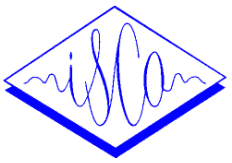

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

Discriminative Training of GMM for Language Identification.

Qu Dan Wang Bingxi
Department of Information Science of Information Engineering University
No. 306, P. O. Box 1001, Zhengzhou, 450002, China
Email: qudan@msn.com bingxiwang@163.com

Abstract

In this paper, a discriminative training procedure for a Gaussian Mixture Model (GMM) language identification system is described. The proposal is based on the Generalized Probabilistic Descent (GPD) algorithm and Minimum Classification Error Rates formulated to estimate the GMM parameters. The evaluation is conducted using the OGI multi-language telephone speech corpus. The experimental results show such system is very effective in language identification tasks.

1. Introduction

Automatic Language Identification (LID) is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. With expanding global partnership, there is an increasing demand for communications cross the boundaries of different languages. Automatic language identification has important applications in domains of information retrieval and military including automatic translation services, emergency services, etc[1]. In information services, multi-lingual services can be provided in information query, which must give users the multi-lingual cues to make a choice among the languages. Language identification systems must distinguish the user's language so as to offer services of multi-lingual category. Classical examples of this type of services include tourism information, emergency services as well as shopping, banking and stock deals. For example, AT&T offers a Language Line interpreter service to help police department to handle emergency calls. Automatic language identification can also be used as a pre-processing end of multi-lingual machine translation systems, and communication systems that can transform one language to another straightforwardly. With the advent of the information era and the popularization of the Internet, language identification shows ever-increasing applicable values, and as such highly effective studies have been conducting internationally.

Gaussian Mixture Models (GMM) have been shown to be effective for robust speaker-independent language identification[2] using conventional training methods based on maximum likelihood (ML) estimation. ML estimation is based on the training data from the same language, but do not take into account the data from other competing languages. 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 Minimum Classification Error (MCE) criterion. This paper describes an effective language identification system base on discriminative training of GMM.

2. Gaussian Mixture Model

Gaussian mixed model is essentially a multi-dimensional probability density function which attempts to model the probability density function of a feature vector, $\vec{x}$,by the weighted combination of multi -variate Gaussian densities:

$$p(x_t \mid l) = \sum_{i=1}^{M} p(l_i)\, p[x_t \mid l_i, m_i, \Sigma_i] \qquad (1)$$

where x_t is D-dimensional observing vector, $p(l_i)$, $i=1,2,...,M$ are mixing weights representing the probability of each Gaussian distribution, and $\sum_{i=1}^{M} p(l_i) = 1$, $p[x_t \mid l_i, m_i, \Sigma_i]$ is D- dimensional Gaussian distribution defined by the corresponding means m_i and covariance matrices Σ_i, namely:

$$p[x_t \mid l_i, m_i, \Sigma_i] = N(x_t, m_i, \Sigma_i) = \frac{1}{(2p)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x_t - m_i)^T \Sigma_i^{-1} (x_t - m_i)\right\} \qquad (2)$$

There are M Gaussian distributions in all with each represented by l_i , i is the mixture index, $i=1,...,M$. Each function is weighted by $p(l_i)$, and then summed to obtain the probabilistic distribution of x_t . l is the model described by $l = \{p_i, \vec{m}_i, \Sigma_i\}$.

The estimation of the GMM parameters is accomplished by an iterative process, termed the Expectation-Maximization (EM) [3]. For more rapid GMM convergence, the mixture means, weights and variances are seeded by statistics determined by a K-means vector quantization estimate of feature vectors.

During recognition, an unknown speech utterance, X, comprising of observations $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T$, is classified by first calculating the average log likelihood that the model produced the unknown speech utterance. This is given as:

$$p(X \mid l) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t \mid l) \tag{3}$$

where l is the GMM. The maximum-likelihood classifier hypothesis, H, can be calculated as

$$H = \arg\max_{l=1}^{L} p(X \mid l_l) \tag{4}$$

where the $L = \{1,2,\ldots, L\}$ is the model index.

3. Discriminative Training of GMM

This section addresses the implementation issues for discriminative estimation of entire parameter set in the context of a GMM-based language identification system. Each language L_j is characterized by a parameter set $\Phi_j = \{l_j\}$ which consists of GMM parameters $l_j = \{w_{j,n}, m_{jn}, \Sigma_{jn} \mid n = 1,2,\ldots,M\}$, $j = 1,2,\ldots,J$. It should be noted that j is the language index and n is the mixture index. The conventional method to estimate the model parameter is EM algorithm according to the Maximization Likelihood criterion (ML). However, the ML approach often does not lead to an optimum performance in classification tasks. An alternative approach to parameter estimation is based on Minimum Classification Error criterion (MCE). 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 GMM for language identification tasks.

Consider a set of training tokens with known language identities $O = \{O^{(l)}\}_{l=1}^{L}$, where each token $O^{(l)}$ is composed of cepstral features $X^{(l)}$ with length $T^{(l)}$. Based on 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, in which model parameters are adjusted iteratively to better represent the statistics of a training database. Below a specific implementation of the GPD algorithm for a GMM-based language identification system[4].

(1) Calculate a set of discriminant functions,

$$g_k(O^{(l)}; \Lambda) = \log p(X^{(l)} \mid l_k) \quad k = 1,2,\ldots, J \tag{5}$$

where discriminant function is the log likelihood score of a training token $O^{(l)}$ on the kth language model l_k.

(2) Calculate the misclassification measure for a training token $O^{(l)}$ from the language k.

$$M_k(O^{(l)}; \Lambda) = -g_k(O^{(l)}; \Lambda) + \log\left\{\frac{1}{J-1}\sum_{s, s\neq k} \exp[g_s(O^{(l)}; \Lambda)h]\right\}^{1/h} \tag{6}$$

where h is a positive real number.

(3) Define the smoothed loss function:

$$L_k(O^{(l)}; \Lambda) = \frac{1}{1 + e^{-g M(O^{(l)}; \Lambda) + r}} \tag{7}$$

where g and r are constants. In general, r is set to be zero, and the parameter g controls the function smoothness. From the Eq.(7), it can be seen that when $M_k(O^{(l)}; \Lambda)$ is far below than zero, it means the correct identification, and when $M_k(O^{(l)}; \Lambda)$ is greater than zero, it means the incorrect recognition. It is obvious that $L_k(O^{(l)}; \Lambda)$ is increasing simply with the $M_k(O^{(l)}; \Lambda)$, Thus the loss function is directly related with the performance of LID system, so the decrease of the loss function means the improvement of the system. MCE training algorithm is the optimizing process of the loss function defined in Eq. (7)

(4) To reduce the loss function, the GPD algorithm is used to adjust the GMM model parameters $l_j = \{w_{jn}, m_{jn}, \Sigma_{jn}\}_{n=1}^{M}$, Denoting any one of the weights $\{w_{jn}\}_{n=1}^{M}$, means $\{m_{jn}\}_{n=1}^{M}$ or covariances $\{\Sigma_{jn}\}_{n=1}^{M}$ by f_j, the new parameter becomes:

$$\bar{f}_j = f_j - e\sum_{l=1}^{L}\sum_{k=1}^{J} \frac{\partial L_k(O^{(l)}; \Lambda)}{\partial f_j^{(i)}} \tag{8}$$

where e is the step size and where

$$\frac{\partial L_k(O^{(l)}; \Lambda)}{\partial f_j} = \frac{\partial L_k(O^{(l)}; \Lambda)}{\partial M_k(O^{(l)}; \Lambda)} \frac{\partial M_k(O^{(l)}; \Lambda)}{\partial g_j(O^{(l)}; \Lambda)} \frac{\partial g_j(O^{(l)}; \Lambda)}{\partial f_j} \tag{9}$$

according to Eq.(5) and Eq. (6) ,

$$\frac{\partial L_k(O^{(l)}; \Lambda)}{\partial M_k(O^{(l)}; \Lambda)} = g\, L_k(O^{(l)}; \Lambda)[1 - L_k(O^{(l)}; \Lambda)] \tag{10}$$

$$\frac{\partial M_k(O^{(l)}; \Lambda)}{\partial g_j(O^{(l)}; \Lambda)} = \begin{cases} -1 & \text{if } j = k \\ \dfrac{\exp[g_j(O^{(l)}; \Lambda)]}{\sum_{s, s\neq k} \exp[g_s(O^{(l)}; \Lambda)]} & \text{if } j \neq k \end{cases} \tag{11}$$

(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 constrains such as the positive definiteness of the covariance matrix $\Sigma_{j,n}$ as well as the stochastic constraints $\sum_n w_{j,n} = 1$. This can be done by transforming these constrained parameters to be an unconstrained domain and then by computing the gradient with respect to the transformed parameters $\tilde{w}_{jn}$, $\tilde{m}_{jn} = \{\tilde{u}_{jnl}\}_{l=1}^{D}$, $\Sigma_{jn} = \left[s_{jnl}^2\right]_{l=1}^{D}$. The following parameter transformations allow us to maintain

the following constrains during parameter adaptation[5].

$$1)\ w_{jn} \to \tilde{w}_{jn} \text{ where } w_{jn} = \frac{e^{\tilde{w}_{jn}}}{\sum_m e^{\tilde{w}_{jm}}} \tag{12}$$

$$2)\ u_{jnl} \to \tilde{u}_{jnl} = \frac{u_{jnl}}{s_{jnl}} \tag{13}$$

$$3)\ s_{jnl} \to \tilde{s}_{jnl} = \log s_{jnl} \tag{14}$$

It can be shown that the gradient computations of individual parameters are of the form:

$$\frac{\partial g_j(O^{(l)};\Lambda)}{\partial \tilde{w}_{j,n}} = \sum_{t=1}^{T(l)} p(n|x_t^{(l)}, l_j)\left[1 - w_{j,n}\right] \tag{15}$$

$$\frac{\partial g_j(O^{(l)};\Lambda)}{\partial \tilde{m}_{j,n}} = \sum_{t=1}^{T(l)} \left[p(n|x_t^{(l)}, l_j)\, \Sigma_{j,n}^{-1/2}\left(x_t - u_{jn}\right)\right] \tag{16}$$

$$\frac{\partial g_j(O^{(l)};\Lambda)}{\partial \tilde{\Sigma}_{j,n}} = \sum_{t=1}^{T(l)} p(n|x_t^{(l)}, l_j)\left\{\Sigma_{j,n}^{-1}[x_t^{(l)} - m_{j,n}][x_t^{(l)} - m_{j,n}]' - I\right\} \tag{17}$$

where, I denotes an $M \times M$ identity matrix and $p(n|x_t^{(l)}, l_j)$ above is defined as Eq.(18).

$$p(n|x_t^{(l)}, l_j) = \frac{w_{jn} N(x_t, m_{jn}, \Sigma_{jn})}{\sum_{q=1}^{M} w_{jq} N(x_t, m_{jq}, \Sigma_{jq})} \tag{18}$$

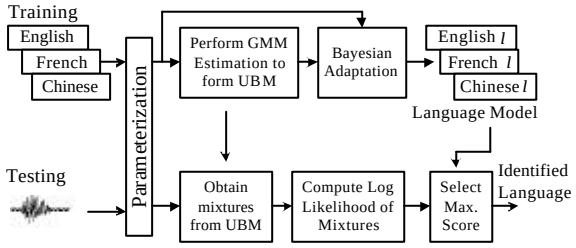

Figure 1 Language identification system using GMM-UBM

4. Reference Experimental System

To evaluate the performance of such system, a reference language identification system base on GMM and ML criterion: language ID system using Gaussian Mixture Model and Universal background Model (GMM-UBM) shown in figure 1. UBM denotes a language- independent distribution. During the training stage, a UBM is generated using all the training data from different languages. Bayesian adaptation algorithm is used to obtain each language model. During testing, calculate the log likelihood of each token and identify the language being used.

5. Experiments and Results

5.1 Experimental speech corpus

The OGI Multi-language Telephone Speech Corpus (OGI-TS) [6]is designed to support research on automatic language identification and multi-language speech recognition. By far, it consists of fluent speech data with fixed glossaries including English, French, Farsi, German, Japanese, Korean, Mandarin Chinese, Spanish, Hindi, Tamil, and Vietnamese. The speech data were produced by more than 100 different native speakers in each of 11 languages through telephone lines. The duration of utterances ranges from 1 to 50 seconds, and the average duration is 13.4 seconds. The choice of languages is based on various factors. These languages represent a range of unrelated languages (e.g. Vietnamese and Tamil and German) as well as languages from the same sub-family (e.g., Germanic languages such as English and German, Romanic languages such as French and Spanish). The languages also include various prosodic features (e.g., Chinese and Vietnamese are tonal languages, Japanese uses pitch-accents and syllabic mora). So the choices of these languages represent a relatively broad expanse. In addition, the corpus content covers 10 aspects of ordinary lives showing its comprehensiveness.

5.2 Experimental results

Experimental data derive from OGI-TS speech corpus mentioned above. Speech signals are sampled at 8kHz with 16 bit resolution. Performances of the proposed approach and reference system were evaluated on the speech signals from 3 languages: English, French and Mandarin Chinese. For each language, 50 speakers are selected as the training set, and the duration of each speaker is about 72 second. The speech are firstly processed to remove the silence parts and then pre-emphasized with the filter $H(z) = 1 - 0.95z^{-1}$. The feature vectors used consist of 15 Perceptual linear prediction coefficients (PLP). Each feature vector is extracted at 16 ms intervals using a 32 ms window of telephone bandwidth speech. Since the experiment involved telephone speech, cepstral mean subtraction is applied to the PLP to reduce the linear channel effects. The corresponding delta coefficients are computed over a window length of 5 frames. Finally the delta coefficients are appended to the features. So for each frame, a 30-dimentional feature vector is calculated.

For the reference system, all the speech data from 150 speakers in 3 languages are used to generate UBM via EM algorithm base on ML criterion. Through Bayesian adaptation, each hypothesized language model is achieved.

During recognition, the difference between score of the hypothesized model and that of the UBM is the score of the test utterance.

For the discriminative training system, the discriminant function and the misclassification measure are calculated for each token from the 150 speakers in 3 languages. The loss function is iteratively adjusted to be minimal to obtain all model parameters. During the MCE training phase, the parameter values used for g , h and the maximum number of iterations N_m were empirically determined to be 0.1, 2.0 and 30, respectively. Additionally, the step size at k th iteration was determined by $e = 0.01/(1-k/N_m)$.

The performance for closed test is shown in table 1 where all test speeches are from the original 150 speakers in 3 languages.

Table 1 LID results for closed test using MCE algorithm

	Language Identified					Ave
Input	En	Ma	Fr	All	ID Rate	ID Rate
En	401	52	27	480	83.54%	86.24%
Ma	21	418	37	476	87.82%	
Fr	41	20	422	483	87.37%	

The task of language identification is also evaluated on the test set. The test speech of English includes 855 speech utterances from other 89 speakers which are different from those of training set. Similarly, Mandarin Chinese and French consist of 202 segments from other 36 speakers and 305 speech segments from other 38 speakers, respectively. And the open test identification rates for English, French and Chinese are 75.79%, 72.28% and 78.36% respectively, as shown in table 2, where the mixture number of GMM is 256. The further experimental results also show the much the mixture number is, the better the recognition is at the cost of more complex computation. So the further improvement for recognition can be achieved at the cost of more complex computation.

Table 2 LID results for open test using MCE algorithm

	Language Identified					Ave
Input	En	Ma	Fr	All	ID Rate	ID Rate
En	648	94	113	855	75.79%	75.47%
Ma	18	146	38	202	72.28%	
Fr	26	40	239	305	78.36%	

The effectiveness of using the MCE method for discriminative estimation of GMM parameters is clearly demonstrated in Table 2. GMM-UBM represents the reference LID system GMMDT represents GMM discriminative training algorithm based on MCE criterion. The results of experiments indicates that the MCE algorithm are able to help in distinguishing between languages with greater accuracy. In our system, the improvement rate is 3.24 %

Table 3 Comparison between the two LID system

Method	Ave LID rate (%)	Improvement (%)
GMM-UBM	73.10	
GMMDT	75.47	3.24

6. Conclusion

This study discusses method of incorporating out-of-class information directed into language identification system. This task is accomplished by using MCE criterion and GPD algorithm. The overall language model parameters are achieved through minimizing the misclassification errors among different languages. The performance of such system is evaluated using English, French and Mandarin Chinese from OGI-TS. The average identification rate based on MCE discriminative training is 75.47% with 3.24% superior to original GMM-UBM based on ML criterion. Experimental results show discriminative training of GMM is very effective in language identification.

7. Acknowledgements

We thank Doctor Wei-Ho Tsai for valuable communications about the MCE aspects of this paper and several valuable ideas to this research.

Reference

[1] Y. K. Muthusamy, E. Barnard and R. A. Cole, "Reviewing Automatic Language Identification", IEEE Signal Processing Magazine, October 1994.

[2] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Processing, vol. 4, pp. 31-44, 1996

[3] D.A. Reynolds, and R.C. Rose, Rosust text-independence speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Processing, vol.3, No.1, pp72-83.

[4] W. H. Tsai, W. W. Chang, "Discriminative training of Gaussian mixture bigram models with applications to Chinese dialect identification", Speech Comm., Vol. 36, pp. 317-326, 2002.

[5] B. H. Juang, W. Chou, C. H. Lee, "Minimum classification error rate methods for speech recognition", IEEE Trans. Speech Audio Processing, Vol.5, pp257-265.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI Multi-language telephone speech corpus. Technical report, Center for Spoken Language Understanding Oregon Graduate Institute of Science and Technology, Portland, 1993.