Segmental probability distribution model approach for isolated Mandarin syllable recognition

J.-L. Shen

Indexing terms: Segmental probability distribution model, Mandarin speech recognition, Information-theoretic distance measurement

Abstract: A segmental probability distribution model (SPDM) approach is proposed for fast and accurate recognition of isolated Mandarin syllables. Instead of the conventional frame-based approach such as the hidden Markov model (HMM), the model matching process in the proposed SPDM is evaluated segment-by-segment based on information-theoretic distance measurements. The training and recognition procedures for the SPDM are developed first. Several distance measurement criteria, including the Chernoff distance, Bhattacharyya distance, Patrick–Fisher distance, divergence and a Bayesian-like distance, are used, and formulations and comparative results are discussed. Experimental results show that, compared to the widely used sub-unit based continuous density HMM, the proposed method leads to an improvement of 15.27% in the error rate, with a 12-fold increase in recognition speed and less than three quarters of the mixture requirements.

1 Introduction

For a practical automatic speech recognition system, fast and accurate recognition is crucial. However, in most situations, there is a trade-off between the recognition accuracy and speed, i.e. the higher the recognition accuracy required, the more recognition time is needed [1–3]. In this paper, a segmental probability distribution model (SPDM) approach is proposed. In this approach, the computational complexity of the training and recognition procedures is greatly reduced. In addition, improved recognition accuracy can be achieved, compared to the hidden Markov model (HMM) based approach, for recognition of the highly confusing isolated Mandarin syllables.

The Chinese language is not alphabetic, and the input of Chinese characters into computers remains a difficult and unsolved problem. Voice input is believed to be a very attractive solution. Mandarin Chinese is a monosyllabic-structured tonal language [1]. Although there are at least 100 000 commonly used words, composed of more than 10 000 commonly used characters, the total number of phonetically allowed syllables is only 1345. Moreover, each Mandarin syllable is assigned a tone, and there are four lexical tones and one neutral tone. If the differences in tones are disregarded, these 1345 Mandarin syllables can be reduced to 408 different base syllables. As the tones can be separately recognised using primarily pitch contour information, the recognition of all 408 Mandarin base syllables is believed to be the key problem for large vocabulary Mandarin speech recognition, due to the monosyllabic structure of Mandarin Chinese.

The hidden Markov model (HMM) approach has been tested with high recognition rates for Mandarin base syllable recognition [1]; the similarity between a test utterance and the acoustic models is measured frame-by-frame with a Viterbi-searched optimal path [2]. Although HMMs offer a fine stochastic representation of speech production, their computational load, both in training and recognition, is extremely high. As an alternative, a segmental probability model (SPM) has been shown to be very suitable for Mandarin base syllable recognition, especially considering the monosyllabic structure of the Chinese language [4]. The SPM is very similar to continuous density HMM (CHMM), except that the state transition probabilities are deleted and a linear warping function is used to divide the syllable utterances into N states. In other words, the stochastic state transition behaviour in HMMs is replaced by a deterministic process in SPM, and the output distributions are also represented by Gaussian mixtures. As shown in our preliminary results [4], comparable recognition rates with the HMM approach can be obtained in the SPM approach, with greatly reduced computational complexity, when the same model configurations are used.

In this paper, a segmental probability distribution model (SPDM) approach is proposed to further reduce the computational complexity for fast and accurate recognition of Mandarin syllables. The utterance to be recognised is first divided into N segments, using the linear warping function, as in the SPM approach. Each segment is then represented by an associated probability distribution (PD), for example, a unimodal Gaussian distribution. These probability distributions are used to measure the similarity between this utterance and the acoustic models with some information-theoretic distance measurement criterion. In other words, the similarity between training and testing speech spectra is measured in terms of the distance between their associated probability distributions. In this way, the recognition process can be evaluated segment-by-segment instead of frame-by-frame [5]. The computational load is significantly reduced because the processing

IEE Proceedings online no. 19982313

Paper first received 30th June 1997 and in final revised form 18th August 1998

The author is with the Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan, Republic of China

time is proportional to the segment number N instead of the total frame number T of this utterance and explicitly $N << T$.

Experimental results show that not only can the recognition speed be greatly improved, but the recognition accuracy can also be maintained by carefully choosing the distance measurement criterion. A family of information-theoretic distance measurement criteria, including the Chernoff distance, Bhattacharyya distance, Patrick–Fisher distance, divergence and a Bayesian-like distance, are used and compared. The training and recognition procedures for the proposed SPDM are also developed.

2 Formulations of model matching

The similarity between a test utterance O and the acoustic model λ is usually measured as the *a posteriori* probability λ given O, i.e. $p(\lambda|O)$. If the *a priori* probability of λ is assumed to be constant, the similarity measurement $p(\lambda|O)$ can be reduced to the conditional density function $p(O|\lambda)$ by Bayes' theorem.

2.1 Conventional frame-based approach

2.1.1 HMM: In the conventional approach, such as HMM, the similarity is measured frame-by-frame as the following form [2]:

$$p(O|\lambda) = \sum_{all\ S} p(O|\lambda, S)p(S|\lambda) = \sum_{all\ S}\prod_{t=1}^{T} p(o_t|\lambda, s_t)p(S|\lambda) \quad (1)$$

where $S = s_1 s_2 \ldots s_T$ a possible state sequence and $O = o_1 o_2 \ldots o_T$ is the frame sequence with a total of T frames. Here the statistical independence of observations is assumed. If the HMM λ has the set of parameters (A, B, π), where π is the initial state transition probability, $A = a_{s_1 s_2}, \ldots, a_{s_{T-1} s_T}$ is the state transition probability and $B = b_{s_1}, \ldots, b_{s_T}$ is the observation probability, eqn. 1 can be shown as [2]

$$p(O|\lambda) = \sum_{all\ S} \pi_{s_1} b_{s_1}(o_1) a_{s_1 s_2} b_{s_2}(o_2) \ldots a_{s_{T-1} s_T} b_{s_T}(o_T) \quad (2)$$

where $b_{s_t}(o_t) = p(o_t|\lambda, s_t)$, as shown in eqn. 1. Instead, the optimal state sequence with maximum probability is selected using the Viterbi decoding algorithm, and the corresponding probability value is used as the similarity between the test utterance and the acoustic models [2]:

$$p(O|\lambda) = \max_{S} \pi_{s_1} b_{s_1}(o_1) a_{s_1 s_2} b_{s_2}(o_2) \ldots a_{s_{T-1} s_T} b_{s_T}(o_T) \quad (3)$$

2.1.2 SPM: In the SPM, the probability $p(O|\lambda)$ has the following [4]:

$$p(O|\lambda) = b_{s_1}(o_1) b_{s_2}(o_2) \ldots b_{s_T}(o_T) \quad (4)$$

where $s_t = j$, if $f(j-1) < t <= f(j)$. $f(j)$ denotes the ending frame for state (segment) j in an utterance which is predetermined by a linear warping function. In fact, eqn. 4 is a simplified form of eqn. 3, where the decoded state sequence is determined by the linear warping function such that the probability of the state sequence $p(S|\lambda)$ in eqn. 1 can be therefore deleted. Here a simple linear warping function is used to equally segment the utterance into N states.

Figs. 1*a* and *b* show the search spaces needed in the model-matching process for HMM and SPM, respectively, which implies a great reduction in the computational complexity from HMM to SPM. Preliminary experimental results indicate that not only is the computational complexity greatly reduced in SPM, but comparable recognition rates with CHMM can be obtained [4].

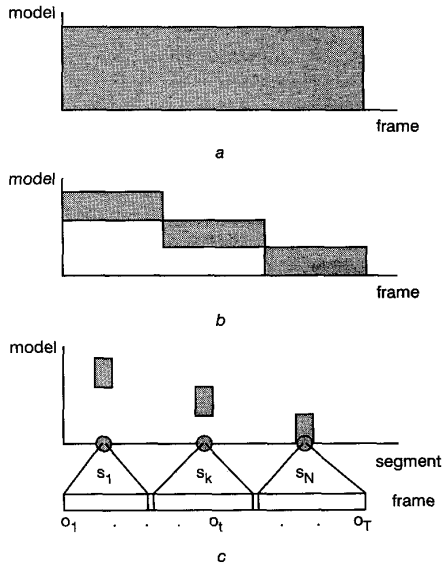

Fig. 1 *Search space for HMM, SPM and SPDM*
a HMM
b SPM
c SPDM

2.2 Segment-based approach: SPDM

In the proposed SPDM approach, the pre-alignment process applied in SPM is first used to divide the utterance into N segments. The observation frame vectors in each segment are then modelled by a PD function. In this way, a sequence of feature vectors is replaced by the parameters of the associated PD and the processing unit is changed from frame to segment. The associated PD for each segment not only reproduces the statistics in this segment, but also captures the time dependency of these observation vectors. Thus, the similarity between a test utterance O and the acoustic model λ can be measured as

$$\begin{aligned} p(O|\lambda) &= p(o_1 o_2 \ldots o_T|\lambda, S) \\ &= p(o_1 \ldots o_{f(1)}|\lambda, s_1) p(o_{f(1)+1} \ldots o_{f(2)}|\lambda, s_2) \\ &\quad \ldots p(o_{f(N-1)+1} \ldots o_T|\lambda, s_N) \\ &= p(G_1|\lambda, s_1) p(G_2|\lambda, s_2) \ldots p(G_N|\lambda, s_N) \end{aligned} \quad (5)$$

where G_j is the corresponding PD for segment j, with a total of N PDs modelling the utterance. Since the prior probability of λ in each segment is assumed to be constant, the similarity $p(O|\lambda)$ is measured as the multiplication of joint probabilities of the distribution and the acoustic models over all segments. This has the following form:

$$p(O|\lambda) = p(G_1, \lambda, s_1) p(G_2, \lambda, s_2) \ldots p(G_N, \lambda, s_N) \quad (6)$$

The similarity between training and testing speech spectra is evaluated in terms of the distortion between their associated PDs, instead of the distance between the individual feature vector and the distribution represented by training speech. In other words, the recognition process depends on the N PDs, instead of the T observation vectors. It is also obvious that the number of probability distributions N is much less than that of the feature vectors T in an utterance. This is why the required recognition time in SPDM can be reduced significantly compared to SPM.

As shown in Fig. 1, the required search spaces are greatly reduced from HMM and from SPM to the proposed SPDM. Two major problems arise in the proposed SPDM approach:

(i) how to derive the acoustic models λ in the SPDM.

(ii) how to evaluate the joint probability $p(Gi, \lambda, s_i)$, $i = 1 \ldots N$.

In the following, we investigate the above two problems, and several solutions are presented and discussed.

3 Training procedure

The block diagram of the training procedure for the SPDM is shown in Fig. 2. Conventionally, each Mandarin syllable is decomposed into the consonant/vowel format like in a western language, and the vowel part includes possible medial and nasal ending [1]. There are 22 context independent (CI) consonants and 41 CI vowels in Mandarin Chinese. These 22 CI consonants can be further expanded into 113 context-dependent (CD) consonants with respect to the beginning phonemes of the following vowels. A linear warping function can be used to divide the syllable utterance into N segments, with equal length for each segment considering the monosyllabic structure of Mandarin Chinese [8]. In this way, the corresponding observation frames that each state (segment) occupies in a Mandarin syllable utterance are easily obtained. For each segment, a PD is then associated with the speech spectra of this segment.

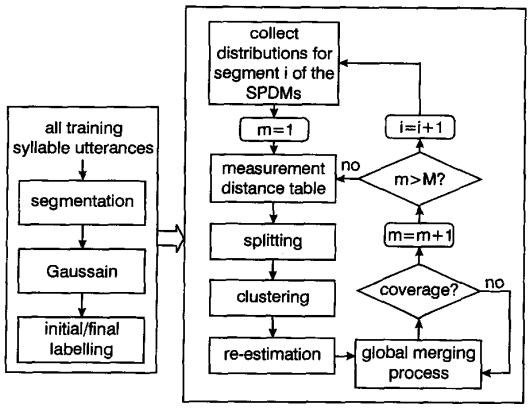

Fig. 2 *Block diagram of training procedure for SPDM*

In this study, the unimodal Gaussian distribution is used to represent the PD such that the corresponding mean vector and covariance matrix can be derived for each segment. In addition, to increase the trainability of the SPDMs, the segment sharing concept is applied. This is where the first few segments of the SPMs with the same CD INITIALs actually bear similar characteristics, and thus can share the same training data; likewise the remaining segments with the same CI FINALs [6]. This is similar to the tied-state methods used in HMM [13]. The modified K-means algorithm is then used to classify these PDs into M mixture components [7]. After the training procedure, those Gaussian distributions with the lowest average measurement distances, i.e. the highest mutual similarities, can be merged into the same mixture component. Compare our method with the segmental K-means method for training a SPM or CHMM, where all training observation samples are vector quantised using the Lloyd algorithm based on the Euclidean distance [13]. In our method, the SPDMs are derived from the Gaussian distributions modelled by these training observation samples, using the modified K-means algorithm based on the information-theoretic distance measurement. As the element to be merged is the Gaussian distribution instead of the sample frame vector, the total training numbers are greatly reduced and the computational load can therefore be reduced tremendously.

4 Recognition procedure

In the recognition procedure, the syllable utterance to be recognised is first divided into N segments and each segment is modelled by a Gaussian distribution, as in the training phase. Taking the logarithm of eqn. 6, the similarity $p(O|\lambda_i)$ between the test utterance and the model λ_i for syllable i can be expressed as

$$p(O|\lambda_i) = \sum_{j=1}^{N} \log p(G_j, \lambda_i, s_j) \tag{7}$$

where G_j is the unimodal Gaussian distribution modelling the speech spectra of the jth segment for this utterance. Here two scoring methods for evaluating the joint probability $p(G_j, \lambda_i, s_j)$ are used:

(i) partitioned distance:

$$p(G_j, \lambda_i, s_j) = \max_{1 \le k \le M} l(DM_k(G_j, \lambda_i, s_j)) \tag{8}$$

(ii) mixture-weighted distance:

$$p(G_j, \lambda_i, s_j) = \sum_{k=1}^{M} w_i(j,k) l(DM_k(G_j, \lambda_i, s_j)) \tag{9}$$

where $w_i(j, k)$ is the mixture gain for adjusting the contribution of each mixture to the similarity, $l(\cdot)$ is the individual similarity for each segment and while $DM_k(\cdot)$ is the measurement distance for the mixture component k.

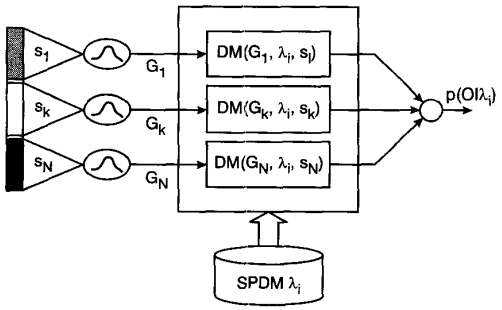

Fig. 3 *Evaluation of similarity measurement in SPDM*

Fig. 3 shows the similarity measurement procedure in the recognition phase of SPDM. Note that lower measurement distance implies higher individual similarity.

In order to maintain the recognition accuracy with a conventional model-matching approach such as HMM or SPM, the information-theoretic distance measurement criterion must be carefully chosen.

5 Information-theoretic distance measurement criteria

In this Section, a family of information-theoretic distance measurement criteria are used to measure the distance between two PDs, including the Chernoff distance, Bhattacharyya distance, divergence, Patrick–Fisher distance and a Bayesian-like distance [11].

5.1 Chernoff distance D_c

Given two classes r_1 and r_2 with the PDs $H_1(x)$ and $H_2(x)$, respectively, the overlap ε between these two distributions can be used as a measure of the similarity between them, which is also called Bayes' error. It is obvious that the higher the value of ε means a greater similarity between these two classes. The mathematical form of the overlap ε can be expressed as

$$\varepsilon_u = \int_x H_1(x)^{1-s} H_2(x)^s dx \tag{10}$$

Using the fact that $\min(H_1(x), H_2(x)) \leq H_1(x)^{1-s} H_2(x)^s$, $0 \leq s \leq 1$, the upper bound of ε, which is called ε_u, can be easily obtained:

$$\varepsilon_u = \int_x \min\left(H_1(x), H_2(x)\right) dx \tag{11}$$

Moreover, if these two PDs are normal (i.e. $N(U_1, \Sigma_1)$ and $N(U_2, \Sigma_2)$, where U_i and Σ_i represent the mean vector and covariance matrix for H_i, respectively) ε_u can be simplified to

$$\varepsilon_u = e^{-D_c} \tag{12}$$

where

$$D_c = \frac{s(1-s)}{2}(U_1 - U_2)^T \times (s\Sigma_1 + (1-s)\Sigma_2)^{-1}(U_1 - U_2) + \frac{1}{2}\log\frac{|s\Sigma_1 + (1-s)\Sigma_2|}{|\Sigma_1|^s|\Sigma_2|^{1-s}}$$

This is called the Chernoff distance [8], which can be used as the distance measurement in eqns. 8 and 9.

In order to obtain the parameter s of the Chernoff distance, two kinds of methods are used. First, the parameter s is fixed to 1/2, which is a special case of the Chernoff distance called the 'Bhattacharyya' distance (D_b). Secondly, the parameter s is optimised empirically. In practice, the parameter s can be designed with a different value for different models or an unique value for all models. Different optimisation criteria can be applied to find the optimal value of s, instead of estimation by experiments.

5.2 Divergence D_d

The divergence is a kind of distance-like criterion from information theory which can be expressed as [9]

$$D_d = E\left\{\log\frac{H_1(x)}{H_2(x)}|r_1\right\} + E\left\{\log\frac{H_2(x)}{H_1(x)}|r_2\right\} \tag{13}$$

where the expected values of the log-likelihood-ratio for classes r_1 and r_2 are used. When $H_1(x)$ and $H_2(x)$ are represented by Gaussian distributions as mentioned above, eqn. 13 can be extended as follows:

$$D_d = \frac{1}{2}(U_1 - U_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(U_1 - U_2) + \frac{1}{2}trace(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I) \tag{14}$$

where $trace(A)$ means the summation of the diagonal terms of the matrix A.

5.3 Patrick–Fisher distance D_p

The Patrick–Fisher distance between the two Gaussian distributions $H_1(x)$ and $H_2(x)$ can be derived from the following measurement criterion, based on the integral of the Euclidean distance for each observation vector x [10]:

$$e^{D_p} = \left[\int_x (H_1(x) - H_2(x))^2 dx\right]^{1/2} = \left[-\frac{2}{\sqrt{2\pi}|\Sigma_1 + \Sigma_2|^{1/2}} \times e^{-\frac{1}{2}(U_1-U_2)^T(\Sigma_1+\Sigma_2)^{-1}(U_1-U_2)} + \frac{1}{\sqrt{2\pi}|2\Sigma_1|^{1/2}} + \frac{1}{\sqrt{2\pi}|2\Sigma_2|^{1/2}}\right]^{1/2} \tag{15}$$

Taking the logarithm of eqn. 15, we can obtain the Patrick–Fisher distance D_p.

5.4 Bayesian-like distance D_{bl}

The Bayesian-like distance is derived from the conventional likelihood function of observing a feature vector sequence $o_1 o_2 \ldots o_T$ for a class r_1 with Gaussian distribution $H_1(x)$. The likelihood function can be measured by Bayes' theorem as in eqn. 1, which has the following form:

$$\prod_{t=1}^{T} p(o_t|r_1) = \frac{1}{\sqrt{2\pi}^T|\Sigma_1|^{T/2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(o_t-U_1)^T\Sigma_1^{-1}(o_t-U_1)} \tag{16}$$

If the frame sequence $o_1 o_2 \ldots o_T$ is modelled by an unimodal Gaussian distribution $N(U_2, \Sigma_2)$, eqn. 16 can be expressed in another form:

$$\frac{1}{\sqrt{2\pi}^T|\Sigma_1|^{T/2}} e^{-\frac{1}{2}(U_1-U_2)^T\Sigma_1^{-1}(U_1-U_2) - \frac{T}{2}trace(\Sigma_2\Sigma_1^{-1})} = e^{-D_{bl}} \tag{17}$$

Thus, taking the minus logarithm of eqn. 17, we can obtain the Bayesian-like distance D_{bl}.

From the above discussion, we can conclude that the similarity between two PDs can be measured based on the following three criteria:

(i) the overlap or unoverlap regions between them such as D_c, D_b, and D_p.

(ii) the discriminant information due to entropy measures, e.g. in D_d, the difference of entropy and cross-entropy of the two PDs is used.

(iii) the likelihood scores, given a probability distribution and the data derived from another distribution.

Note that similar forms can be obtained for two Gaussian distributions, i.e. the distance measurement

can be separated into two terms, where the first term gives class separability due to the mean difference, and the second term gives the class separability due to the covariance difference. Moreover, the first part of the distance measurement due to mean difference is the weighted Euclidean distance (i.e. Mahalanobis distance), in which different weighting factors are derived from the combination of the covariance matrices Σ_1 and Σ_2 of the two distributions for different distance measurement criteria. On the other hand, different formulations using the covariance matrices Σ_1 and Σ_2 are provided for different distance measurement criteria in the second part. Despite theoretical differences for these measures, we can choose between them according to recognition speed and accuracy for the purpose of speech recognition.

6 Experimental results and discussion

6.1 Speech database

The speech database used in all experiments was produced by three speakers. For each speaker, four utterances of each of the 1345 Mandarin tonal syllables were produced in isolation. In all experiments, three utterances of each of the 1345 Mandarin syllables are used in training and one utterance is used in testing for three speakers, respectively. The quoted recognition rates are the average of the rates for each of the speakers. All the speech data are obtained in an office-like laboratory environment. They are low-pass filtered, digitised by an Ariel S-32C DSP board with sampling frequency 16kHz. After end-point detection is performed, a 20ms Hamming window is applied every 10ms with a pre-emphasis factor of 0.95. 14-order mel-frequency cepstral coefficients, derived from the power spectrum filtered by a set of 30 triangular band-pass filters, are used as feature parameters.

6.2 Experiments

6.2.1 Choose the distance measurement criterion: The experimental results with respect to different distance measurement criteria discussed in Section 5 are shown in Table 1, where the partitioned distance is used for evaluating the recognition procedure. It can be found that the Chernoff distance D_c yields the best recognition accuracy, which indicates a recognition rate as high as 91.62%. Here the value of the parameter s in the Chernoff distance is optimised by experiments. Fig. 4 shows the influence of s on the recognition rates, where the value of s ranges from 0.1 to 0.9. In addition, slight degradation on recognition rates is provided using the Bhattacharyya distance D_b (s = 0.5) and Bayesian-like distance D_{bl}, i.e. 91.47% and 91.06%. The recognition rates are also reduced by 3.27% and 4.57% using the divergence D_d and Patrick–Fisher distance D_p, respectively.

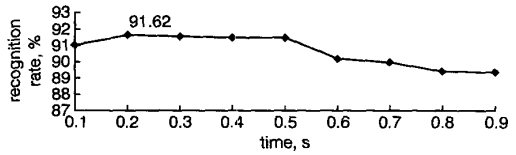

Fig. 4 *Recognition rates with respect to different parameters s in CD-SPDM*

6.2.2 Choose the scoring method in the recognition procedure: Table 2 shows the experimental results for the two scoring methods in the recognition phase, i.e. partitioned distance and mixture-weighted distance, for the SPDM based on the Chernoff distance (CD-SPDM). Note that the required recognition time using the partitioned distance is much less than that using the mixture-weighted distance, with an increased recognition rate of 2.61%. Therefore, the CD-SPDM using the partitioned distance is used in all the following experiments.

6.2.3 Comparative results for various acoustic models: The extensive experimental results for the various SPMs, various CHMMs based on different acoustic units and the CD-SPDMs proposed here are compared in Table 3. The symbol (N, M) in syllable-based CHMM and SPM means N states (segments) and M mixtures per state. In SS-SPMs, a segment shared concept based on the CD consonant/CI vowel format of a Mandarin syllable is applied for SPM, as mentioned in Section 3. In fact, the model configuration of the CD-SPDMs is the same as that of the SS-SPMs, i.e. the first few segments model the 113 CD consonants and the remaining segments model the 41 CI vowels. In addition, two widely used sub-unit-based CHMMs for Mandarin syllable recognition such as consonant/vowel-based CHMM and phone-based CHMM are evaluated as a comparison [12]. In the consonant/vowel-based CHMM, the symbol (n_1, n_2, M) means the first n_1 states model the 113 CD consonants, the remaining n_2 states model the 41 CI vowels and each state contains M mixtures. On the other hand, in phone-based CHMM, 149 right context-dependent (RCD) phone-like units are used. The symbol (N, M) means each RCD phone-like unit is modelled by N states, where each state is represented by M mixtures.

CD-SPDM outperforms SPM both in recognition rates (91.62% against 90.61%) and in speed (0.083 s/syl.

Table 1: Comparative results according to different distance measurement criteria

	Chernoff	Bhattacharyya	Divergence	Patrick-Fisher	Bayesian-like
Recognition rates (%)	91.62	91.47	88.35	87.05	91.06

Table 2: Experimental results for two scoring methods in recognition phase using CD-SPDM with model configuration (2, 3, 3)

Model		Recognition time (s/syl in SPARC 10)	Recognition rates (%)
CD-SPDM	Partitioned distance	0.083	91.62
	Weighted distance	0.103	89.01

Table 3: Comparison of recognition accuracy, speed and number of distributions for various acoustic models with different model configurations

Model	Type	Total number of distributions	Recognition time (s/syl in SPARC 10)	Recognition rates (%)
SPM	(3, 2)	2448	0.996	90.61
SS-SPM	(1, 2, 3)	585	0.214	89.20
	(2, 2, 3)	924	0.254	91.37
	(2, 3, 3)	1047	0.226	92.37
Syllable-based CHMM	(3, 2)	2448	3.846	79.73
Consonant/vowel-based CHMM	(3, 3, 3)	1386	2.012	88.10
	(3, 4, 2)	1006	1.799	88.50
	(3, 4, 3)	1509	2.337	89.74
Phone-based CHMM	(2, 2)	596	1.074	87.24
	(2, 3)	894	1.486	88.60
	(3, 3)	1341	2.262	90.43
CD-SPDM	(1, 2, 3)	585	0.047	87.90
	(2, 2, 3)	924	0.074	90.28
	(2, 3, 3)	10.47	0.083	91.62

against 0.996 s/syl.), with far fewer mixtures required (1047 against 2448). However, compared to SS-SPM, nearly three times the recognition speed can be achieved (0.083 s/syl. against 0.226 s/syl.) in the CD-SPDM at the expense of a 0.75% recognition rate (91.62% against 92.37%) when the same model configurations are used. Furthermore, in comparison with the phone-based CHMM, more than 25 times the recognition speed can be achieved (0.083 s/syl. against 2.262 s/syl.) with a 12.43% error rate reduction (91.62% against 90.43%) using less than 4/5 mixture numbers (1047 against 1341) in the proposed CD-SPDMs. Note that from Table 2 SS-SPM and CD-SPDM can provide the best performance on recognition accuracy and speed, respectively.

6.2.4 Evaluating the training procedure: Table 4 shows the experimental results that the acoustic models in SS-SPMs are directly used to perform the CD-SPDM, in which the recognition rate is reduced from 91.62% to 88.20%. This is because of the unmatched conditions in training and recognition. Accordingly, the effectiveness of the proposed training procedure can be confirmed.

Table 4: Experimental results for evaluating the training procedure

	Recognition rates using CD-SPDM (%)
CD-SPDM	91.62
SS-SPM	88.20

6.2.5 Beam search processing: The beam search is the most well known method to speed up the recognition process in the HMM-based approach [2]. Table 5 exhibits the experimental results with different beam widths for various acoustic models. It is obvious that less beam width in the recognition process implies less recognition time and, in most situations, lower recognition rates. Table 5 shows that when the beam width is set to 400, the recognition rates are reduced by 0.32% and 0.34%, respectively, in phone-based and consonant/vowel-based HMM, and nearly three times the recognition speed can be obtained. However, in comparison with the CD-SPDMs as listed in the last row of Table 5, the recognition speed is nine times higher and the error rates are increased by more than 15%. On the other hand, this beam search method can be also used in the CD-SPDMs. Here the searching element is changed to segment instead of frame, and the total beam width is 408. The experimental results, as shown in the last part of Table 5, indicate that the beam search method can further speed up the recognition procedure in the CD-SPDMs with a slightly reduced recognition rate. As the beam width is set to 70, the required time to recognise a syllable is reduced from 0.083s to 0.067s, and an identical recognition rate of 91.62% with full search can be achieved. Accordingly, in comparison with the most successfully sub-syllabic CHMMs, more than 12 times the recognition speed and 15.27% error rate reduction can be achieved using less than 3/4 of the mixture numbers only in the proposed CD-SPDMs.

Table 5: Experimental results for various acoustic models using beam search method

Model type	Beam width	Time (s/syl)	Rate (%)
Phone-based CHMM (3, 3)	300	0.624	89.02
	400	0.754	90.11
	500	0.881	90.21
	800	1.282	90.38
	full search	2.262	90.43
Consonant/vowel-based CHMM (3, 4, 3)	300	0.657	88.01
	400	0.792	89.40
	500	0.936	89.59
	800	1.429	89.72
	full search	2.337	89.74
CD-SPDM (2, 3, 3)	30	0.060	90.65
	50	0.063	91.32
	60	0.065	91.47
	70	0.067	91.62
	full search	0.083	91.62

In the last experiment as shown in Fig. 5, we applied the beam search method to the SPDMs based on different information-theoretic distance measurement criteria. Similar trends as with the CD-SPDMs can be obtained, i.e. the error rate increase is less than 0.5% when the beam width is set higher than 50. However, around 1% recognition rate reduction can be achieved when the beam width is reduced to 30.

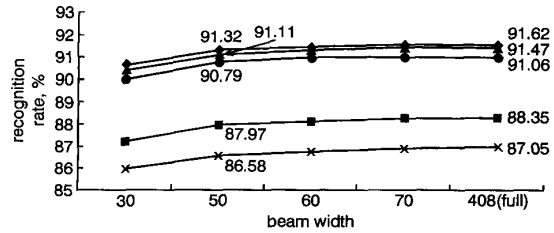

Fig. 5 *Recognition results with respect to different beam width for different information-theoretic distance measurement criteria*
—◆— Chernoff —▲— Bhattacharyya —■— divergence
—×— Patrick–Fisher —○— Bayesian-like

7 Conclusions

A segmental probability distribution model (SPDM) approach for Mandarin syllable recognition has been proposed. Instead of conventional frame-by-frame distortion measures, the recognition process was evaluated segment-by-segment based on information-theoretic distance measurements. A family of distance measurement criteria were used and compared, including the Chernoff distance, Bhattacharyya distance, Patrick–Fisher distance, divergence and a Bayesian-like distance. Experimental results show that not only can the recognition time be reduced tremendously, but also improved recognition rates and fewer mixture requirements can be achieved in the proposed SPDM as compared to the widely used sub-unit based CHMMs.

8 Acknowledgments

The author would like to thank Dr. Lin-shan Lee for his valuable suggestions and discussions, the reviewers for their insightful comments.

9 References

1 LEE, L.S., TSENG, C.Y., GU, H.Y., LIU, F.H., CHANG, C.H., LIN, Y.H., LEE, Y.M., TU, S.L., HSIEH, S.H., and CHEN, C.H.: 'Golden Mandarin(I) - a real-time Mandarin speech dictation machine for Chinese language with very large vocabulary', *IEEE Trans. Speech Audio Process.*, 1993, **1**, (2), pp. 158–179

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

3 OSTENDORF, M., DIGALAKIS, V.V., and KIMBALL, O.A.: 'From HMM's to segment models: a united view of stochastic modeling for speech recognition', *IEEE Trans., Speech Audio Process.*, 1996, **4**, (5), pp. 360–378

4 LYU, R.Y., HONG, I.C., SHEN, J.L., LEE, M.Y., and LEE, L.S.: 'Isolated Mandarin base-syllable recognition based upon the segmental probability model (SPM)', *IEEE Trans. Speech Audio Process.*, 1998, **6**, (3), pp. 293–299

5 SHEN, J.L., and LEE, L.S.: 'A Chernoff distance based segmental probability model (CD-SPM) approach for Mandarin syllable recognition'. Proceedings of Eurospeech, Madrid, Spain, September 1995, pp. 1491–1494

6 SHEN, J.L., WANG, H.M., LYU, R.Y., and LEE, L.S.: 'Incremental speaker adaptation using phonetically balanced training sentences for Mandarin syllable recognition based on segmental probability models'. Proceedings of international conference on *Spoken language process*, Tokyo, Japan, 1994, pp. 443–446

7 WILPON, J.G., and RABINER, L.R.: 'A modified K-means clustering algorithm for use in isolated word recognition', *IEEE Trans. Acoustics, Speech Signal Process.*, 1985, **ASSP-33**, (3), pp. 587–594

8 FUKUNAGA, K.: 'Introduction to statistical pattern recognition (Academic Press, Chap. 3, San Diego, 1990)

9 PATRICK, E.A., and FISHER, F.P.: 'Nonparametric feature selection', *IEEE Trans., Information Theory*, 1969, **15**, pp. 577–584

10 KULLBACK, S.: 'Information theory and statistics' (Wiley, New York, 1959)

11 LEE, Y.T.: 'Information-theoretic distortion measures for speech recognition', *IEEE Trans. Signal Process.*, 1991, **39**, (2), pp. 330–335

12 LYU, R.Y., WANG, H.M., and LEE, L.S.: 'A comparison of different units applied to isolated/continuous large vocabulary Mandarin speech recognition'. Proceedings of international conference on *Computer processing of oriental languages*, May 1994, (Korea),pp. 211–214

13 LEE, C.H., GIACHIN, E., RABINER, L.R., PIERACCINI, R., and ROSENBERG, : 'Improved acoustic modeling for large vocabulary continuous speech recognition', *Computer Speech Language*, 1992, **6**, pp. 103–127