ELSEVIER

Speech Communication 36 (2002) 317–326

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Discriminative training of Gaussian mixture bigram models with application to Chinese dialect identification

Wuei-He Tsai, Wen-Whei Chang *

Department of Communications Engineering, National Chiao-Tung University, Hsinchu, Taiwan, ROC

Received 21 July 2000; received in revised form 1 September 2000; accepted 6 October 2000

Abstract

This study focuses on the parametric stochastic modeling of characteristic sound features that distinguish languages from one another. A new stochastic model, the so-called Gaussian mixture bigram model (GMBM), that allows exploitation of the acoustic feature bigram statistics without requiring transcribed training data is introduced. For greater efficiency, a minimum classification error (MCE) algorithm is employed to accomplish discriminative training of a GMBM-based Chinese dialect identification system. Simulation results demonstrate the effectiveness of the GMBM for dialect-specific acoustic modeling, and use of this model allows the proposed system to distinguish between the three major Chinese dialects spoken in Taiwan with 94.4% accuracy. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Gaussian mixture bigram model; Minimum classification error algorithm; Chinese dialect identification

1. Introduction

Until recently, research in Chinese language processing was almost exclusively aimed at voice dictation of Mandarin Chinese (Lee, 1997). However, hundreds of Chinese dialects exist and may be linguistically divided into seven major groups, namely, Mandarin, Holo, Hakka, Wu, Yue, Xiang and Gan (Ramsey, 1987). There are minor differences within each group, but there are major differences between groups of such magnitude that the groups sound mutually unintelligible. The interconnections between the Chinese dialects are in fact as complicated as those among the family of Romance languages, such as French, Spanish and Italian. This motivated our research into trying to build a system capable of recognizing the identities of spoken Chinese dialects. During the initial stage of developing our dialect-ID system, special efforts were made to discriminate among three major Chinese dialects spoken in Taiwan, Mandarin, Holo and Hakka. Although most statements made about Mandarin syntax are also applicable to Holo and Hakka, the three dialects differ significantly in pronunciation and vocabulary. The dialectal differences arise not only from variations in phonetic inventory and phonotactics, but also from the tonal system employed in individual dialects. Traditional descriptions of spoken Chinese divide syllables into two parts: basic syllables, which are realized by combinations of consonants and vowels, and tones, which are realized by contrasting variations in pitch contours. According to a recent estimate (Tsai, 1997), there are 408

* Corresponding author.

0167-6393/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00090-X

basic syllables and 4 tones in Mandarin, 808 basic syllables and 7 tones in Holo, and 708 basic syllables and 6 tones in Hakka.

Designing a reliable language and/or dialect identification scheme requires that stochastic models be used to summarize some of the most relevant aspects of language acoustics. Previous work (Hazen and Zue, 1997) suggested that language characteristics are represented in the segmental and prosodic features of speech utterances. Segmental features can be acoustic–phonetic, which refers to the acoustic realizations of phonemes, or phonotactic, which refers to the rules governing combination of various phonemes in a language. Prosodic information is encoded in the pitch, amplitude and duration variations that span across segments. There have been numerous recognition schemes proposed to accomplish the language identification (language-ID) task, with varying degrees of success (Muthusamy et al., 1994). In the phonotactic components of most language-ID systems, one or more phonetic recognizers are used for tokenizing speech utterances into sequences of phonemes or broad phonetic classes, followed by a set of *n*-gram language models (House and Neuburg, 1977; Zissman, 1996; Hazen and Zue, 1997). However, porting such phonetic approaches to the problem of Chinese dialect-ID may present its own set of problems. This is mainly because phonetic recognizers have to be trained and evaluated on a phonetically labeled database, which is not available for every dialect. For Chinese spoken languages, the existing phonetic transcription system was designed only with Mandarin in mind and hence proves to be insufficient to accommodate other dialects such as Holo and Hakka.

In this study, we present a theoretical framework of Gaussian mixture bigram models (GMBMs) and use them to characterize the temporal and spectral evolution of the speech signal. The main attraction of GMBMs arises from the fact that the observations used in dialect-specific modeling are extracted directly from the acoustic features, instead of using the imperfect outputs of the front-end phonetic recognizer. Thus the model parameters of GMBMs can be estimated without any transcription of training utterances; allowing us to circumvent the difficult task due to the lack of an agreed-upon system appropriate for transcribing all of the dialects. Furthermore, we believe that the language-discriminating power of the system can be improved by using the GMBM, which integrates time correlation on acoustic frames into the model structure. Although using acoustic frame dependency was previously proposed for speech recognition (Wellekens, 1987; Deng, 1994), their emphases were placed upon hidden Markov models with model parameters estimated according to the maximum likelihood (ML) criterion. The basic problem with ML estimation is that each model is trained independently of the others and hence cannot obtain model parameters which maximize classification accuracy. A better solution to parameter estimation is based on the minimum classification error (MCE) criterion (Juang et al., 1997). Discriminative training by the MCE method has been successfully applied to various kinds of classifier frameworks including the hidden Markov model, dynamic time warping, and neural networks. In this study, we present a specific implementation of the MCE training procedure in the context of a GMBM-based dialect-ID system.

The rest of this paper is organized as follows. Section 2 presents the stochastic framework required for solving the problem of Chinese dialect-ID. Section 3 introduces the basic formulations of a Gaussian mixture bigram model. In Section 4, we address the parameter estimation problem using the expectation-maximization algorithm. Details of the MCE training procedure required for discriminative estimation of the GMBM parameters are provided in Section 5. Section 6 presents Chinese dialect-ID results obtained using different design approaches. Finally, Section 7 provides a short summary.

2. Problem formulation

Although Chinese dialects differ significantly from each other with respect to their phonologies and vocabularies, the task of exploiting these characteristics involves high-level linguistic knowledge that usually requires skilled linguists fluent in all of the dialects. To accomplish this task is not only impractical, but also limits the system's applicability to identify dialects for which the vocabulary and linguistic rules are not well specified. Motivated by the above concern, we attempted to examine whether acoustic features can be incorporated directly into the dialect-ID process without requiring human-supplied linguistic knowledge. Toward this end, it is a prerequisite to establish stochastic models that summarize some of the most relevant aspects of language acoustics.

To begin, let $\boldsymbol{D} = \{D_j\}_{j=1}^{J}$ represent a set of J Chinese dialects to be identified. A dialect-ID system takes test utterances as input and produces the identities of the dialects being spoken as outputs. Choosing an appropriate representation of acoustic information is the first step in applying statistical methods to solving the dialect-ID problem. The specific types of feature measurements considered here are mel-cepstral features and pitch-based features. The reasons are as follows. First, the primary difference for the tones is in the pitch contours and the tones are essentially independent of the spectral envelope parameters of the syllables. Second, the fact that Chinese is a tonal language suggests the combination of segmental and prosodic features is likely to help in distinguishing between dialects with greater accuracy. Speech signals were preprocessed to extract mel-cepstral features every 20-ms Hamming-windowed frames with 10-ms frame shifts. Let $\boldsymbol{X}^{(1)} = \{\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_{T^{(1)}}^{(1)}\}$ denote a sequence of $T^{(1)}$ feature vectors, each of which consists of 10 mel-frequency cepstral coefficients and their first derivatives. Pitch measurements were extracted from voiced speech segments using the simple inverse filter tracking (SIFT) algorithm (Markel, 1972). Let $\boldsymbol{X}^{(2)} = \{\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \ldots, \boldsymbol{x}_{T^{(2)}}^{(2)}\}$ denote a sequence of $T^{(2)}$ feature vectors representing the pitch and differential pitch.

The dialect-ID system operates in two phases: training and recognition. In the training phase, pitch and mel-cepstral features were extracted from speech signals and then used to train stochastic models for every dialect to be recognized. During recognition, dialects were identified by matching test utterances with the stochastic model of each dialect and by calculating the probability $p(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)} \,|\, D_j)$ for each dialect D_j. According to the maximum likelihood decision rule, the identifier should decide in favor of a dialect $\hat{D}$ satisfying

$$\hat{D} = \arg\max_j \log p(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)} \,|\, D_j). \tag{1}$$

In practical realization, the search process can be aided by taking advantage of the statistical independence between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. This appears due to the fact that mel-cepstrum and pitch are separately characterized by the speakers' vocal tract shapes and excitation signals. The dialect-ID process can thus be simplified as follows:

$$\hat{D} = \arg\max_j \sum_{i=1}^{2} \beta_j^{(i)} \log p(\boldsymbol{X}^{(i)} \,|\, \lambda_j^{(i)}), \tag{2}$$

where $\beta_j^{(i)}$ and $\lambda_j^{(i)}$ denote the weighting coefficient and the stochastic model associated with the feature sequence $\boldsymbol{X}^{(i)}$, respectively. A block diagram of the proposed dialect-ID system is shown in Fig. 1.

3. Gaussian mixture bigram model

The effectiveness of the dialect-ID system crucially depends on how well the feature representation $\boldsymbol{X}^{(i)}$ and the stochastic model $\lambda_j^{(i)}$ capture the relevant information for discriminating between different dialects. In this section, we will develop the theoretical framework of so-called GMBM. The use of the bigram model is motivated by previous experiments showing that the sequential statistics of acoustic features could be exploited as an efficient approach to language-ID. For notational convenience, the feature measurement $\boldsymbol{X}^{(i)}$ and its associated model $\lambda_j^{(i)}$ will be denoted as $\boldsymbol{X}$ and λ here, respectively. Given the model λ, the probability of the observation sequence $\boldsymbol{X}$ is determined according to the following expression:

$$p(\boldsymbol{X} \,|\, \lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}, \lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}, \lambda), \tag{3}$$

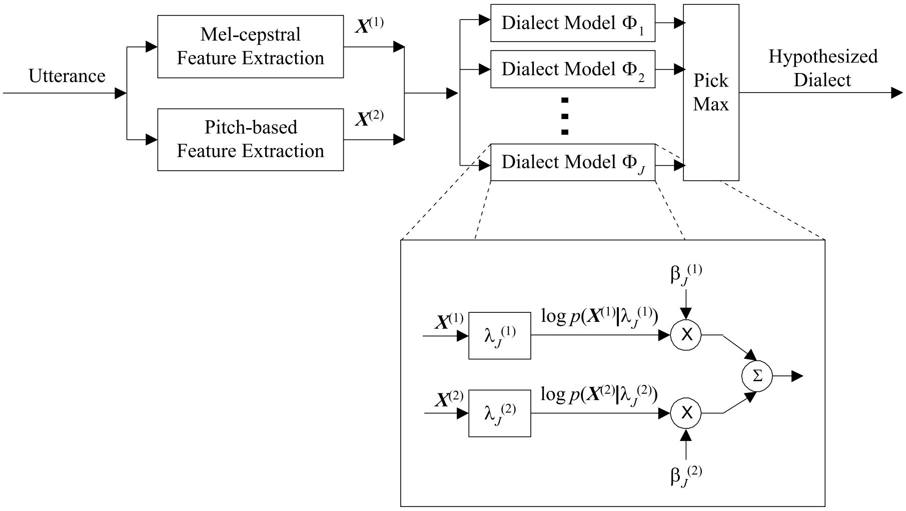

Fig. 1. Illustration of the proposed dialect-ID system.

where we have assumed that the probability of observing $\boldsymbol{x}_t$ depends only on the immediately preceding vector $\boldsymbol{x}_{t-1}$.

Depending upon the choice of density function $p(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}, \lambda)$, a number of different bigram model structures can be realized. One method is to perform discrete observation bigram modeling after feature representations have been converted into sequences of discrete symbols via vector quantization (VQ) (Hazen and Zue, 1993; Harbeck and Ohler, 1999). The basic advantage of this approach is that the bigram models are trained and evaluated on codebook sequences, which can be trained without any transcription of training data. However, its applicability may be restricted due to the fact that observations used in bigram modeling are not extracted directly from acoustic feature measurements, but from a discrete set of representative templates. To compensate for this shortcoming, we propose to represent the feature distribution using a probabilistic mixture model based on a weighted sum of component Gaussian densities. In using this approach, it is assumed that the feature space is characterized by a set of broad acoustic classes and, moreover, within each class the acoustic frame dependency is modeled by a component Gaussian density. The bigram density function $p(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}, \lambda)$ assumes the following form in our current model implementation:

$$p(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}, \lambda) = \sum_{n=1}^{N} w_n f_{\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1},n,\lambda}(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1}, n, \lambda), \qquad (4)$$

where w_n denotes the mixture weight subject to the constraint $\sum_{n=1}^{N} w_n = 1$. The well-known Bayes' formula of probability theory allows us to rewrite the nth mixture component density as

$$f_{\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1,n,\lambda}}(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t-1,n,\lambda}) = \frac{f_{\boldsymbol{x}_t,\boldsymbol{x}_{t-1} \,|\, n,\lambda}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1} \,|\, n, \lambda)}{f_{\boldsymbol{x}_{t-1} \,|\, n,\lambda}(\boldsymbol{x}_{t-1} \,|\, n, \lambda)}. \qquad (5)$$

To permit theoretical analysis, we further assume that $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$ are two vectors of jointly Gaussian random variables; the joint density is then

$$f_{\boldsymbol{x}_t,\boldsymbol{x}_{t-1} \,|\, n,\lambda}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1} \,|\, n, \lambda) = \frac{|\boldsymbol{C}_n|^{-1/2}}{(2\pi)^M} \exp\left[-\frac{1}{2}(z_t - \boldsymbol{m}_n)' \boldsymbol{C}_n^{-1} (z_t - \boldsymbol{m}_n) \right], \qquad (6)$$

where prime denotes vector transpose and $z_t = [\boldsymbol{x}_{t-1}\ \boldsymbol{x}_t]'$ is an augmented feature vector associated with mean vector $\boldsymbol{m}_n$ and covariance matrix $\boldsymbol{C}_n$. From this it can be shown that the conditional density of $\boldsymbol{x}_t$ given $\boldsymbol{x}_{t-1}$ is also Gaussian in the form of

$$f_{\boldsymbol{x}_t\,|\,\boldsymbol{x}_{t-1},n,\lambda}(\boldsymbol{x}_t\,|\,\boldsymbol{x}_{t-1},n,\lambda) = \mathcal{N}(\boldsymbol{x}_t,\mu_n,\boldsymbol{\Sigma}_n)$$
$$= \frac{|\boldsymbol{\Sigma}_n|^{-1/2}}{(2\pi)^{M/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}_t-\mu_n)'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{x}_t-\mu_n)\right], \tag{7}$$

where $\mathcal{N}(\cdot)$ represents an M-variate Gaussian density with mean vector $\boldsymbol{\mu}_n$ and covariance matrix $\boldsymbol{\Sigma}_n$ given by

$$\boldsymbol{\mu}_n = \boldsymbol{\theta}_n + \boldsymbol{B}_n\boldsymbol{x}_{t-1}, \tag{8}$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{C}_n^{(4)} - \boldsymbol{C}_n^{(3)}\boldsymbol{C}_n^{(1)^{-1}}\boldsymbol{C}_n^{(2)}. \tag{9}$$

In Eq. (8), we have

$$\boldsymbol{\theta}_n = \boldsymbol{m}_n^{(2)} - \boldsymbol{C}_n^{(3)}\boldsymbol{C}_n^{(1)^{-1}}\boldsymbol{m}_n^{(1)}, \tag{10}$$

$$\boldsymbol{B}_n = \boldsymbol{C}_n^{(3)}\boldsymbol{C}_n^{(1)^{-1}}, \tag{11}$$

where the $M \times 1$ matrix $\boldsymbol{m}_n^{(k)}$ and the $M \times M$ matrix $\boldsymbol{C}_n^{(k)}$ are derived, respectively, from $\boldsymbol{m}_n$ and $\boldsymbol{C}_n$ as follows:

$$\boldsymbol{m}_n = \begin{bmatrix} \boldsymbol{m}_n^{(1)} \\ \boldsymbol{m}_n^{(2)} \end{bmatrix}_{2M\times 1}, \tag{12}$$

$$\boldsymbol{C}_n = \begin{bmatrix} \boldsymbol{C}_n^{(1)} & \boldsymbol{C}_n^{(2)} \\ \boldsymbol{C}_n^{(3)} & \boldsymbol{C}_n^{(4)} \end{bmatrix}_{2M\times 2M}. \tag{13}$$

4. Maximum likelihood parameter estimation

In this section, we address the estimation problem which involves optimizing the GMBM parameters to match the distribution of training feature vectors. Within the GMBM framework, the parameter set λ consists of all the mixture weights w_n, mixture Gaussian mean vectors $\boldsymbol{\mu}_n$, and mixture Gaussian covariance matrices $\boldsymbol{\Sigma}_n$. The traditional approach to parameter estimation is based on the ML principle. Given a set of training feature vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, the aim of ML estimation is to find the model parameters which maximize the GMBM likelihood, i.e.,

$$\lambda_{\mathrm{ML}} = \arg\max_{\lambda} \log \prod_{t=1}^{T} p(\boldsymbol{x}_t\,|\,\boldsymbol{x}_{t-1},\lambda)$$
$$= \arg\max_{\lambda} \log \prod_{t=1}^{T}\sum_{n=1}^{N} w_n\mathcal{N}(\boldsymbol{x}_t,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n). \tag{14}$$

Toward this end, the expectation-maximization (EM) algorithm (Dempster et al., 1977) is applied here to estimate the model parameters which guarantees a monotonic increase in the likelihood.

Starting with an initial model λ, the new model $\bar{\lambda}$ is estimated by maximizing the auxiliary function

$$Q(\lambda,\bar{\lambda}) = \sum_{t=1}^{T}\sum_{n=1}^{N} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)$$
$$\times \log p(n,\boldsymbol{x}_t\,|\,\boldsymbol{x}_{t-1},\bar{\lambda}), \tag{15}$$

where

$$p(n,\boldsymbol{x}_t\,|\,\boldsymbol{x}_{t-1},\bar{\lambda}) = \bar{w}_n\mathcal{N}(\boldsymbol{x}_t,\bar{\boldsymbol{\mu}}_n,\bar{\boldsymbol{\Sigma}}_n) \tag{16}$$

and

$$p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda) = \frac{w_n\mathcal{N}(\boldsymbol{x}_t,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)}{\sum_{q=1}^{N} w_q\mathcal{N}(\boldsymbol{x}_t,\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q)}. \tag{17}$$

On each EM iteration, the reestimation formulas derived for individual parameters of the nth mixture density are of the form

$$\bar{w}_n = \frac{1}{T}\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda), \tag{18}$$

$$\bar{\boldsymbol{\theta}}_n = \frac{\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)(\boldsymbol{x}_t-\boldsymbol{B}_n\boldsymbol{x}_{t-1})}{\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)}, \tag{19}$$

$$\bar{\boldsymbol{B}}_n = \left[\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)(\boldsymbol{x}_t-\theta_n)\boldsymbol{x}'_{t-1}\right]$$
$$\times \left[\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)\boldsymbol{x}_{t-1}\boldsymbol{x}'_{t-1}\right]^{-1}, \tag{20}$$

$$\bar{\boldsymbol{\Sigma}}_n = \frac{\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)(\boldsymbol{x}_t-\boldsymbol{\mu}_n)(\boldsymbol{x}_t-\boldsymbol{\mu}_n)'}{\sum_{t=1}^{T} p(n\,|\,\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\lambda)}. \tag{21}$$

The new model $\bar{\lambda}$ then becomes λ for the next iteration and the reestimation process is repeated until the likelihood reaches a fixed value.

5. Discriminative training of a GMBM-based dialect-ID system

This section addresses the implementation issues for discriminative estimation of the entire parameter set in the context of a GMBM-based dialect-ID system. Each dialect D_j is characterized by a parameter set $\Phi_j = \{(\beta_j^{(i)}, \lambda_j^{(i)}), i = 1, 2\}$ which consists of weighting coefficients $\beta_j^{(i)}$ and the GMBM parameters $\lambda_j^{(i)} = \{w_{j,n}^{(i)}, \boldsymbol{\theta}_{j,n}^{(i)}, \boldsymbol{B}_{j,n}^{(i)}, \boldsymbol{\Sigma}_{j,n}^{(i)} \mid n = 1, 2, \ldots, N\}$ for the two measurements, one representing the mel-cepstral features ($i = 1$) and the other representing the pitch-based features ($i = 2$). The approach employed in the previous section was to estimate the model parameters using the EM algorithm according to the ML criterion. However, the ML approach often does not lead to an optimum performance in classification tasks. An alternative approach to parameter estimation is based on MCE criterion (Juang et al., 1997). The major advantage of the MCE approach is that discrimination between different models can be improved by incorporating out-of-class information during training. In this study, the MCE algorithm was extended to accomplish discriminative estimation of the model parameters of GMBMs.

Consider a set of training tokens with known dialect identities $\boldsymbol{O} = \{\boldsymbol{O}^{(l)}\}_{l=1}^{L}$, where each token $\boldsymbol{O}^{(l)}$ is composed of mel-cepstral features $\boldsymbol{X}^{(l,1)}$ with length $T^{(l,1)}$ and pitch-based features $\boldsymbol{X}^{(l,2)}$ with length $T^{(l,2)}$. Based on $\boldsymbol{O}$, the goal of the MCE estimation is to find the identifier parameter set $\Lambda = \{\Phi_1, \Phi_2, \ldots, \Phi_J\}$ such that the probability of misclassifying all training tokens is minimized. A typical approach in this direction is the generalized probabilistic descent (GPD) algorithm (Katagiri et al., 1991), in which model parameters are adjusted iteratively to better represent the statistics of a training database. Below we present a specific implementation of the GPD algorithm for a GMBM-based dialect-ID system.

1. Calculate a set of discriminant functions,

$$\mathscr{G}_k(\boldsymbol{O}^{(l)}; \Lambda) = \sum_{i=1}^{2} \beta_k^{(i)} \log p(\boldsymbol{X}^{(l,i)} \mid \lambda_k^{(i)}), \quad k = 1, 2, \ldots, J. \tag{22}$$

2. Calculate the misclassification measure for a training token $\boldsymbol{O}^{(l)}$ from dialect D_k,

$$\mathscr{M}_k(\boldsymbol{O}^{(l)}; \Lambda) = -\mathscr{G}_k(\boldsymbol{O}^{(l)}; \Lambda) + \log \left\{ \frac{1}{J-1} \sum_{s, s \neq k} \exp\left[\mathscr{G}_s(\boldsymbol{O}^{(l)}; \Lambda)\eta\right] \right\}^{1/\eta}, \tag{23}$$

where η is a positive real number.

3. Calculate the smoothed loss function

$$\mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda) = \frac{1}{1 + \mathrm{e}^{-\gamma \mathscr{M}_k(\boldsymbol{O}^{(l)}; \Lambda)}}, \tag{24}$$

where the parameter γ controls the function smoothness.

4. To reduce the loss function, the GPD algorithm is used to adjust the weighting coefficients $\beta_j^{(i)}$ and the GMBM parameters $\lambda_j^{(i)}$. Denoting either by $\phi_j^{(i)}$, the new parameter becomes

$$\bar{\phi}_j^{(i)} = \phi_j^{(i)} - \epsilon \sum_{l=1}^{L} \frac{\partial \mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \phi_j^{(i)}}, \tag{25}$$

where ϵ is the step size and where

$$\frac{\partial \mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \phi_j^{(i)}} = \frac{\partial \mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \mathscr{M}_k(\boldsymbol{O}^{(l)}; \Lambda)} \frac{\partial \mathscr{M}_k(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)}; \Lambda)} \frac{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \phi_j^{(i)}}. \tag{26}$$

According to Eqs. (23) and (24), we have

$$\frac{\partial \mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \mathscr{M}_k(\boldsymbol{O}^{(l)}; \Lambda)} = \gamma \mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda)[1 - \mathscr{L}_k(\boldsymbol{O}^{(l)}; \Lambda)]. \tag{27}$$

$$\frac{\partial \mathscr{M}_k(\boldsymbol{O}^{(l)}; \Lambda)}{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)}; \Lambda)} = \begin{cases} -1 & \text{if } j = k \\ \dfrac{\exp[\mathscr{G}_k(\boldsymbol{O}^{(l)}; \Lambda)\eta]}{\sum_{s, s \neq j} \exp[\mathscr{G}_s(\boldsymbol{O}^{(l)}; \Lambda)\eta]} & \text{if } j \neq k. \end{cases} \tag{28}$$

5. We next want to derive the expressions for the partial derivatives of the discriminative function with respect to individual parameters. Further restriction, however, must be imposed on the

parameter adjustment to accommodate various constraints such as the positive definiteness of the covariance matrix $\boldsymbol{\Sigma}_{j,n}^{(i)}$ as well as the stochastic constraints $\sum_n w_{j,n}^{(i)} = 1$ and $\sum_i \beta_j^{(i)} = 1$. This can be done by transforming these constrained parameters to an unconstrained domain and then by computing the gradient with respect to the transformed parameters $\tilde{w}_{j,n}^{(i)}$, $\tilde{\beta}_j^{(i)}$ and $\tilde{\boldsymbol{\Sigma}}_{j,n}^{(i)}$ (Juang et al., 1997; Chengalvarayan and Deng, 1997). It can be shown that the gradient computation of individual parameters are of the form

$$\frac{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)};\Lambda)}{\partial \tilde{w}_{j,n}^{(i)}} = \beta_j^{(i)} \sum_{t=1}^{T^{(l,i)}} \left[p(n \,|\, \boldsymbol{x}_t^{(l,i)}, \boldsymbol{x}_{t-1}^{(l,i)}, \lambda_j^{(i)}) - w_{j,n}^{(i)} \right], \tag{29}$$

$$\frac{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)};\boldsymbol{\Lambda})}{\partial \boldsymbol{\theta}_{j,n}^{(i)}} = \beta_j^{(i)} \sum_{t=1}^{T^{(l,i)}} p(n \,|\, \boldsymbol{x}_t^{(l,i)}, \boldsymbol{x}_{t-1}^{(l,i)}, \lambda_j^{(i)}) \boldsymbol{\Sigma}_{j,n}^{(i)^{-1}} \left[\boldsymbol{x}_t^{(l,i)} - \mu_{j,n}^{(i)} \right], \tag{30}$$

$$\frac{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)};\boldsymbol{\Lambda})}{\partial \boldsymbol{B}_{j,n}^{(i)}} = \beta_i^{(j)} \sum_{t=1}^{T^{(l,i)}} p(n \,|\, \boldsymbol{x}_t^{(l,i)}, \boldsymbol{x}_{t-1}^{(l,i)}, \lambda_j^{(i)}) \boldsymbol{\Sigma}_{j,n}^{(i)^{-1}} \left[\boldsymbol{x}_t^{(l,i)} - \boldsymbol{\mu}_{j,n}^{(i)} \right] \boldsymbol{x}_{t-1}^{(l,i)'}, \tag{31}$$

$$\frac{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)};\boldsymbol{\Lambda})}{\partial \tilde{\boldsymbol{\Sigma}}_{j,n}^{(i)}} = \frac{1}{2} \beta_j^{(i)} \sum_{t=1}^{T^{(l,i)}} p(n \,|\, \boldsymbol{x}_t^{(l,i)}, \boldsymbol{x}_{t-1}^{(l,i)}, \lambda_j^{(i)}) \cdot \left\{ \boldsymbol{\Sigma}_{j,n}^{(i)^{-1}} \left[\boldsymbol{x}_t^{(l,i)} - \boldsymbol{\mu}_{j,n}^{(i)} \right] \left[\boldsymbol{x}_t^{(l,i)} - \boldsymbol{\mu}_{j,n}^{(i)} \right]' - \boldsymbol{I} \right\}, \tag{32}$$

$$\frac{\partial \mathscr{G}_j(\boldsymbol{O}^{(l)};\boldsymbol{\Lambda})}{\partial \tilde{\beta}_j^{(i)}} = \beta_j^{(i)} \left[\log p(\boldsymbol{X}^{(l,i)} \,|\, \lambda_j^{(i)}) - \sum_{s=1}^{2} \beta_j^{(s)} \log p(\boldsymbol{X}^{(l,s)} \,|\, \lambda_d^{(s)}) \right], \tag{33}$$

where $\boldsymbol{I}$ denotes an $M \times M$ identity matrix and $p(n \,|\, \boldsymbol{x}_t^{(l,i)}, \boldsymbol{x}_{t-1}^{(l,i)}, \lambda_j^{(i)})$ above is defined as Eq. (17).

6. Experimental results

Extensive computer simulations have been conducted to test the validity of the proposed dialect-ID system. Our effort began with the collection of a speech corpus that consisted of speech utterances in Mandarin, Holo and Hakka. The corpus was produced by eight speakers, 5 males and 3 females, with the length of each utterance being about 15 s. For this study we attempted to avoid speaker bias by using speakers who are fluent in all of the three dialects studied. Our text materials consisted of 30 folk-tale passages in colloquial style. Each passage consisted of 26–32 Chinese characters, and the texts were grouped by passages into sets of 20 and 10. The set with 20 passages was used for training, and the set with 10 passages was used for testing. All the speakers were asked to read the text three times, once in each of the three dialects. Thus the number of training utterances was 480 (20 passages × 8 speakers × 3 dialects), and the number of test utterances was 240 (10 passages × 8 speakers × 3 dialects). All of the utterances in the speech corpus were recorded in a relatively quiet environment, and then sampled at 16 kHz with 16-bit precision.

Although we expected that the combined use of pitch and mel-cepstral features would increase dialect-discriminating power, it was worth evaluating the relative contribution towards dialect-ID that each feature measurement provided. For example, this can be done by setting $(\beta_j^{(1)}, \beta_j^{(2)}) = (1, 0)$ in Eq. (2) for the case where only access to mel-cepstral feature $\boldsymbol{X}^{(1)}$ was available. A preliminary experiment was first performed to examine the dependency of dialect-ID results on the number of Gaussian mixtures used in the GMBM. A series of dialect-ID results using ML-trained GMBM are summarized in Fig. 2, with the parameters being the number N of mixture components and the feature measurement used. Our general conclusion is that the GMBM-based system yields better performance with an increase in the number N of Gaussian mixtures, but the performance has a tendency to become flattened for high N. In the sequel, we empirically chose the GMBM with $N = 16$ and $N = 8$ to model the mel-cepstral and pitch-based features, respectively.

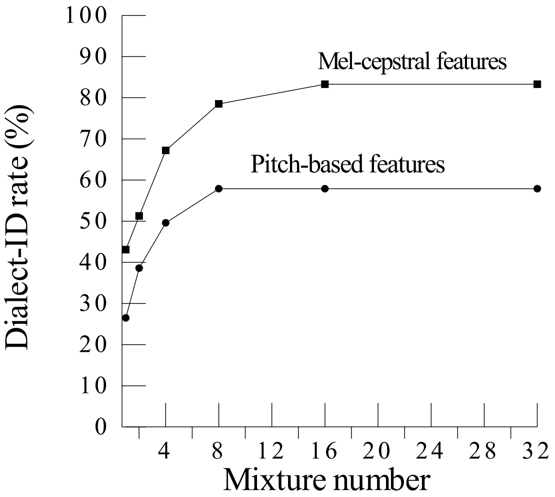

Fig. 2. Dialect-ID performance of the ML-trained GMBMs.

Tables 1 and 2 give dialect-ID results that compare the case where the GMBM was trained using the ML algorithm against the case that was trained using the MCE algorithm. The performance was evaluated in terms of the confusion matrix, the rows of which correspond to the dialects actually being spoken and the columns indicate the dialects identified. During the MCE training phase, the parameter values used for γ, η, and the maximum number of iterations N_{m} were empirically determined to be 1.0, 2.0 and 30, respectively. Additionally, the step size at kth iteration was determined by $\epsilon = 0.01/(1 - k/N_{\mathrm{m}})$. Compared with the ML method, the effectiveness of using the MCE method for discriminative estimation of GMBM parameters is clearly demonstrated. The results of these experiments also indicated that while the mel-cepstral features are useful in Chinese dialect-ID, other sources of information are likely to help in distinguishing between dialects with greater accuracy. The importance of incorporating prosodic information was reflected in the observation that using only pitch information allows the system to identify three dialects with 56.9% accuracy.

For purposes of comparison, we also investigated two other modeling techniques which necessitate no phonetically labeled database. The first approach included two subsystems, the first of which used a vector quantizer to tokenize the incoming utterance into sequences of discrete codebook symbols, and the second used phonotactically motivated language models to identify target dialects. The language model used for this experiment was an interpolated bigram model with parameters estimated according to the relative frequency method (Hazen and Zue, 1997). Hereafter we will refer to this system as VQBM. Fig. 3 shows the dialect-ID results of the VQBM system for a VQ codebook size ranging from 8 to 128. In it we see that compared with the GMBM, the VQBM is less successful when applied to distinguish among Chinese dialects. The better performance associated with the GMBM can be attributed to the fact that it not only provides a smooth approximation to the feature distribution,

Table 1
Dialect-ID results based on ML-trained GMBMs

	Recognition		
Actual	Mandarin	Holo	Hakka
(a) *Using mel-cepstral features*			
Mandarin	**0.83**	0.02	0.15
Holo	0.01	**0.88**	0.11
Hakka	0.10	0.09	**0.81**
(b) *Using pitch-based features*			
Mandarin	**0.54**	0.38	0.08
Holo	0.30	**0.58**	0.12
Hakka	0.11	0.29	**0.59**

Table 2
Dialect-ID results based on MCE-trained GMBMs

	Recognition		
Actual	Mandarin	Holo	Hakka
(a) *Using mel-cepstral features*			
Mandarin	**0.91**	0.02	0.07
Holo	0.01	**0.90**	0.09
Hakka	0.07	0.06	**0.87**
(b) *Using pitch-based features*			
Mandarin	**0.56**	0.36	0.08
Holo	0.27	**0.62**	0.11
Hakka	0.10	0.28	**0.62**

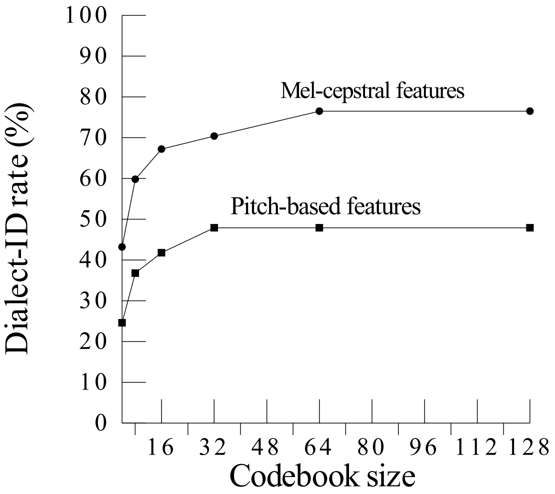

Fig. 3. Dialect-ID performance of the VQBM approach.

its components also clearly detail the multi-modal nature of the density. The second approach we tested was motivated by previous experiments (Zissman, 1996) showing that languages can be distinguished by means of Gaussian mixture models (GMMs). For each dialect, the ML algorithm was used to create two GMMs: one for the mel-cepstral features and the other for the pitch-based features. Fig. 4 shows the dialect-ID results of the GMM-based system for various number of Gaussian mixtures. A comparison between Figs. 2 and 4 revealed that the GMBM is preferred to the GMM for use in modeling of language acoustics. The main reason is that the GMBM takes into account the time correlation on acoustic frames, as opposed to the independence assumption made in the GMM. The investigation further showed that the improvement is especially prominent for GMBM modeling of pitch-based features, indicating that pitch contour dynamics are highly useful in automatic identification of tonal languages such as Chinese.

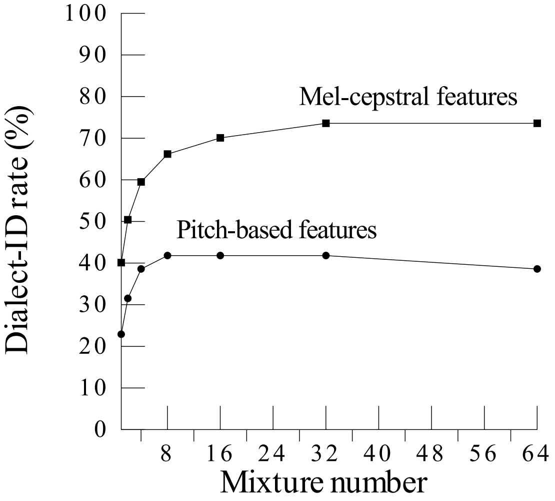

Fig. 4. Dialect-ID performance of the ML-trained GMMs.

Table 3
Dialect-ID results of the overall system trained using MCE algorithm

	Recognition		
Actual	Mandarin	Holo	Hakka
Mandarin	**0.95**	0.02	0.03
Holo	0.01	**0.96**	0.03
Hakka	0.03	0.04	**0.93**

Until now, we only considered the situation where dialects were identified solely by means of either mel-cepstral or pitch-based features. The next step in this investigation concerned the performance improvement that may result from combining mel-cepstral and pitch information within a unified framework. Toward this end, we attempted to determine the most likely dialect hypothesis using Eq. (2). The classifier parameters were jointly estimated using the GPD algorithm based on the MCE criterion, so that the misclassification error rate could be minimized. Table 3 lists a confusion matrix showing the dialect-ID results for the system that examined differences between dialects at mel-cepstral and pitch levels. In it we see that pitch information provides additional performance gain by acting as a secondary source of dialect-discriminating information. The top-choice accuracy was measured to obtain a recognition score of 94.4%.

7. Conclusions

This study discussed methods of incorporating acoustic information directly into the Chinese dialect-ID system without requiring transcribed training data. This task was accomplished by using a weighted sum of Gaussian mixture densities to characterize the bigram statistics of mel-cepstral as well as pitch-based features. Simulation results indicate that this new method is superior to the vector codebook approach, which performs phonotactic analysis on discrete codebook symbols. One enhancement that further increases discrimination across dialects is the use of the MCE algorithm in estimating the classifier parameters. While this study only presented experimental results for the Chinese dialect-ID task, the design techniques used to refine the acoustic characterization can be applied to more general problems in language identification and other speaker identification problems.

Acknowledgements

This research was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC89-2213-E009-188. The authors are very grateful to the unknown reviewers for their careful readings of this paper and their constructive suggestions.

References

Chengalvarayan, R., Deng, L., 1997. Use of generalized dynamic feature parameters for speech recognition. IEEE Trans. Speech Audio Processing 5 (3), 232–242.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39, 1–38.

Deng, L., 1994. Integrated optimization of dynamic feature parameters for hidden Markov modeling of speech. IEEE Signal Processing Lett. 1 (4), 66–69.

Harbeck, S., Ohler, U., 1999. Multigrams for language identification. In: Proceedings EUROSPEECH'99.

Hazen, T.J., Zue, V.W., 1993. Automatic language identification using a segment based approach. In: Proceedings EUROSPEECH'93, pp. 1303–1306.

Hazen, T.J., Zue, V.W., 1997. Segment-based automatic language identification. J. Acoust. Soc. Amer. 101 (4), 2323–2331.

House, A.S., Neuburg, E.P., 1977. Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. J. Acoust. Soc. Amer. 62, 708–713.

Juang, B.H., Chou, W., Lee, C.H., 1997. Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Processing 5 (3), 257–265.

Katagiri, S., Lee, C.H., Juang, B.H., 1991. New discriminative algorithms based on the generalized probabilistic descent method. In: Proceedings of the IEEE-SP Workshop on Neural Network for Signal Processing, pp. 299–308.

Lee, L.S., 1997. Voice dictation of mandarin chinese. IEEE Signal Processing Magazine 14, 63–101.

Markel, J.D., 1972. The SIFT algorithm for fundamental frequency estimation. IEEE Trans. Audio and Electroacoustics 20, 129–137.

Muthusamy, Y.K., Barnard, E., Cole, R.A., 1994. Reviewing automatic language identification. IEEE Signal Processing Magazine 4, 33–41.

Ramsey, S.R., 1987. The Languages of China. Princeton University Press, Princeton, NJ.

Tsai, W.H., 1997. A study of speaker-independent Chinese dialect identification. Master's Thesis, National Chiao-Tung University, Taiwan, ROC.

Wellekens, C., 1987. Explicit time correlation in hidden Markov models for speech recognition. In: Proceedings ICASSP'87, pp. 384–386.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Processing 4, 31–44.