Isolated Mandarin syllable recognition using segmental features

S. Chang
S.-H. Chen

Indexing terms: Speech recognition, Acoustic segments, Mandarin base syllables

Abstract: A segment-based speech recognition scheme is proposed. The basic idea is to model explicitly the correlation among successive frames of speech signals by using features representing contours of spectral parameters. The speech signal of an utterance is regarded as a template formed by directly concatenating a sequence of acoustic segments. Each constituent acoustic segment is of variable length in nature and represented by a fixed dimensional feature vector formed by coefficients of discrete orthonormal polynomial expansions for approximating its spectral parameter contours. In the training, an automatic algorithm is proposed to generate several segment-based reference templates for each syllable class. In the testing, a frame-based dynamic programming procedure is employed to calculate the matching score of comparing the test utterance with each reference template. Performance of the proposed scheme was examined by simulations on multi-speaker speech recognition for 408 highly confusing isolated Mandarin base-syllables. A recognition rate of 81.1% was achieved for the case using 5-segment, 8-reference template models with cepstral and delta–cepstral coefficients as recognition features. It is 4.5% higher than that of a well-modelled 12-state, 5-mixture CHMM method using cepstral, delta cepstral, and delta-delta cepstral coefficients.

1 Introduction

Many acoustic properties of speech signal have been considered in the past for speech recognition in order to achieve high recognition rate. They include short-time stationarity, voicing/unvoicing, temporal correlation, coarticulation, etc. Among them, the property of strong correlation over time is an important but seldom used property. In most developed speech recognition systems, a sequence of feature vectors, derived frame-by-frame by using LPC-based or filterbank analysis techniques, is directly used to represent the input speech signal. The correlation among contiguous frames of speech signal is nearly ignored. For instance, the traditional hidden Markov model (HMM) [1], which assumes that the observed feature vectors within a state are locally IID (independently and identically distributed), it is known to be weak in capturing such a property. It is, therefore, desirable to improve the performance of a speech recognition system by additionally incorporating the temporal correlation information.

Some efforts had been done on this problem in the past few years. A modified HMM which took the correlation among feature vectors of several successive frames into account by defining a new emission probability was proposed in Reference 2. In References 3 and 4, linear predictive HMM was proposed to relax the IID assumption of the traditional HMM by introducing the correlation among contiguous frames into the modelling of the state observation probability. A layered neural network was employed in Reference 5 to implement speech frame prediction for taking advantages of long-term temporal correlation of speech signal. All the above methods handled the problem by modifying the structures of models describing the input speech signal. An alternative way of studying this problem is to define a new representation of speech signal to incorporate the spectral correlation over time. A fixed-length representation of a variable-length observed segment based on a resampling transformation was used in Reference 6 to capture the spectral/temporal structure of speech signal over the time interval of a phone.

In this paper, a segment-based speech recognition scheme is proposed. Instead of taking frames as the basic processing units done in most conventional methods, this approach employs acoustic segments as basic processing units. Speech signal of an utterance is then regarded as a template formed by concatenating a sequence of acoustic segments. Each segment roughly represents an acoustic event like a state of HMM and is characterised by a variable-length vector sequence of spectral features. A fixed dimensional feature vector, referred to as 'segmental feature vector', is then extracted from each segment for capturing the temporal correlations of spectral parameters in the segment and is taken as the recognition feature of the system. The utterance is therefore modelled as a segment-based template characterised by a sequence of segmental feature vectors. In the training, an automatic training algorithm is employed to generate several segment-based reference templates for each class. In the testing, a dynamic programming procedure is employed to calculate the matching score of comparing the test utterance with each segment-based reference template. Comparing with all conventional frame-based speech

This work has been supported by Telecommunication Laboratories and in part by the National Science, Council, Taiwan, Republic of China.

Paper 1648K (E5), first received 21st June 1993 and in revised form 24th October 1994

The authors are with the Department of Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

recognition techniques, the proposed approach possesses a distinctive property that the characteristics of strong temporal correlation in speech signal is explored in feature extraction and implicitly embedded in the recognition features. We therefore expect the system to perform better on speech recognition.

The performance of the proposed scheme was examined by simulations on multispeaker speech recognition for 408 highly confusing isolated Mandarin base-syllables. Since there are no common databases for Mandarin syllables recognition, it is difficult to compare our results with other works. Actually, in the past few years, many researches worked on isolated Mandarin base-syllables recognition were devoted to speaker-dependent cases. We shall therefore discuss some major achievements in those studies. In Reference 15, HMMs trained by a special direct-concatenation approach were used for syllable recognition. The average recognition rate for two male speakers was 90.69%. In Reference 16, a discriminative training procedure based on the generalised probabilistic descent (GPD) algorithm was applied to train the subsyllable HMM models of Mandarin initials and finals. A recognition rate of 93.3% was achieved for the case of recognising utterances of a female speaker. A hierarchical neural network recogniser based on a C/V segmentation algorithm was explored in Reference 17. A recognition rate of 90.14% was obtained. In Reference 18, a one-class-one-net neural network with modified selective update algorithm was proposed. The recognition rate was 85%.

2 Proposed segment-based speech recognition scheme

The block diagram of the proposed speech recognition scheme is shown in Fig. 1. It operates on two phases: the training phase (see Fig. 1*a*) and the recognition phase (see Fig. 1*b*). Speech signal is first preprocessed to extract a spectral feature vector for each frame. Then, in the training phase, a two-stage algorithm is employed to generate some segment-based reference templates for each syllable using a large training set. Each segment-based reference template is expressed in the form of a sequence of segmental feature vectors. In the first stage of the training algorithm, initial reference templates are generated by first dividing all training utterances into a predetermined number of segments using a minimum distortion (MD) segmentation algorithm [7, 8] and then finding several representative templates for each class by using a modified k-means clustering algorithm [9]. In the second stage, a two-step iterative procedure called a joint segmentation/clustering algorithm, which alternatively performs partition/resegmentation and template-updating, is employed to refine all reference templates. In the recognition phase, a segment-based template matching is first performed to calculate the matching score of comparing the input testing utterance with each reference template. It is a joint optimisation procedure which simultaneously extracts segmental feature vectors, determines segment boundaries, and calculates the matching score. Then, a final decision is made by simply choosing the syllable associated with the reference template with best matching score as the recognised syllable. Several key issues of the proposed approach are detailed as follows.

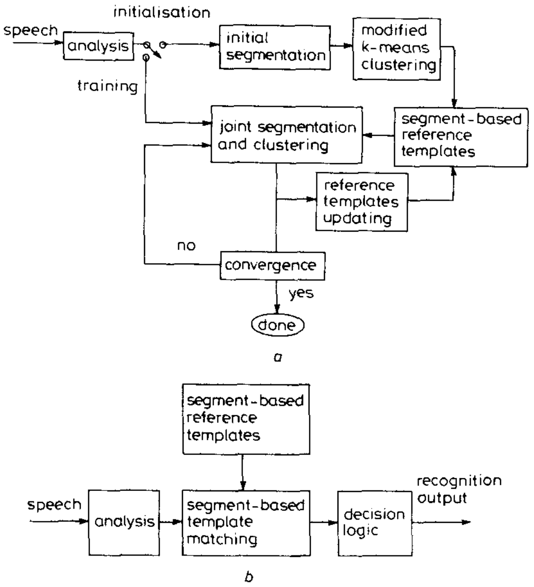
initialisation
speech
analysis
training
initial segmentation
modified k-means clustering
joint segmentation and clustering
segment-based reference templates
reference templates updating
no
convergence
yes
done
a
segment-based reference templates
speech
analysis
segment-based template matching
decision logic
recognition output
b

Fig. 1 *Block diagram of the segment-based speech recognition scheme*
a Training phase
b Recognition phase

2.1 Segmental feature vector

A feature vector of K spectral parameters is calculated for each frame by speech analysis. The speech signal is then characterised by the resulting sequence of feature vectors. In the proposed scheme, the feature vector sequence of an utterance is first divided into a predetermined number of variable-length segments. A fixed dimensional feature vector is then extracted from each segment. The new feature vector is simply referred to as 'segmental feature vector' for being distinguished from the conventional frame-based feature vector. In our implementation, a segmental feature vector is composed of K sets of coefficients of discrete orthonormal polynomial expansions. To be more specific, the sequence of each feature parameter in a given segment is regarded as a feature contour. Each feature contour is then represented by several low-order coefficients of a discrete orthonormal polynomial expansion. Due to the fact that the segment roughly represents a stable acoustic event like a state of HMM, all its feature contours are expected to be smooth enough for being well-approximated by the reconstructed contours obtained by orthonormal polynomial expansions using coefficients up to the second order. This contour approximation method was demonstrated to be suitable for a stable acoustic event in our previous work [11]. The basis functions of the orthonormal polynomial expansion are normalised, in length, to [0, 1] and expressed as

$$\Phi_0\left(\frac{i}{N}\right) = 1$$

$$\Phi_1\left(\frac{i}{N}\right) = \left[\frac{12N}{N+2}\right]^{1/2}\left[\left(\frac{i}{N}\right) - \frac{1}{2}\right]$$

$$\Phi_2\left(\frac{i}{N}\right) = \left[\frac{180N^2}{(N-1)(N+2)(N+3)}\right]^{1/2} \times \left[\left(\frac{i}{N}\right)^2 - \left(\frac{i}{N}\right) + \frac{N-1}{6N}\right] \quad (1)$$

for $i = 0, 1, \ldots, N$, where $N+1$ is the length of the segment. A feature contour $F_k(i/N)$ can then be approximated by the smoothed reconstructed contur

$$\hat{F}_k\left(\frac{i}{N}\right) = \sum_{j=0}^{2} a_{k,j} \times \Phi_j\left(\frac{i}{N}\right) \quad \text{for } 0 \leqslant i \leqslant N \tag{2}$$

where

$$a_{k,j} = \frac{1}{N+1} \sum_{i=0}^{N} F_k\left(\frac{i}{N}\right) \times \Phi_j\left(\frac{i}{N}\right) \quad \text{for } j = 0, 1, 2 \tag{3}$$

are coefficients of orthonormal polynomial expansion representing the feature contour $F_k(i/N)$. The segmental feature vector of the segment is therefore formed by these $3K$ coefficients.

2.2 Distance measure

Before discussing the joint segmentation/clustering algorithm used in the training phase to derive a set of segment-based reference templates for recognition, the distance measure between two variable-length segments must be defined first. We now consider two segments $\boldsymbol{X} = [x_0, x_1, \ldots, x_{\tau_x}]$ and $\boldsymbol{Y} = [y_0, y_1, \ldots, y_{\tau_y}]$, where x_i and y_j are K-dimensional feature vectors, and $\tau_x + 1$ and $\tau_y + 1$ denote the corresponding lengths of these two segments, respectively. We refer to $\boldsymbol{X}$ as the test segment and to $\boldsymbol{Y}$ as the reference segment. Let $\boldsymbol{A}_x$ and $\boldsymbol{A}_y$ be the two segmental feature vectors representing $\boldsymbol{X}$ and $\boldsymbol{Y}$. The matching distance of mapping $\boldsymbol{Y}$ to $\boldsymbol{X}$ is then defined as

$$D_{Seg}(X, Y) = \sum_{i=0}^{\tau_x} \sum_{k=1}^{K} \left[F_k^x\left(\frac{i}{\tau_x}\right) - \hat{F}_k^y\left(\frac{i}{\tau_x}\right)\right]^2 \tag{4}$$

where $F_k^x(i/\tau_x)$ is the kth original feature contour of X and $\hat{F}_k^y(i/\tau_x)$ denotes the kth smoothed feature contour reconstructed using $\boldsymbol{A}_y$. According to the orthogonality property, $D_{Seg}(X, Y)$ can be decomposed into two parts: (i) the contour-fit distortion D_{cf}

$$D_{cf}(X) = \sum_{i=0}^{\tau_x} \sum_{k=1}^{K} \left[F_k^x\left(\frac{i}{\tau_x}\right) - \hat{F}_k^x\left(\frac{i}{\tau_x}\right)\right]^2 \tag{5}$$

between the original test feature contour $\hat{F}_k^x(i/\tau_x)$ and its orthogonally expanded reconstructed version $\hat{F}_k^x(i/\tau_x)$, and (ii) the dissimilarity measure between two smoothed reconstructed segments

$$\begin{aligned}\hat{D}_{Seg}(X, Y) &= (\tau_x + 1)(\boldsymbol{A}_x - \boldsymbol{A}_y)^t(\boldsymbol{A}_x - \boldsymbol{A}_y)\\ &= \sum_{i=0}^{\tau_x} \sum_{k=1}^{K} \left[\hat{F}_k^x\left(\frac{i}{\tau_x}\right) - \hat{F}_k^y\left(\frac{i}{\tau_x}\right)\right]^2 \end{aligned}\tag{6}$$

We note that, from above definition, a segment-based time warping is implicitly embedded in the orthonormal polynomial expansion of calculating the matching distance that maps the reference segment to the test segment.

2.3 Path constraints

Because the time scales of a test token and a reference template are, in general, not perfectly aligned, it is important for speech recognition to tolerate possible durational deviations between them. In our segment-based speech recognition approach, the problem of time alignment is solved by segment-based nonlinear time warping of reference template with some constraints introduced on finding the optimal matching path associated with the least accumulated matching distance between the test token and the reference template. Two types of constraints, referred to as local duration constraints and global path constraints, were used. They are similar to the constraints used in the conventional dynamic time warping (DTW) method [10]. Consider a reference template with S segments. Let the durations of these S segments be $\tau_1, \tau_2, \ldots, \tau_S$. Let the total length of the first s segments be denoted as $L_s = \sum_{i=1}^{s} \tau_i$. To avoid excessively compressing or expanding the time scale, two local duration constraints were set via defining a maximum expansion factor E_{MAX} and a maximum compression factor E_{MIN}. In our study, $E_{MAX} = 2$ and $E_{MIN} = 0.5$. By these two local duration constraints, the duration of a test segment being admitted to match with the sth segment of the reference template is restricted to the range $[E_{MIN}\tau_s, E_{MAX}\tau_s]$. With these two parameters given, the global path constraints that confine the region of optimal path searching can be expressed by the following two equations:

$$1 + E_{MIN}(L_s - 1) \leqslant t_s \leqslant 1 + E_{MAX}(L_s - 1) \tag{7}$$

$$M + E_{MAX}(L_s - N) \leqslant t_s \leqslant M + E_{MIN}(L_s - N) \tag{8}$$

for $s = 1, 2, \ldots, S$, where t_s is the time index of the possible ending point of the sth segment of the test token, and M and N are the durations of the test token and the reference template, respectively. Confined by those path constraints, the search space for finding the optimal path of template matching is greatly reduced.

2.4 Reference template generation

We now consider the training process of generating segment-based reference templates of syllables from a large training set. For simplicity, segment-based reference templates are referred to in the following as reference templates. Assume that the number of segments in every reference template is the same and determined in advance as S. A two-stage algorithm is proposed to generate several reference templates for each syllable class from the ensemble of training tokens of the syllable. In the first stage, all training tokens are first segmented into S segments by an MD segmentation algorithm [7, 8]. Pairwise matching distances are then calculated for all training tokens belonging to the same syllable class. To avoid the problem caused by unconsistent segmentation, all path constraints discussed above are applied in the calculation of pair-wise matching distances. A modified k-means clustering algorithm [9] is then applied to generate a set of initial reference templates.

Then, in the second stage, a joint segmentation/clustering algorithm, which alternatively performs partition/resegmentation and template-updating, is employed to refine all reference templates. It is a two-step iterative procedure. The first step of the iterative procedure is to optimally partition and resegment all training tokens by using the given set of reference templates. This is implemented by first calculating matching distances of comparing each training token with all reference templates, and then assigning it to the cluster associated with the reference template with minimum matching distance. Specifically, for a training utterance $\boldsymbol{TU}$ with N frames, the matching distance of comparing it with a reference template $\boldsymbol{R}_j$ is calculated by

$$D(\boldsymbol{TU}, \boldsymbol{R}_j) = \min_{\tau_1, \ldots, \tau_S} \left\{\frac{1}{N} \sum_{s=1}^{S} D_{Seg}(\boldsymbol{tu}_s, \boldsymbol{r}_{j,s})\right\} \tag{9}$$

where τ_s is the duration of the sth segment, $\boldsymbol{tu}_s$, of $\boldsymbol{TU}$, $\boldsymbol{r}_{j,s}$ is the sth segment of $\boldsymbol{R}_j$, and S is the number of segments of $\boldsymbol{R}_j$. It is noted that the segment durations of $\boldsymbol{TU}$ must satisfy the constraint $\sum_{s=1}^{S} \tau_S = N$. After calculating all matching distances, we then assign $\boldsymbol{TU}$ to cluster j^* if

$$D(\boldsymbol{TU}, \boldsymbol{R}_{j^*}) \leqslant D(\boldsymbol{TU}, \boldsymbol{R}_j) \quad \forall j = 1, 2, \ldots, J, j \neq j^* \tag{10}$$

where J is the number of reference templates of the syllable to which $\boldsymbol{TU}$ belongs. Besides, the set $\{\tau_s, 1 \leqslant s \leqslant S\}$ which achieves the minimum distance $D\ (\boldsymbol{TU}, \boldsymbol{R}_{j*})$ is taken as the optimal resegmentation of $\boldsymbol{TU}$. Actually, the optimal partition and resegmentation is realised by an efficient dynamic programming procedure.

The second step of the iterative procedure is to update all reference templates based on the results of partition and re-segmentation obtained in the first step. We first collect all training tokens belonging to the same cluster together and then find a new representative reference template for each cluster based on the minimum mean square error (MMSE) criterion. Because all training tokens have been properly segmented in the first step, this can be simply done by averaging all training tokens in the cluster segment-by-segment. In fact, a duration-weighted averaging is applied in our realisation to satisfy the MMSE criterion. Specifically, the sth segmental feature vector, $\boldsymbol{A\hat{r}}_s^j$, of the jth new reference template is calculated by

$$\boldsymbol{A\hat{r}}_s^j = \frac{\sum_{i=1}^{I(j)} \tau_{i,s} \times \boldsymbol{A}_{i,s}}{\sum_{i=1}^{I(j)} \tau_{i,s}} \qquad (11)$$

where $\tau_{i,s}$ and $\boldsymbol{A}_{i,s}$ are, respectively, the duration and the segmental feature vector of the sth segment of the ith training token, and $I(j)$ is the number of training tokens belonging to the jth cluster. Besides, a set of supplementary parameters, $\{\hat{\tau}_s^j, 1 \leqslant s \leqslant S\}$, corresponding to the durations of those S segments of the new reference template, has to be updated. Because durational parameters are only applied to constrain the searching path in the calculation of matching distance, any set $\{\hat{\tau}_s^j, 1 \leqslant s \leqslant S\}$ which makes all optimal resegmentations, $\{\tau_{i,s}, 1 \leqslant s \leqslant S\}$, $i = 1, 2, \ldots, I(j)$, of training tokens in the jth cluster in the current iteration satisfy both types of local duration constraints and global path constraints is said to be an admissible set. For simplicity, we choose from all admissible sets the one whose components are closest to the average segment durations of optimal resegmentations in the current iteration as the new durational parameters. Because the partition/resegmentation optimisation procedure in the first step is based on a minimum distance criterion and the updating operation in the second step is based on an MMSE criterion, this ensures that the iterative procedure of the second stage will result in generating a sequence of sets of reference templates with a nonincreasing average matching distance of representing training tokens in the ensemble of the syllable class. According to the generalised Lloyd algorithm [12, 13], it is guaranteed to converge.

3 Experimental results

The performance of the proposed recognition scheme was examined by simulations on a multispeaker speech recognition task for all 408 highly confusing isolated Mandarin base syllables. The database used in our simulations consists of two parts: (i) ten repetitions of utterances of all 408 base-syllables uttered by a single female speaker and (ii) three repetitions of utterances of all 408 base-syllables uttered by each of 12 speakers including eight male and four female. The female who produced the first part of the database is also a speaker of the second part. So we combined these two parts into one in our simulations. The first eight repetitions of utterances of the first part and 24 repetitions of utterances (two repetitions for each speaker) of the second part were chosen as the training data, and all others were used as the testing data. All speech signals were sampled at a rate of 8 kHz and preemphasised with the digital filter, $1\text{–}0.95z^{-1}$. It is then analysed for each Hamming-windowed frame of 32 ms with 8 ms frame shift. 12 Mel-scale cepstral coefficients and their derivatives were then extracted and used as the recognition features.

Since there were no other researches worked on this database in the past, we used the recognition results of the continuous HMM (CHMM) method which uses left-to-right n-state models with m-mixture Gaussian distributions as the bench mark for performance comparison. The CHMM for each syllable was trained by the Baum–Welch reestimation method with an initial model obtained by the followng steps. First, all training tokens were individually divided into n segments by an MD algorithm. [7, 8]. Each segment corresponds to a state. Then, the LBG algorithm [12] was used to partition all observation vectors in each state into m clusters. Vectors in each cluster were modelled by a Gaussian distribution. The mean vector and the (diagonal) covariance matrix of the Gaussian distribution were then estimated for each cluster. The transition probabilities and the mixture weights were also estimated.

The performance of the proposed segment-based speech recognition scheme was first examined for the case of using five reference templates for each syllable. Various number of segmental feature vectors were used for each reference template. In this study, only 12 Mel-scale cepstral coefficients and 12 data cepstral coefficients were used as spectral features. The Top 5 recognition rates for the experiments of using 4, 5, 6, and 8 segmental feature vectors are shown in Table 1. It is noted that Top n recognition rate means that the correct syllable class is appeared in the best n candidates of recognising the test utterance. Basically, Top n recognition rate is important for a Mandarin speech recogniser when we consider to incorporate a language model into it. As described previously, the purpose of the proposed scheme is to properly handle the correlation among contiguous frames within an acoustic segment for improving speech recognition performance. But, it is known [3, 4] that the capability of a CHMM in dealing with speech-frame correlation could also be enhanced by using a larger number of states and/or adding some higher order derivative information as additional observation features. We therefore in the following examine the performance of the CHMM method. Experiments using different number of states per syllable were done. The mixture number of Gaussian distributions in each state was fixed at five. Two cases of using different sets of observation features were tested. In the first case, only 12 Mel-scale cepstral and 12 delta cepstral coefficients were used. In the second case, all features of 12 Mel-scale cepstral, 12 delta cepstral, and 12

Table 1: Recognition results of the segment-based scheme with five reference templates per syllable*

No. of segments	Top 1, %	Top 2, %	Top 3, %	Top 4, %	Top 5, %
4	76.6	90.1	94.6	96.5	97.5
5	77.9	90.5	94.6	96.5	97.5
6	78.6	91.3	95.2	96.9	97.2
8	79.6	91.4	95.2	96.8	97.8

* Recognition features include 12 cepstral and 12 delta cepstral coefficients.

Table 2: Recognition results of the CHMM with 5-mixture Gaussian distributions per syllable*

No. of states	Top 1, %	Top 2, %	Top 3, %	Top 4, %	Top 5, %
4	69.9	84.3	90.7	93.6	95.4
6	70.8	85.4	91.1	94.0	95.6
8	71.6	85.8	91.8	94.5	96.0
10	72.8	87.1	92.4	94.9	96.3
12	73.1	86.9	92.1	94.9	96.7

* Recognition features include 12 cepstral and 12 delta cepstral coefficients

Table 3: Recognition results of the CHMM with 5-mixture Gaussian distributions per syllable*

No. of states	Top 1, %	Top 2, %	Top 3, %	Top 4, %	Top 5, %
4	73.6	87.3	92.9	95.3	97.0
6	74.6	88.5	93.7	95.9	96.9
8	74.4	88.4	93.7	95.7	96.9
10	75.4	88.9	93.7	96.0	97.2
12	76.6	89.8	94.4	96.5	97.5

* Recognition features include 12 cepstral, 12 delta cepstral, and 12 delta–delta cepstral coefficients

delta-delta cepstral coefficients were used. Tables 2 and 3 show the recognition results of the CHMM method for these two cases, respectively. It is found from these two tables that the performance of the CHMM method can be improved by increasing the number of states as well as by adding higher order derivative information. However, as comparing Table 1 with Tables 2 and 3, we find that the proposed method outperforms the well-modelled CHMM method. Actually, only five segmental feature vectors were needed for each reference template in the proposed scheme to perform better than the 12-state CHMM method. We can therefore conclude that the proposed segmental feature vector can more efficiently model the correlation among contiguous frames within an acoustic segment so as to improve the recognition performance.

The performance of the proposed method using all features of 12 Mel-scale cepstral, 12 delta cepstral, and 12 delta–delta cepstral coefficients was also tested. The number of reference templates was still fixed at five. The recognition results of experiments using four, five and six segmental feature vectors for each reference template are listed in Table 4. By comparing Table 4 with Table 1, we find that the difference between the recognition rates of these two cases is insignificant. This result shows that adding delta–delta cepstral coefficients has no effect on improving the recognition performance in the proposed method. This is not a surprising result because the information of delta–delta cepstral coefficients has already implicitly contained in the segmental feature vectors extracted from Mel-scale cepstral and delta cepstral coefficients. We can therefore treat delta–delta cepstral coefficients as dummy features in the proposed method. This experimental result also shows that the segmental feature vector could more efficiently capture the frame correlation than directly using higher order derivative information.

Table 4: Recognition results of the segment-based scheme with five reference templates per syllable*

No. of segments	Top 1, %	Top 2, %	Top 3, %	Top 4, %	Top 5, %
4	76.8	90.2	94.5	96.4	97.2
5	77.9	90.6	94.7	96.3	97.5
6	78.5	91.3	95.2	96.9	97.7

* Recognition features include 12 cepstral, 12 delta cepstral, and 12 delta–delta cepstral coefficients

The effect of increasing the number of reference templates in each syllable class was then examined. Basically, the performance of a speech recogniser will be improved as we increase the number of model parameters because the variability of speech signals can be more accurately modelled. However, model parameters should not be over-expanded to prevent from making the training data insufficient for properly estimating themselves. In this study, the cases of using eight reference templates per syllable with four, five and six segmental feature vectors were tested. Only 12 Mel-scale cepstral and 12 delta cepstral coefficients were used as spectral features. The recognition rates are displayed in Table 5. A Top 1 recognition rate of 81.1% was achieved. As comparing Table 5 with Table 1, we find that the recognition rate increased as the number of reference templates per syllable was increased. For performance comparison, the effect of increasing the mixture number of Gaussian distributions in the CHMM method was also tested. Here, all 12 Mel-scale cepstral, 12 delta cepstral and 12 delta–delta cepstral coefficients were used. Table 6 shows the recognition results of experiments using eight mixtures of Gaussian distributions for the two cases of using 6 and 8 states. Comparing Table 6 with Table 3, we find that the performance of the CHMM method became to drop as the mixture number of Gaussian distributions had been increased from five to eight. It shows from above discussions that the proposed method has the advantage of more efficiently using training data to estimate its model parameters than the CHMM method. For the convenience of performance comparison, the Top 1 recognition rates of the best cases of the proposed method (listed in Table 5) and the CHMM (listed in Table 3) versus the number of model parameters are displayed in Fig. 2. Is it worth noting that the model parameters of the proposed method include segment feature vectors and duration constraint parameters. For CHMM, the model parameters include the mean vectors, the diagonal covariance matrices, the mixture weights, and the transition probabilities. It can be seen from Fig. 2 that the proposed

Table 5: Recognition results of the segment-based scheme with eight reference templates per syllable*

No. of segments	Top 1, %	Top 2, %	Top 3, %	Top 4, %	Top 5, %
4	79.6	91.4	95.4	96.8	97.6
5	81.1	92.3	95.5	96.8	97.8
6	80.9	92.1	95.8	97.1	97.9

* Recognition features include 12 cepstral and 12 delta cepstral coefficients

Table 6: Recognition results of the CHMM with 8-mixture Gaussian distributions per syllable*

No. of states	Top 1, %	Top 2, %	Top 3, %	Top 4, %	Top 5, %
6	72.4	86.1	91.4	94.1	95.5
8	72.7	86.5	91.6	94.0	95.6

* Recognition features include 12 cepstral, 12 delta cepstral, and 12 delta–delta cepstral coefficients

method significantly outperforms the CHMM method as approximately the same number of model parameters was used.

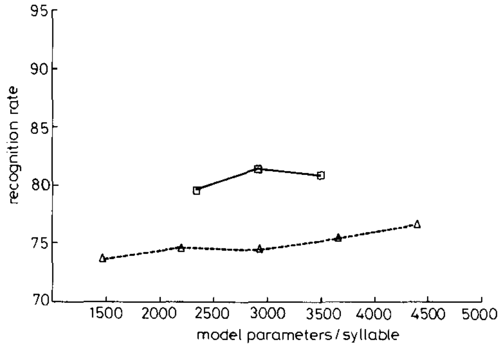

Fig. 2 *Recognition rates against number of model parameters for the segment-based recognition scheme and the CHMM method*

□—□ Segment-based recognition scheme
Δ---Δ CHMM

Due to the fact that the HMM method is also popular in continuous speech recognition, we discuss in the following the possibility of extending the proposed method to the task of recognising 408 syllables in continuous Mandarin speech. The simplest way is to directly extend the method by using syllable models represented by segmental reference templates. In the training, the well-known segmental *k*-means algorithm, which is commonly used for train HMM models, can be modified and applied to generate segmental reference templates of syllables from a set of continuous training utterances. Similarly, the one-stage algorithm can also be modified in the recognition test to find the best recognised syllable sequence for a given input utterance. So the proposed method is equally applicable to continuous Mandarin speech recognition. A preliminary study on this work has confirmed the feasibility of the approach and shown its advantages [14].

4 Conclusions

A novel segment-based speech recognition scheme has been discussed. It differs from conventional frame-based methods on taking segmental features representing spectral parameter contours as the recognition features to take advantage of the strong temporal correlation of speech signal. Effectiveness of the approach on speech recognition has been demonstrated using a challenge task of recognising 408 highly confusing isolated Mandarin base-syllables. Experimental results have confirmed that it outperforms the conventional CHMM method using well-modelled 12-state, 5-mixture HMM models.

5 References

1 RABINER, L.R.: 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proc. IEEE*, Feb. 1989, **77**, (2), pp. 257–286

2 WELLEKENS, C.J.: 'Explicit time correlation in hidden Markov models for speech recognition'. Proc. ICASSP-87, 1987, pp. 384–386

3 KENNY, P., LENNIG, M., and MERMELSTEIN, P.: 'A linear predictive HMM for vector-valued observations with applications to speech recognition', *IEEE Trans.*, Feb. 1990, **ASSP-38**, (2), pp. 220–225

4 WOODLAND, P.C.: 'Hidden Markov models using vector linear prediction and discriminative output distributions'. Proc. ICASSP-92, 1992, pp. 509–512

5 DENG, L., HASSANEIN, K., and ELMASTRY, M.: 'Neural-Network architecture for linear and nonlinear predictive hidden Markov models: application to speech recognition'. Proceedings of the IEEE workshop on *Neural networks for signal processing*, 1991, pp. 411–421

6 OSTENDORF, M., and ROUKOS, S.: 'A stochastic segment model for phoneme-based continuous speech recognition', *IEEE Trans.*, Dec. 1989, **ASSP-37**, (12), pp. 1857–1869

7 BRIDLE, J.S., and SEDGWICK, N.C.: 'A method for segmenting acoustic patterns with application to automatic speech recognition'. Proc. ICASSP-77, 1977, pp. 656–659

8 SVENDSEN, T., and SOONG, F.K.: 'On the automatic segmentation of speech signals'. Proc. ICASSP-87, 1987, pp. 77–80

9 WILPON, J.G., and RABINER, L.R.: 'A modified *k*-means clustering algorithm for use in isolated work recognition', *IEEE Trans.*, Jun. 1985, **ASSP-33**, (3), pp. 587–595

10 MYERS, C., RABINER, L.R., and ROSENBERG, A.E.: 'Performance tradeoffs in dynamic time warping algorithms for isolated word recognition', *IEEE Trans.* Dec. 1980, **ASSP-28**, (6), pp. 623–635

11 CHANG, S., CHEN, S.H., CHUNG, C.J., and HONG, V.: 'A low data rate LPC Vocoder using contour quantization'. Proceeding of EUSIPCO-92 sixth European signal processing conference, 1992, pp. 459–462

12 LINDE, Y., BUZO, A., and GRAY, R.M.: 'An algorithm for vector quantizer design', *IEEE Trans.*, 1980, **COM-28**, pp. 84–95

13 SABIN, M.J., and GRAY, R.M.: 'Global convergence and empirical consistency of the generalized Lloyd algorithm', *IEEE Trans.*, 1986, **IT-32**, pp. 148–155

14 CHANG, S., HONG, V., and CHEN, S.H.: 'Segment-based speech recognition for continuous Mandarin speech'. International workshop on *Speech processing in Japan*, 1993, pp. 23–28

15 LIU, F., LEE, Y., and LEE, L.S.: 'A direct-concatenation approach to train hidden Markov models to recognize the highly confusing Mandarin syllables with very limited training data', *IEEE Trans. Speech Audio Process*, 1994, **1**, pp. 113–119

16 CHANG, P.C.: 'A study on discriminative training for speech recognition'. PhD thesis, National Chiao Tung University, 1993

17 WANG, J.F., WU, C.H., CHANG, S.H., and LEE, J.Y.: 'A hierarchical neural network model based on a C/V segmentation algorithm for isolated Mandarin speech recognition', *IEEE Trans. Signal Process.*, 1991, **39**, pp. 2141–2145

18 JOU, I.C., HU, M.S., and JUANG, Y.T.: 'Mandarin syllables recognition based on one class one net neural network with modified selective update algorithm'. IEEE international workshop on *Intelligent signal processing and communication systems*, 1992, pp. 577–591