through the LPC-analysis and the trace segmentation component to the time-warping component which matches the input utterance against speaker-adaptive references and chooses one of possible word classes based on the minimum distance criterion.

3 A Sequential Method for Averaging Sequence of Feature Vectors

This section describes a sequential method for averaging multiple sequences of vectors. It is based upon the method for making a new averaged sequence from two sequences of vectors differing in their length. We reported its generalized version in [3] and present the essential part of the method here. Let $A=a(1)a(2)\ldots a(T_a)$ and $B=b(1)b(2)\ldots b(T_b)$ be the two sequences to be merged, and $g(\ )$ be the time-warping function resulting when the match between A and B has been evaluated by the time-warping algorithm. The following steps (1)–(3) and a weighting factor w form a new sequence of vectors, $C=c(1)c(2)\ldots c(T_c)$.

1) T_c is the largest integer not greater than $wT_a+(1-w)T_b+0.5$,

2) Given an integer $k(1 \leqq k \leqq T_c)$, we can determine the two smallest integers i and j, and the two largest integers i' and j' subjecting to the following conditions;

$$j=g(i) \text{ and } j'=g(i'), \tag{a}$$

$$k-0.5 \leqq wi+(1-w)j \text{ and } k+0.5>wi'+(1-w)j'. \tag{b}$$

3) From a subsequence of A, $a(i)a(i+1)\ldots a(i')$ and a subsequence of B, $b(j)b(j+1)\ldots b(j')$, a new component vector $c(k)$ of C is calculated as follows;

$$c(k)=\frac{w}{i'-i+1}[a(i)+a(i+1)+\ldots+a(i')] + \frac{1-w}{j'-j+1}[b(j)+b(j+1)+\ldots+b(j')].$$

This averaging operation is denoted by

$$C=M(A, B, w).$$

Let A_n $(n=1, 2, \ldots, N)$ denote N sequences of vectors to be averaged. The following

formulae define the averaged sequence of vectors A.

$$A(1)=A_1,$$
$$A(k)=M(A(k-1), A_k, (k-1)/k), \qquad (k=2, 3, \ldots, N),$$
$$A \quad =A(N).$$

For each word class, we make a guideline template by applying this averaging method to a set of utterances collected from several speakers.

4 Rules for Segmentation and for Concatenation

4.1 Rules for segmentation

We shall describe rules as to where the guideline templates would be segmented. The structure of Japanese syllables are classified into the following eight patterns: CV, CVV, CVN, CVQ, CSV, CSVV, CSVN and CSVQ. C, V, N and S mean a consonant, a vowel, a moraic nasal and a semivowel, respectively. The symbol Q designates the phoneme specific to Japanese, called 'sokuon (moraic silence)', which is acoustically realized as a rapid transition followed by a rather long closure when it precedes a plosive, and assimilates the following consonant when it precedes a fricative. In each of the eight patterns the first consonant may be omitted.

Those syllables are divided as follows:

$$\mathrm{C(S)V} \rightarrow \mathrm{C(S)V-V+}, \qquad \mathrm{C(S)VV} \rightarrow \mathrm{C(S)V-V+},$$
$$\mathrm{C(S)VN} \rightarrow \mathrm{C(S)V-VN-N*}, \quad \mathrm{C(S)VQ} \rightarrow \mathrm{C(S)V-VQ*}.$$

The symbol '—' means that phonemes 'V' and 'N' are segmented at the center of their sustained part, and the symbol '+' means that if the syllable under consideration precedes a sustained consonant, the latter half portion of a vowel combines the following consonant which will be cut at its center, and that if the syllable precedes a plosive, the segmentation is made just before the plosive. The symbol '*' indicates that whatever a consonant follows the syllable, the segmentation is made just before the consonant.

4.2 Rules for concatenation

Japanese syllables in the form of 'C(S)V' are called monosyllables. A set of monosyllables does not include all the CV-segments produced according to the rules for segmentation. So we have established the following rules as to which monosyllable would be associated with a CV-segment in the shift matching.

1) A monosyllable 'C(S)V' is associated with a CV-segment in the form of 'C(S)V'.

2) Time-reversed patterns of monosyllables are used for those CV-segments like 'VC', 'VN' and 'VQ*(C)' which are not included in the inventory of monosyllables. That is, a monosyllable 'CV' is associated with 'VC', 'nV' or 'mV' with 'VN' and 'CV' with 'VQ*(C)'.

3) If a syllable is followed by a single vowel, we have a CV-segment in the form of 'V_1V_2' with which no monosyllable can be associated. In this case a segment for concatenation is derived from the CV-segment 'V_1V_2' in the same way

as we used in [6].

We should add some comments for these rules. The rule 2) is established because the last half part of a vowel in a word is more affected by the following consonant than by the preceding one. The on-glide from a vowel to a consonant differs physically from the off-glide from the latter to the former, but it is reported [7] that both the glides are perceptually similar to each other. However there is an exception to this rule. If a consonant is either dental or alveolar and a vowel is either /i/ or /u/, the consonant is palatalized. Therefore, a time-reversed pattern of a segment 'CV' would be remarkably different from a segment 'VC'. If a vowel (/i/ or /u/) is bounded by two unvoiced consonants, the vowel is mostly devocalized and its voiced part vanishes. In such a case it is difficult to synthesize a word template only by the rules stated.

5 Recognition Experiments

The proposed method for synthesizing speaker-adaptive references was evaluated through an application to the recognition of the vocabulary of 52 Japanese city names. The syllabic structures of these words are so simple that the semivowel /j/ is not included in a word. Thus we have needed only 68 simple monosyllables in the form of 'CV' to synthesize the templates.

In order to provide training and test utterances, 20 male talkers (referred to as test talkers) spoke the words in the vocabulary twice and 68 simple monosyllables once in a sound booth. Thus we have two sets of the word utterances for each talker. We used the first set of spoken words and the set of monosyllables to adapt the word recognition system, and the second set to test it. In addition, five separate talkers (referred to as design talkers) spoke the words in the vocabulary once, which we used to prepare the speaker-independent guideline templates.

We conducted three word recognition experiments; speaker-adaptive, speaker-independent and speaker-dependent ones. The last two were performed for comparison. In the first experiment we synthesized several sets of speaker-adaptive word templates changing the values of μ from 0 to 1. In the case of $\mu=1$ the averaging operation is not made between a concatenated word pattern and its corresponding guideline template, and in the case of $\mu=0$ a resulting word pattern is a guideline template itself. In the second experiment we used as the reference patterns the speaker-independent guideline templates created in the first phase of the word recognition system. In the third experiment we newly produced a set of speaker-dependent templates for each test talker by the time-normalization of his first set of spoken words. For each talker, his second set of spoken words were used to test the word recognition system throughout these three experiments. A distance quantization unit, that is, a distance between two successive samples after the time-normalization, is also commonly selected to be 0.3 for the guideline templates and 0.6 for spoken words and monosyllables. These values were de-

Table 1. Experimental Results.

experiments		rates (%)
speaker-adaptive		
$\mu=$	0.3	99.0
	0.4	99.1
	0.5	99.0
	0.7	98.9
	1.0	98.1
speaker-independent		97.5
speaker-dependent		99.4

termined through a preliminary experiment. Since the guideline templates are created by the averaging operation, their temporal change is smoother than that of natural utterances. The smaller value would, therefore, be selected for the guideline templates.

6 Experimental Results and Discussions

Table 1 shows the results of the experiments, which are means calculated for twenty test talkers. The best word recognition rate was 99.1% for $\mu=0.4$ and the worst was 98.0% for $\mu=1.0$ in the speaker-adaptive experiment, while the rate was 99.4% and 97.5% in the speaker-dependent and speaker-independent ones respectively.

The statistical F-test showed that the hypothesis that there be no difference between the rate in the best speaker-adaptive case and the one in the speaker-independent case could be rejected with the significance level of 0.001. On the other hand no difference is observed between the rate in the worst case and the one in the speaker-independent case. Furthermore, as indicated in Table 1, the recognition rates are improved in a wide range of μ by averaging. This concludes that the proposed method is effective and stable for the speaker-adaptation.

7 Conclusion

The method for the speaker-adaptation in an isolated word recognition system has been described and evaluated through the application to the recognition of the vocabulary of 52 Japanese city names. The experiments conducted based on the this procedure have shown that it is effective for the speaker adaptation. However, the vocabulary used in this study is small and does not contain words including syllables in the form of CSV. It should be tested for a larger vocabulary containing the words with the syllables in this type in the future work.

REFERENCES

1) Sakoe, H. and Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans., Acoust., Speech, Signal Processing, Vol. ASSP-26, Feb. 1978, pp. 43–49.
2) Rabiner, L. R., Levinson, S. E. and Rosenberg, A. E., "Speaker-Independent Recognition of Isolated Word Using Clustering Techniques," IEEE Trans., Acoust., Speech, Signal Processing, Vol. ASSP-27, Aug. 1979, pp. 336–349.
3) Niimi, Y., "A Method for Forming Universal Reference Patterns in an Isolated Word Recognition System," Proc. of the 4-th Int. Joint Conf. on Pattern Recognition, 1978, pp. 1022–1024.
4) Rabiner, L. R., Levinson, S. E. and Sondi, M. M., "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition," The Bell Syst. Tech. J., Vol. 62, 1983, pp. 1035–1074.
5) Rabiner, L. R., Rosenberg, A. E., Wilpon, J. G. and Zampini, "A Bootstrapping Training Technique for Obtaining Demisyllable Reference Patterns," J. Acoust. Soc. Am., Vol. 71, 1982, pp. 1588–1595.
6) Niimi, Y., Kitamura, N. and Kobayashi, Y., "Speaker Adaptation in an Isolated Word Recognition System," Proc. of the 2nd WESTPAC, 1985, pp. 366–371.
7) Agui, T. and Hosomura, T., "The Study on the Reversibility of Phoneme Dyads," J. Acoust. Soc. Jpn., Vol. 31, 1975, pp. 521–528.

(Aug. 31, 1987, received)