STUDIA PHONOLOGICA XIX (1985)

Speaker Adaptation in an Isolated Word Recognition System

Yasuhisa NIIMI, Norio KITAMURA and Yutaka KOBAYASHI

Summary

A speaker-dependent isolated word recognition system can attain a high recognition rate. It is, however, not easy to train such a system with a large vocabulary, and is necessary to adapt the system to a specific speaker with a small set of his training samples. We present a new method for the speaker adaptation in an isolated word recognition system and give some preliminary experimental results of its application. The word recognition system is based upon the time-warping algorithm. The reference patterns used in the matching process are adapted to a new speaker in the following two steps.

(1) The utterances of a word class spoken by multiple speakers are averaged to create a common reference of the class.

(2) This common refernece is transformed so that its base points (usually selected from characteristic parts of vowels and consonants) coincide with speaker-dependent spectra extracted from the training utterances of the speaker.

Some experiments were conducted to recognize 65 Japanese city names spoken by 20 adult male speakers. The common references were made from the utterances spoken by five seperate speakers. The recognition rate was 96.6% for the speaker-adapted references, while it was 96.4% for the speaker-dependent references and 95.7% for the common references. These results show that the proposed method is effective for the speaker adaptation.

1. Introduction

This paper describes a new method for the speaker adaptation 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-

Yasuhisa NIIMI (新美康永): Associate Professor, Department of Computer Science, Kyoto Institute of Technology.

Norio KITAMURA (北村典生): Graduate Student, 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.

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 required can be prohibitive. In addition it is not feasible to store sets of references for a large number of potential users. It is, therefore, desirable to develop a speaker-indiependent system, or a speaker-adaptive system which can adapt itself to a new speaker very rapidly.

A number of approaches have been made to the development of a speaker-independent system. One of such approaches is to employ for a word class multiple templates created using clustering techniques[2], or a speaker-independent templates produced through averaging a number of utterances[3]. A recent progress in this context is to model an acoustic manifestation of a word in a hidden Markov process[4]. All of these approaches require a great number of training samples for a word class.

A representative approach to the development of a speaker-adaptive system is to create the word templates by concatenation of demisyllables[5] or phoneme-like symbols. The demisyallables or the templates for the phoneme-like symbols are extracted from a set of training utterances of a specific speaker.

The investigation described in this paper is in the second direction and concerns how to adapt to a new speaker the references used in the isolated word recognition system based upon the time-warping algorithm. Our fundamental idea is that the static properties of a speech signal have more amount of information on a speaker than the dynamic properties. In other words, we consider that the spectral characteristics in stationary parts of the speech signal provide sufficient information to adapt the references to a speaker. According to this idea we create a set of speaker-adaptive references in the following two steps.

(1) The utterances spoken by multiple speakers are averaged to create a common (speaker-independent) reference for a word class, and a characteristic point is manually located on the stationary part of each phoneme in the reference. It is called a *base point*.

(2) Corresponding to these base points, speaker-dependent spectra (called *target vectors* hereafter) are extracted from the training uttereances of a speaker to whom the word recognition system is intended to adapt itself. A speaker-adaptive reference of a word class is formed by the linear transformation of the common reference so that its base points coincide with the corresponding speaker-dependent target vectors.

Some experiments were conducted to recognize 65 Japanese city names spoken by 20 adult male speakers. The common references were made from the utterances spoken by five seperate speakers. The recognition rate was 96.6% for the speaker-adaptive references, while it was 96.4% for the speaker-dependent references and 95.7% for the common references. These results show that the proposed method is effective for the speaker adaptation.

Section 2 gives a brief description of the isolated word recognition system in which the proposed method for the speaker adpatation works. Section 3 discusses a sequential method for averaging vector sequences differing in their length. This method is used to create a set of common references. Section 4 discusses the method for adapting the common references to a new speaker. This is the central part of what we intend to report in this paper. Section 5 describes the preliminary experiments conducted to test the speaker adaptation procedure and some discussions on the obtained results.

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 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. Following the automatic end point detection logic that captures the interval containing the word using those parameters, 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. Thus the input word is represented by a sequence of 20-dimensional vectors.

The speech recognition system works in the three modes: making the common references, adapting them to a new speaker and recognizing the utterances of the speaker. In the first mode, the output of the LPC-analyzer is switched to #1 in Figure 1, and the designing utterances collected from several speakers are fed into the averaging component to produce the common references on which the base points are marked. In the second mode, an analyzed output is switched to #2 and the information on the target vectors concerning a new speaker is fed into the speaker adaptation component to create the references adjusted to the speaker. In the third mode, an incoming speech signal is directly input through the switch #3 to the time-warping component which matches the input utterance against several references and chooses one of possible word class based upon the

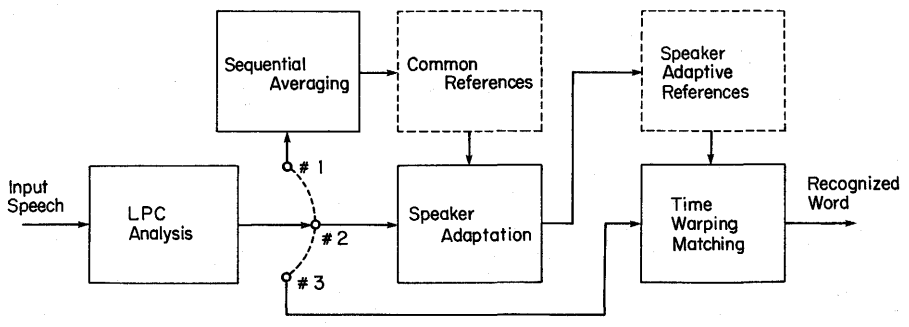

Fig. 1. An overview of the speaker-adaptive word recognition system.

minimum distance criterion.

3. A Sequential Method for Averaging the Sequences of 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 $A=a(1)a(2)...a(T_a)$ and $B=b(1)b(2)...b(T_b)$ 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, $C=c(1)c(2)...c(T_c)$.

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

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

(a) $j=g(i)$ and $j'=g(i')$,

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

Figure 2 depicts the geometrical relation among these conditions.

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

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

This averaging operation is denoted by

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

Let A_n $(n=1,2,...,N)$ denote N sequences of vectors to be averaged. The following formulae define the averaged sequence of vectors A.

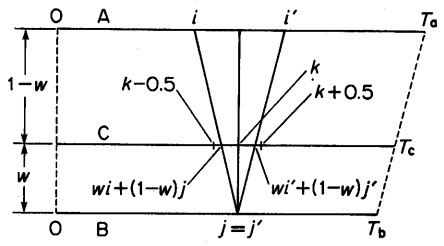

Fig. 2. The geometrical relation among the parameters i, i', j, j', k and w which determine a new component vector $c(k)$ of C.

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

4. SEPAKER ADAPTATION OF A COMMON REFERENCE

Figure 3 schematically shows the process to adapt a common reference to a specific speaker. For the sake of an easy explanation all the speech patterns are represented as one dimensional patterns in the figure. Let $c(t)$ (curve I) and $a(t)$ (curve II) denote respectively the common reference of a word class and its speaker-adaptive version. Given the base points T_i's ($i=1,2,...,N$) on $c(t)$ and the target vectors a_i's ($i=1,2,...,N$), the common reference $c(t)$ is transformed so as to pass through these points at the times T_i's, and the values of $a(t)$ can be calculated in the interval between T_i and T_{i+1} by adding to $c(t)$ the displacement vector $\Delta(t)$ produced by the linear interpolation of the two difference vectors $a_i-c(T_i)$ and $a_{i+1}-c(T_{i+1})$. That is,

$$a(t)=c(t)+\Delta(t),$$

where

$$\Delta(t)=[\Delta_i(T_{i+1}-t)+\Delta_{i+1}(t-T_i)]/(T_{i+1}-T_i),$$
$$\Delta_i=a_i-c(T_i),$$
$$\Delta_{i+1}=a_{i+1}-c(T_{i+1}).$$

We will describe in the next section how to select the base points of the word and how to extract the target vectors from the training utterances of the speaker to whom the common reference is intended to adapt itself.

5 RECOGNITION EXPERIMENTS AND RESULTS

The proposed method for the speaker adaptation was evaluated through an

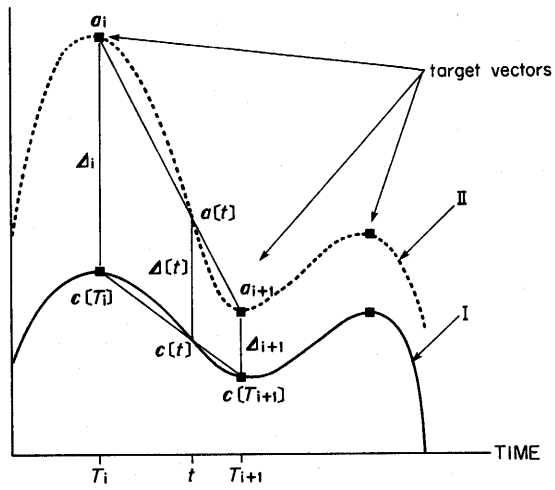

Fig. 3. Speaker adaptation of a common refeernce. Curves I and II indicate a common reference and an adapted reference respectively.

application to the recognition of the vocabulary of 65 Japanese city names. It was used to make a set of the speaker-adaptive references for a speaker. In order to provide training and test utterances, each of twenty five male speakers spoke 65 Japanese city names twice in a sound booth. Thus there are two sets of the utterances for each speaker.

The two experiments were conducted; the first was designed to investigate the effectiveness of the proposed method for the speaker adaptation, and the second to examine the effect of the location and the number of the base points on the speaker adaptation.

In the first experiment were compared the three types of references; speaker-dependent, speaker-independent and speaker-adaptive ones. For each speaker his first set of utterances was used as his first type of references. The set of common references made by the averaging method described in Section 3 was adopted as the second type of references. The first utterance sets produced by the five of twenty five speakers contributed to this averaging process. These five speakers were selected because they spoke the words clearly and carefully. The third type of references was derived from the common references by modifying their base points with speaker specific target vectors after the method mentioned in Section 4.

In this preliminary experiment carried out to evaluate the potential effectiveness of the proposed method, the speaker-dependent target vectors were extracted from the first set of utterances spoken by each speaker because the problem of coarticulation should be avoided. The base points were manually located on the sustained part of each phoneme for a common reference by visual inspections of its spectral representation. Then the time-warping algorithm was performed to match between a training utterance and a common reference belonging to the same word class. The vector in the training utterance related to each base point through this matching was selected as the corresponding target vector. The information on those base points and their target vetors are sufficient to complete the speaker-adaptive references.

In the first experiment a base point was selected on each phoneme included in a word, and all the base points contributed to the modification of a common reference. On the other hand, the second experiment was performed to examine the effect of the location and the number of the base points on a word. So the target vectors were extracted corresponding to the following four sets of the base points: (A) all the base points, (B) the ones on vowels and voiced consonants, (C) the ones on vowels and unvoiced fricative (affricate) cosonants, and (D) the ones only on vowels. In each case the silent parts before the explosion were adopted as base points and in the three cases except (A) the excluded base points were not modified (fixed as they were) in the transformation of the common references.

Figure 4 shows the results of the first experiment conducted using, as the

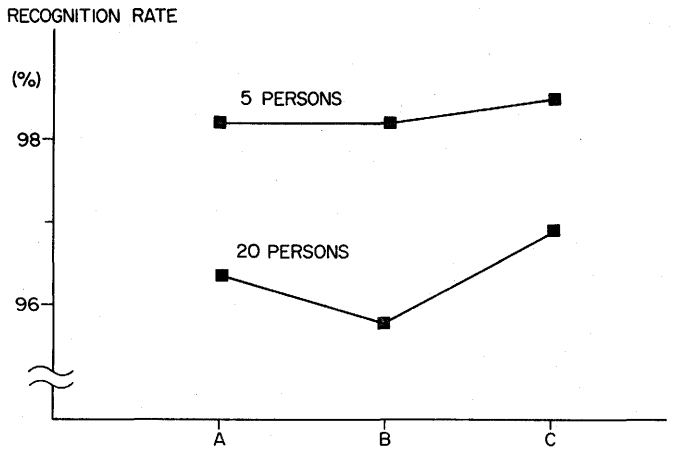

Fig. 4. Recognition rates for three types of references: (A) speaker-dependent, (B) speaker-independent and (C) speaker-adaptive ones.

test utterances, the second set of utterances produced by twenty speakers whose utterances were not included in constructing the common references. The averaged recognition rate was 96.9% for the speaker-adaptive references, while it was 96.4% for the speaker-dependent references and 95.7% for the speaker-independent references. The statistical F-test showed that the hypothesis that there be no difference between the rates for the speaker-adaptive references and for the speaker-independent ones could be rejected with the significance level of 0.05. This concludes that the proposed method is effective for the speaker adaptation.

Some comments should be made for the results indicated in Figure 4. It is expected that the speaker-dependent references whould give a higher recognition rate than the speaker-adaptive ones. The experiment, however, resulted in a reverse relation. Although there was no significant defference between these recognition rates, the two reasons for this could be considered.

(1) The speaker dependent references employed in this experiment were not of high quality because only an utterance was used for each word class instead of averaging a few utterances, and

(2) On the other hand, the smoothing effect in creating the common references might be of benefit to absorbing the intra-speaker variability of the utterances.

But we consider that more detail analysis of the sources of recognition errors is necessary to reach a definitive conclusion.

Figure 4 also shows the results of the recognition for the second sets of utterances of the five speakers whose first sets of utterances were used for making the common references. For all the three types of references their avaraged recognition rates were higher by about 2% than those of the remaining speakers. This is due to the fact that their utterances were selected for making the common re-

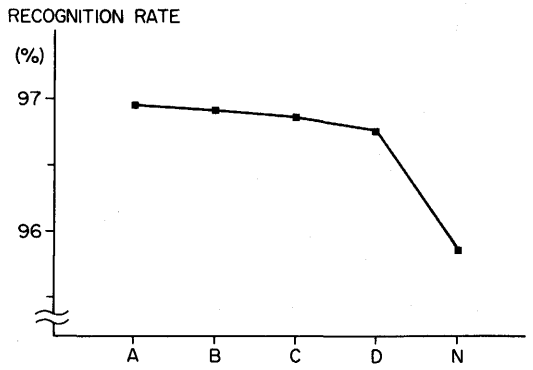

Fig. 5. The relation of the sets of base points to be modified to the recognition rates. A, B, C and D are explained in the text. N is the rate for speaker-independent references.

ferences because they spoke clearly and carefully.

Figure 5 shows the average recognition rates obtained in the second experiment which was conducted to examine the effect of the location and the number of the base points on the speaker adaptability. In this experiment, the test utterances were also the second set of utterances of the remaining 20 speakers who did not contribute to the formation of the common references. For comparison, Figure 5 includes the average recognition rate for the speaker-independent references. The figure shows that a decrease on the number of the base points used to modify a common reference leads to a gradual degradation of the recognition rates, but in comparison of these rates with the one in the speaker-independent case, it can be inferred that the adaptation in the vowel parts brings an essential improvement of the recognition rates.

6. 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 65 Japanese city names. The experiments conducted based upon this procedure have shown that it is effective for the speaker adaptation, but can be considered as the first pass in the direction of developing a speaker-adaptive speech recognition system. In order to avoid the problem of coarticulation, the same vocabulary as the one to be recognized has been used to train the system in the experiments. Our final goal is to employ a seperate small set of the utterances, for example, the set of monosyllabic sounds. We are currently continuing the study in this direction.

Referenoes

1) Sakoe, H. and Chiba, S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans., Acoust., Speech, and 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, and 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., "A Bootstrapping Training Technique for Obtaining Demisyllable Reference Patterns," J. Acoust. Soc. Am., Vol. 71, 1982, pp. 1588–1595.

(Aug. 31, 1985, received)