RC24953 (W1003-002) March 1, 2010
Other

IBM Research Report

Training Universal Background Models for Speaker Recognition

Mohamed Kamal Omar, Jason Pelecanos
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Training Universal Background Models for Speaker Recognition

Mohamed Kamal Omar and Jason Pelecanos

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{mkomar,jwpeleca}@us.ibm.com

Abstract

Universal background models (UBM) in speaker recognition systems are typically Gaussian mixture models (GMM) trained from a large amount of data using the maximum likelihood criterion. This paper investigates three alternative criteria for training the UBM. In the first, we cluster an existing automatic speech recognition (ASR) acoustic model to generate the UBM. In each of the other two, we use statistics based on the speaker labels of the development data to regularize the maximum likelihood objective function in training the UBM. We present an iterative algorithm similar to the expectation maximization (EM) algorithm to train the UBM for each of these regularized maximum likelihood criteria. We present several experiments that show how combining only two systems outperforms the best published results on the English telephone tasks of the NIST 2008 speaker recognition evaluation.

1. Introduction

Improved user security in speech-driven telephony applications can be achieved with automatic speaker verification. Current automatic speaker verification systems face significant challenges caused by adverse acoustic conditions. Telephone band limitation, channel/transducer variability, as well as natural speech variability have a negative impact on the performance of speaker verification systems. Degradation in the performance of these systems due to inter-session variability has been one of the main challenges to the deployment of speaker verification technologies. We investigate in this work how integrating more information about the development and test sets into the speaker recognition system may improve its performance and robustness.

In this work, we propose two main approaches for training the UBM. In the first, the UBM is constructed by using the Kullback-Leibler (KL) distance as a measure for clustering the Gaussian components of an ASR acoustic model. This approach attempts to exploit the context-dependent phonetic information of the ASR acoustic model in estimating the UBM parameters. Subsequently, this method is called the phonetically inspired UBM (PIUBM) approach. The approach is motivated by the fact that many of the speaker characteristics are conditioned on some phonetic units or phonetic classes and therefore may be better modeled using a UBM trained with an explicit modeling of these units and classes. Examples of using ASR systems for speaker recognition include modeling the speakers using maximum likelihood linear regression (MLLR) transforms generated by an ASR model [1] and using counts of n-gram words or phones generated from an ASR transcription of the audio [2, 3, 4].

In the second approach, we examine two discriminative regularizations of the maximum likelihood objective function for estimating the UBM parameters. Most speaker verification systems use maximum likelihood estimation (MLE) or Bayesian methods to estimate the parameters of the UBM. The popularity of MLE is attributed to the existence of efficient algorithms to implement it, such as the expectation-maximization (EM) algorithm [5]. It is also attributed to its consistency and asymptotic efficiency, if the true probability density function (PDF) belongs to the admissible set of parameterized PDF models [6]. However, as we do not know the true PDF, we can not guarantee a small approximation error. A small approximation error can be achieved by using a complex structure of the hypothesized models that can approximate a large set of PDFs. On the other hand, this increases the computational and conceptual complexity of the system, increases the required amount of training data to obtain a robust estimate of the model parameters, and deteriorates the generalization ability of the model. Discriminative training offers an alternative that estimates the model parameters to optimize an estimate of the training data recognition error. Unlike maximum likelihood estimation of the UBM parameters, optimizing a discriminative criterion can be made directly related to any weighted sum of the false alarm and the miss probabilities such as the Equal Error Rate (EER) or the minimum Detection Cost Function (DCF). Discriminative training has been used in SVM-based speaker recognition systems [7]. However, in GMM-based SVM speaker recognition systems, the UBM parameters are estimated using the EM algorithm to maximize the likelihood of the training data. One of the prominent examples of discriminative training of the UBM parameters is using the Maximum Mutual Information (MMI) criterion to optimize the UBM parameters of an automatic language identification system. This discriminatively trained UBM [8] provided noteworthy improvements compared to the MLE UBM. In this work, we integrate information about the speakers of the development set into the objective functions used for training the UBM discriminatively and describe efficient iterative algorithms to estimate the UBM parameters.

In the next section, we describe the main architecture of the speaker verification system used in this work. In Section 3, we formulate the problem and describe our objective criterion for the PIUBM approach. In Section 4, the details of estimating the UBM parameters to optimize the two regularized maximum likelihood objective functions are described. The experiments performed to evaluate the performance of the systems are described in Section 5. Finally, Section 6 contains a discussion of the results and future research.

2. The speaker verification system

In this work, the speaker recognition systems are based on the use of GMM supervectors. These GMM supervectors are formed from the concatenation of the MAP [9, 10] adapted means that are normalized according to a mapping proposed in [7]. Nuisance Attribute Projection (NAP) [7] is applied to remove supervector directions that correspond to large intra-speaker variability. In all the systems reported in this work, 128 nuisance directions were removed. These nuisance directions, as per our submission in the NIST 2008 speaker recognition evaluation [11], are based on the principal components extracted from the average within-class covariance matrix [12].

Before score normalization, the output scores of the speaker verification systems can be represented by some kind of generalized inner product of two vectors representing the verification and the enrollment utterances [7]. This can be described by the relation

$$s = \Phi_e^T \mathbf{K} \Phi_v, \tag{1}$$

where Φ_e is the supervector representing the enrollment utterance, Φ_v is the supervector representing the verification utterance, $\mathbf{K}$ is the NAP projection matrix, and s is the score corresponding to this pair of utterances. Both Φ_e and Φ_v are vectors in a high dimensional space of dimension equal to the product of the feature vector dimension and the number of Gaussian probability density functions in the UBM.

For each utterance, the mean based supervector is generated by concatenating functions of the adapted Gaussian means into a supervector. A GMM with K mixture components is used to construct the high-dimensional supervectors for the enrollment utterance, Φ_e, and the verification utterance, Φ_v. These supervectors are constructed as follows

$$\Phi_k = \sqrt{w_k}\boldsymbol{\Sigma}_k^{-\frac{1}{2}}\left(\mu_k^{adapt} - \mu_k^{ubm}\right), \tag{2}$$

$$\Phi = \left[\Phi_1^T \Phi_2^T \ldots \Phi_K^T\right]^T, \tag{3}$$

where w_k is the weight of the kth Gaussian component in the GMM, μ_k^{adapt} is the MAP adapted mean for this component, μ_k^{ubm} is the universal background model (UBM) mean for this component, and $\boldsymbol{\Sigma}_k$ is the diagonal covariance matrix of the kth Gaussian component in the GMM. We use the single iteration MAP adaptation presented by Reynolds [10] to generate the utterance-specific adapted means, $\{\mu_k^{adapt}\}$, from the UBM means, $\{\mu_k^{ubm}\}$.

For all the systems reported in this work, the UBM consists of 1024 mixture components. The UBM of the baseline system is trained using Maximum Likelihood (ML) training [5, 13]. Both Z-Norm and T-Norm [14] score normalization approaches were applied separately for each gender. Further details about the various systems are described in the experiments section.

3. PIUBM approach

This approach was first applied to nonnative speaker and accent detection in [15]. It achieved the best published results on both tasks on the Fisher and the CSLU-FAE databases respectively. In this system, the UBM is estimated directly from the acoustic model of the ASR system by using K-means clustering. A symmetric variant of the Kullback-Leibler (KL) distance between two Gaussian components is used as a distance measure in the K-means clustering algorithm to achieve the final clustering of the ASR acoustic model to a UBM of 1024 Gaussian components. This novel method for UBM construction is applied to ASR acoustic models trained in the feature-based minimum phone error (FMPE) feature space [16].

The process of K-Means model training, in this context, consists of specifying a set of K Gaussian components, $U = (u_1, u_2, \ldots, u_K)$, that minimize the average distortion d of C Gaussian components, $G = (g_1, g_2, \ldots, g_C)$, which correspond to the ASR acoustic model. The average distortion is specified by

$$d = \frac{1}{C}\sum_{c=1}^{C}\min_{k=1}^{K} d(u_k, g_c), \tag{4}$$

where $d(u_k, g_c) = KL(u_k, g_c) + KL(g_c, u_k)$, and $KL(u_k, g_c)$ is the KL distance between u_k and g_c.

To minimize the average distortion in each iteration of the K-means clustering, the update equations for the mean and the variance for each dimension of each Gaussian component of the UBM are

$$\hat{\mu}_{kf} = \frac{1}{N_k}\sum_{n\in S_k} m_{nf}, \tag{5}$$

$$\hat{\sigma}_{kf}^2 = \sqrt{\frac{\sum_{n\in S_k}\left(v_{nf} + (\mu_{kf} - m_{nf})^2\right)}{\sum_{n\in S_k}\frac{1}{v_{nf}}}}, \tag{6}$$

where S_k is the set of indices of ASR Gaussian components assigned to the kth Gaussian of the UBM, N_k is the size of this set, μ_{kf} is the current mean for the fth dimension of the kth Gaussian of the UBM, m_{nf} and v_{nf} are the mean and variance respectively for the fth dimension of the nth ASR Gaussian. The weight of each UBM Gaussian is set equal to the normalized number of ASR Gaussian components assigned to it.

4. Regularized ML training approach

The UBM in speaker verification systems is typically a Gaussian mixture model (GMM) trained on a large amount of data using the EM algorithm. In this work, two methods for training the UBM are investigated. In both methods, the UBM parameters are estimated by adding a regularization term to the maximum likelihood objective function. In the first method, the UBM parameters are trained using an objective function that favors a sparse representation for each speaker in the training data. In the second, the regularization term favors larger values for target trial scores and smaller values for imposter trial scores. In the following, we discuss the two approaches in detail.

4.1. Sparse speaker representation approach

Estimating the UBM parameters using maximum likelihood training does not take into consideration the available speaker labels of the training data. In this approach, we add a regularization term to the likelihood objective function to ensure the sparsity of the speaker supervector representation. The parameters of the UBM are updated using an EM-like algorithm to maximize the regularized maximum likelihood objective function. To ensure the sparsity of the supervector representation for each speaker in the training data, the average of the supervector representation of the utterances of the speaker is used to represent the speaker. Increasing the sparsity of the speaker representation is equivalent to minimizing the l_0 norm of the speaker supervector. The l_0 norm of a vector is equal to the number of non-zero elements of the vector. The l_0 norm of a speaker supervector, Φ_q, is

$$\|\Phi_q\|_0 = \sum_{y=1}^{Y} f(\phi_{qy}), \tag{7}$$

where

$$f(z) = \begin{cases} 0 & z = 0 \\ 1 & \text{otherwise} \end{cases},$$

and Y is the dimension of the supervector representation which equals the product of the number of Gaussian components in the UBM and the feature vector dimension.

Given the supervector representation in Equations 2 and 3, it can be shown that minimizing the l_0 norm of the speaker's supervector representation is equivalent to maximizing the l_2 norm of the vector of estimates of the expected posterior probabilities of the Gaussian components given the speaker training data which is given by

$$\|\gamma_q\|_2^2 = \sum_{k=1}^{K} {\gamma_q^k}^2, \tag{8}$$

where K is the number of Gaussian components in the UBM, $\gamma_q^k = \frac{1}{N_q}\sum_{i=1}^{N_q}\gamma_{qi}^k$ is an estimate of the expected posterior probability of the kth Gaussian component of the UBM given an observation from speaker q, $\gamma_{qi}^k = P(k|X_{qi})$ is the posterior probability of the kth Gaussian component of the UBM given the ith observation of speaker q, N_q is the number of observations from speaker q. Adding the sum over all speakers of the l_2 norm of the expected posterior probabilities of the Gaussian components given the speaker training data as a regularization term, the objective function to be maximized is

$$O = L + \lambda \sum_{q=1}^{Q}\sum_{k=1}^{K} {\gamma_q^k}^2, \tag{9}$$

where $\lambda > 0$, L is the log likelihood of the training data, λ is the regularization parameter, and Q is the number of training speakers.

We use an iterative algorithm similar to the EM algorithm to estimate the UBM parameters that maximize the objective function in Equation 9. It can be shown that the update equations for the mean and the variance for each dimension of each Gaussian component of the UBM are

$$\hat{\mu}_{kf} = \frac{\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\beta_{qi}^k x_{qif}}{\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\beta_{qi}^k}, \tag{10}$$

$$\hat{\sigma}_{kf}^2 = \frac{\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\beta_{qi}^k \left(x_{qif} - \mu_{kf}\right)^2}{\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\beta_{qi}^k}, \tag{11}$$

where

$$\beta_{qi}^k = \gamma_{qi}^k \left(1 + \lambda\left(\gamma_q^k - \sum_{l=1}^{N_q}\sum_{g=1}^{K}\gamma_q^g\gamma_{ql}^g\right)\right), \tag{12}$$

and x_{qif} is the fth dimension of the ith observation vector of speaker q in the development data. It can be shown also that the update equation for the weight of each Gaussian component of the UBM is

$$\hat{w}_k = \frac{\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\beta_{qi}^k}{\sum_{g=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\beta_{qi}^g}. \tag{13}$$

This approach is called the sparse speaker representation (SSR) approach in the following sections.

4.2. Discriminative regularization approach

Estimating the UBM parameters using maximum likelihood training does not directly target reducing the speaker verification errors on the training data. In this approach, we add a regularization term to the log-likelihood objective function to reduce the value of the imposter scores and increase the value of the target scores. The parameters of the UBM are updated using an EM-like algorithm to maximize the regularized maximum likelihood objective function

$$O = L - \lambda_t \sum_{r=1}^{T} e^{a_t - b_t s_{tr}} - \lambda_p \sum_{j=1}^{J} e^{a_p + b_p s_{pj}}, \tag{14}$$

where $\lambda_t > 0$, $\lambda_p > 0$, λ_t is the target-trials regularization parameter, λ_p is the imposter-trials regularization parameter, s_{tr} is the rth target score, s_{pj} is the jth imposter score, a_t, b_t are the parameters of the target regularization function, a_p, b_p are the parameters of the imposter regularization function, T is the number of target scores, and J is the number of imposter scores. The parameters of the target and imposter regularization functions are estimated on a held-out set to provide proper conditioning of the target and imposter scores respectively. In the experiments reported here, we used the same value for both λ_t and λ_p which is double the value that ensures all the variances of the UBM Gaussian components are positive. Also the target and imposter scores are the speaker recognition scores without NAP compensation and without ZT normalization. We investigated using the NAP-compensated and ZT-normalized scores in the objective function but we keep the discussion and the results in this work to the simpler case of scores without NAP compensation and without ZT normalization.

We use an iterative algorithm similar to the EM algorithm to estimate the UBM parameters that maximize the objective function in Equation 14. It can be shown that the update equations for the mean and the variance for each dimension of each Gaussian component of the UBM are

$$\hat{\mu}_{kf} = \frac{\sum_{u=1}^{U}\sum_{i=1}^{N_u}\left(\alpha_{ui}^k - w_k B_{uk} A_{uki}\right) x_{uif}}{\sum_{u=1}^{U}\sum_{i=1}^{N_u}\left(\alpha_{ui}^k - w_k B_{uk} A_{uki}\right)}, \tag{15}$$

$$\begin{aligned}\hat{\sigma}_{kf}^2 &= \frac{\sum_{u=1}^{U}\sum_{i=1}^{N_u}\alpha_{ui}^k\left(x_{uif} - \mu_{kf}\right)^2}{\sum_{u=1}^{U}\sum_{i=1}^{N_u}\alpha_{ui}^k} \\ &\quad - \frac{w_k \sum_{u=1}^{U} D_{ukf} E_{ukf}}{\sum_{u=1}^{U}\sum_{i=1}^{N_u}\alpha_{ui}^k},\end{aligned} \tag{16}$$

where

$$\alpha_{ui}^k = \gamma_{ui}^k \left(1 + \left(H_u^k - \sum_{l=1}^{N_u}\sum_{g=1}^{K} H_u^g \gamma_{ul}^g\right)\right), \tag{17}$$

$$A_{uki} = \frac{\gamma_{ui}^k}{\sum_{i=1}^{N_u}\gamma_{ui}^k + R}, \tag{18}$$

$$A_{uk} = \sum_{i=1}^{N_u} A_{uki}, \tag{19}$$

$$\begin{aligned}B_{uk} &= \sum_{v \in U_{q_u}} b_t \lambda_t e^{a_t - b_t s_{uv}} A_{vk} \\ &\quad - \sum_{v \notin U_{q_u}} b_p \lambda_p e^{a_p + b_p s_{uv}} A_{vk},\end{aligned} \tag{20}$$

$$\begin{aligned} D_{ukf} &= \frac{\sum_{i=1}^{N_u} \gamma_{ui}^k \left(x_{uif} - \mu_{kf}\right)}{\sum_{i=1}^{N_u} \gamma_{ui}^k + R}, & (21) \\ E_{ukf} &= \sum_{v \in U_{q_u}} b_t \lambda_t e^{a_t - b_t s_{uv}} D_{vkf} \\ &\quad - \sum_{v \notin U_{q_u}} b_p \lambda_p e^{a_p + b_p s_{uv}} D_{vkf}, & (22) \\ H_u^k &= \frac{w_k}{\sum_{i=1}^{N_u} \gamma_{ui}^k + R} \\ &\quad \sum_{i=1}^{N_u} \sum_{f=1}^{F} \frac{E_{ukf}}{\sigma_{kf}^2} \left(x_{uif} - \mu_{kf} - D_{ukf}\right), & (23) \end{aligned}$$

x_{uif} is the fth dimension of the ith observation vector of the development data for utterance u, U is the total number of utterances in the training data, N_u is the total number of observations for utterance u, γ_{ui}^k is the posterior probability of the kth Gaussian component given the ith observation vector of utterance u, U_{q_u} is the set of all other utterances belonging to the speaker q_u which utterance u belongs to, F is the dimension of the feature vector, and R is the map adaptation relevance factor. It can be shown also that the update equation for the weight of each Gaussian component of the UBM is

$$\hat{w}_k = \frac{\sum_{u=1}^{U} \sum_{i=1}^{N_u} \alpha_{ui}^k}{\zeta - \frac{w_k}{2} \sum_{u=1}^{U} \sum_{f=1}^{F} \frac{D_{ukf} E_{ukf}}{\sigma_{kf}^2}}, \qquad (24)$$

where the value of ζ is estimated to satisfy the constraint $\sum_{k=1}^{K} w_k = 1$. This approach is called the discriminative regularization (DR) approach in the following sections.

5. Experiments

The three previously discussed methods to train the UBM parameters were evaluated on the English tasks of the core condition of the NIST 2008 Speaker Recognition Evaluation (SRE) [11] and compared to the MLE UBM baseline systems.

The development data set consists of a combination of audio from the NIST 2004 speaker recognition database, the Switchboard II Phase III corpora, the NIST 2006 speaker recognition database, and the NIST 2008 interview development set. The collection contains 13770 utterances: 6038 utterances of male speakers and 7732 of female speakers. The total number of speakers in the development data is 1769 speakers: 988 female speakers and 781 male speakers. The development set or a subset of it was used to estimate the UBM parameters, to estimate the expected within-class covariance matrix over all speakers for NAP compensation, as well as for gender-dependent ZT-norm score normalization.

5.1. Baseline system

The front-end features consist of 36 dimensional features forged from 12 cepstral coefficients and their corresponding delta and delta-delta features. There are 24 filters in the filter bank, over a frequency range of 125-3800 Hz, used to generate these cepstral coefficients with a 32ms window and a 10ms frame shift. Feature warping is applied to the resulting feature vectors [17] to reduce linear channel and slowly varying additive noise effects. Each utterance in both the training and the testing data is represented by a GMM mean based supervector of dimension 36864. This representation was generated using a UBM of 1024 Gaussian components by MAP adaptation. The system performance was measured at two operating points, namely in terms of the Equal-Error Rate (EER) and the minimum Detection Cost Function (DCF) as defined in the evaluation plan [11].

In all experiments, we used the GMM-based setup described in Section 2 which generates a score for each pair of utterances using the inner product of the corresponding GMM based mean supervectors after applying NAP compensation to the supervectors. ZT-normalization is applied to these scores to generate the final scores.

5.2. PIUBM system

In the first set of experiments, the 1024 Gaussian component UBM for the baseline is trained using the whole 13770-utterance development set. On the other hand, for the PIUBM system, the UBM is generated by clustering the 250K Gaussian components of the English telephone conversational ASR acoustic model to 1024 Gaussian components. In the following, we describe the ASR system and then details about the experimental setup.

5.2.1. ASR system overview

The 40-dimension features for the IBM ASR system are estimated from sequences of 13-dimensional perceptual linear prediction (PLP) features by using a linear discriminant analysis (LDA) projection, and then applying a maximum likelihood linear transformation (MLLT). The acoustic model consists of 250K diagonal-covariance Gaussian components. In the context of speaker-adaptive training, vocal tract length normalization (VTLN) and feature-space maximum likelihood linear regression (FMLLR) are used. For the feature-based minimum phone error (FMPE) baseline, an FMPE transform is applied on top of the utterance-specific FMLLR transforms. A single pass of MLLR adaptation is also performed. The language model is a 72K-vocabulary interpolated back-off 4-gram language model.

5.2.2. Testing setup

Three systems are compared in this set of experiments: the baseline system using the 36 MFCC-based speaker recognition frontend, a system with a UBM trained on the ASR FMPE frontend using the EM algorithm, and the PIUBM which uses a UBM generated from the ASR acoustic models using the clustering algorithm described in Section 3. The MLE 1024 Gaussian component UBM for both the baseline and the ASR frontend systems are trained using the whole 13770-utterance development set. The results are reported on the English tasks of the core condition of the NIST 2008 speaker recognition evaluation. The description of these tasks is provided in Table 1. As shown in Table 2, the performance of the two systems which use the ASR frontend features outperform the baseline system on the Int-Tel, the Tel-Mic, and the Int-Int-S tasks. The results in Table 2 show also that the PIUBM system outperforms the other two systems significantly on the Tel-US and Tel-Eng tasks.

5.3. Regularized maximum likelihood systems

Our focus in this set of experiments is on the Int-Int-All interview task of the NIST08 evaluation. We achieved the best baseline system, for the system architecture described in Section 2, on this task by using only the NIST 2008 development data for training the UBM parameters and the whole 13770-utterance development data for estimating the NAP rejected subspace and the gender-dependent ZT score normalization. As shown in the baseline results of Tables 2 and 3, significant gains on the interview tasks in both EER and minimum DCF are achieved

Task	Description
Int-Int-All	Interview speech in training and test.
Int-Int-S	Interview speech from the same (lapel) microphone in training and test.
Int-Int-D	Interview speech from different microphones in training and test.
Int-Tel	Interview speech in training and telephone speech in test.
Tel-Mic	Telephone speech in training and telephone microphone speech in test.
Tel-Eng	English language telephone speech in training and test (any variety).
Tel-US	English language telephone speech spoken by a native US English Speaker in training and test.

Table 1: Description of the English NIST 2008 core condition evaluation tasks reported in our experiments.

	Performance minDCF (x10^3) and EER (%) (in parentheses)						
System	Int-Int-All	Int-Int-S	Int-Int-D	Int-Tel	Tel-Mic	Tel-Eng	Tel-US
Baseline	23.9 (4.6)	2.0 (0.8)	24.2 (4.6)	37.5 (10.3)	28.8 (7.4)	15.6 (3.5)	15.4 (4.4)
ASR Frontend	21.4 (4.8)	0.7 (0.4)	22.2 (5.0)	31.8 (7.6)	21.8 (6.4)	16.4 (3.4)	15.4 (4.4)
PIUBM	23.4 (5.3)	1.7 (0.3)	24.5 (5.5)	30.7 (8.6)	22.1 (6.7)	12.7 (2.7)	11.6 (3.0)

Table 2: The results on NIST 2008 English core condition tasks comparing the baseline system with systems utilizing ASR features and the PIUBM.

by using the NIST 2008 interview development set for training the UBM instead of the whole 13770-utterance development set. To keep the comparisons fair, only the NIST 2008 development data is used for training the UBM parameters of the sparse speaker representation and the discriminative regularization systems. As shown in Table 3, no significant gains in EER and minimum DCF on the Tel-Eng and the Tel-US tasks are obtained by using either the sparse speaker representation or the discriminative regularization objective functions to train the UBM parameters. This can be explained by the fact that the NIST 2008 development data, which was used for estimating the UBM parameters, is interview data only and did not have any telephone utterances. On the other hand, significant gains in EER and minimum DCF are obtained by using either the sparse speaker representation or the discriminative regularization objective functions to train the UBM parameters on all other tasks which have interview data or microphone data as shown in Table 3. Table 3 shows also that the discriminative regularization system outperforms the sparse speaker representation system on all tasks that have interview data or microphone data. The results on the mixed condition tasks of Int-Tel and Tel-Mic in Table 3 show more significant gains for the DR system versus the SSR system compared to the results on the interview tasks.

5.4. Combination of the two approaches

In this set of experiments, we combine with equal weights the scores of the PIUBM system and the discriminative regularization system. As shown in Table 4, significant improvements on the three interview tasks are obtained by combining the two systems. As far as we know, the result on the Int-Int-S task is the best result published on this task.

The results in Table 4 show also significant improvements on the telephone and mixed condition tasks from the combination of the two systems. As far as we know, the results on the telephone tasks of Tel-Eng and Tel-US are significantly better in both EER and minimum DCF than the best published combination results on these tasks. The fact that these results are achieved by combining two systems only may be attributed to the use of diverse systems with different features, UBM training data, and objective functions for training the UBM parameters.

6. Conclusions

The sparse speaker representation system consistently outperforms the baseline system on the English NIST 2008 core condition tasks. The discriminative regularization system also consistently outperforms the baseline system with a maximum likelihood UBM on the same tasks. In both cases, the improvement is achieved by integrating information about the development speakers into the estimation of the UBM parameters. Integrating context-dependent phonetic information into the training of the UBM parameters is demonstrated to be useful as well but at the expense of making the speaker recognition system language-dependent. In the PIUBM system, estimating the UBM parameters using an ASR telephone English acoustic model provided the best single-system performance on the telephone tasks of the NIST 2008 evaluation task. Combining the PIUBM system with the discriminative regularization system at the score level gives the best published performance on the English telephone tasks of the NIST 2008 evaluation and significant gains on the other tasks compared to the individual systems. We plan to integrate together the information about the training speakers and the context-dependent speech units in training the UBM parameters and compare the results to using either information by itself as reported in this work.

7. References

[1] A. Stolcke, S. Kajarekar, L. Ferrer, and E. Shriberg, “Speaker recognition with session variability normalization based on MLLR adaptation transforms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, 2007.

[2] G. Doddington, “Speaker recognition based on idiolectal differences between speakers,” in *Eurospeech*, 2001, vol. 4, pp. 2521–2524.

[3] M. Kohler, W. Andrews, J. Campbell, and J. Hernandez-Cordero, “Phonetic refraction for speaker recognition,” in

	Performance minDCF (x10^3) and EER (%) (in parentheses)						
System	Int-Int-All	Int-Int-S	Int-Int-D	Int-Tel	Tel-Mic	Tel-Eng	Tel-US
Baseline	19.4 (4.2)	2.9 (0.9)	19.3 (4.1)	37.5 (7.9)	32.3 (7.7)	15.6 (3.5)	16.4 (4.1)
SSR	16.2 (3.0)	2.2 (0.7)	16.4 (3.1)	34.4 (7.2)	26.8 (7.1)	13.9 (3.4)	14.2 (3.9)
DR	15.9 (2.7)	1.9 (0.6)	16.1 (2.8)	29.1 (7.1)	25.6 (7.2)	14.0 (3.4)	14.1 (4.1)

Table 3: The results on the English NIST 2008 core condition tasks comparing the baseline system with the sparse speaker representation (SSR) and discriminative regularization (DR) systems.

	Performance minDCF (x10^3) and EER (%) (in parentheses)						
System	Int-Int-All	Int-Int-S	Int-Int-D	Int-Tel	Tel-Mic	Tel-Eng	Tel-US
PIUBM	23.4 (5.3)	1.7 (0.3)	24.5 (5.5)	30.7 (8.6)	22.1 (6.7)	12.7 (2.7)	11.6 (3.0)
DR	15.9 (2.7)	1.9 (0.6)	16.1 (2.8)	29.1 (7.1)	25.6 (7.2)	14.0(3.4)	14.1 (4.1)
Combination	13.7 (2.7)	1.3 (0.3)	14.2 (2.7)	20.3 (5.1)	15.5 (4.7)	9.7 (2.1)	9.3 (2.1)

Table 4: The results on the English NIST 2008 core condition tasks comparing the individual systems of the PIUBM and the discriminative regularization (DR) to their combination.

Workshop on Multilingual Speech and Language Processing, 2001.

[4] D. Klusacek, J. Navratil, D. Reynolds, and J. Campbell, "Conditional pronunciation modeling for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[6] H. Poor, *An Introduction to Signal Detection and Estimation*, New York: Springer-Verlag, 1994.

[7] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[8] L. Burget, P. Matejka, and J. Cernocky, "Discriminative training techniques for acoustic language identification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 209–212, 2006.

[9] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[10] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[11] National Institute of Standards and Technology, "NIST speech group website," *http://www.nist.gov/speech*, 2008.

[12] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," *International Conference on Spoken Language Processing*, 2006.

[13] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[14] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[15] M. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[16] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 961–964, 2005.

[17] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 213–218, 2001.