Large Margin GMM for discriminative speaker verifi cation

Reda Jourani, Khalid Daoudi, Régine André-Obrecht, Driss Aboutajdine

Multimed Tools Appl manuscript No.
(will be inserted by the editor)

Large Margin GMM for discriminative speaker verification

Reda Jourani · Khalid Daoudi · Régine André-Obrecht · Driss Aboutajdine

Received: date / Accepted: date

Abstract Gaussian mixture models (GMM), trained using the generative criterion of maximum likelihood estimation, have been the most popular approach in speaker recognition during the last decades. This approach is also widely used in many other classification tasks and applications. Generative learning in not however the optimal way to address classification problems. In this paper we first present a new algorithm for discriminative learning of diagonal GMM under a large margin criterion. This algorithm has the major advantage of being highly efficient, which allow fast discriminative GMM training using large scale databases. We then evaluate its performances on a full NIST speaker verification task using NIST-SRE'2006 data. In particular, we use the popular Symmetrical Factor Analysis (SFA) for session variability compensation. The results show that our system outperforms the state-of-the-art approaches of GMM-SFA and the SVM-based one, GSL-NAP. Relative reductions of the Equal Error Rate of about 9.33% and 14.88% are respectively achieved over these systems.

Keywords Large margin training · Gaussian mixture models · discriminative learning · speaker recognition · session variability modeling

R. Jourani · R. André-Obrecht
SAMoVA Group, IRIT - UMR 5505 du CNRS
University Paul Sabatier, 118 Route de Narbonne, Toulouse, France
E-mail: {jourani, obrecht}@irit.fr

K. Daoudi
GeoStat Group, INRIA Bordeaux-Sud Ouest
351, cours de la libération, Talence. France
E-mail: khalid.daoudi@inria.fr

R. Jourani · D. Aboutajdine
Laboratoire LRIT. Faculty of Sciences, Mohammed 5 Agdal University
4 Av. Ibn Battouta B.P. 1014 RP, Rabat, Morocco
E-mail: aboutaj@fsr.ac.ma

1 Introduction

Most of state-of-the-art speaker recognition systems rely on the generative training of Gaussian Mixture Models (GMM) using maximum likelihood estimation and maximum a posteriori estimation (MAP) [1]. A speaker independent world model (UBM) is first trained with the Expectation-Maximization algorithm from hundreds of hours of speech data. The parameters of that model are then MAP adapted to the feature distribution of a target speaker. In speaker recognition applications, mismatch between the training and testing conditions can decrease considerably the performances. The session variability remains the most challenging problem to solve. The Factor Analysis techniques [2,3], e.g., Symmetrical Factor Analysis (SFA) [4,5], were proposed to address that problem in GMM based systems, by compensating for speaker and channel variability in the GMM supervector space.

The generative training of the GMM does not however directly optimize the classification performance. It was therefore of interest to develop alternative discriminative approaches that address directly the classification problem [6,7], as they should lead to better performances than generative methods. For instance, Support Vector Machines (SVM) combined with GMM supervectors are among state-of-the-art approaches in speaker verification [8,9]. The Nuisance Attribute Projection (NAP) [10] compensation technique is designed for the SVM based systems. NAP is a pre-processing method that aims to remove the directions of undesired sessions variability, before the SVM training.

Recently a new discriminative approach for multiway classification has been proposed, the Large Margin Gaussian mixture models (LM-GMM) [11]. As in SVM, the parameters of LM-GMM are trained by solving a convex optimization problem. However they differ from SVM by using ellipsoids to model the classes directly in the input space, instead of half-spaces in an extended high-dimensional space (no kernel trick/matrix is required). While LM-GMM have been used in speech recognition, they have not been used in speaker recognition (to the best of our knowledge). In an earlier work [12], we proposed a simplified version of LM-GMM which exploit the fact that traditional GMM systems use diagonal covariances and only the mean vectors are MAP adapted. We then applied this simplified version to a "small" speaker identification task. While the resulting training algorithm is more efficient than the original one, we found however that it is still not efficient enough to process large databases such as in NIST Speaker Recognition Evaluation (NIST-SRE) campaigns [13].

In order to address this problem, we propose in this paper a new approach for fast training of Large-Margin GMM which allows efficient processing in large scale applications. To do so, we exploit the fact that in general not all the components of the GMM are involved in the decision process, but only the k-best scoring components. We also exploit the property of correspondence between the MAP adapted GMM mixtures and the UBM mixtures. In order to show the effectiveness of the new algorithm, we carry out a full NIST speaker verification task using NIST-SRE'2006 (core condition) data. We evaluate our fast algorithm in a Symmetrical Factor Analysis compensation scheme, and

we compare it with the NAP compensated GMM supervector Linear Kernel (GSL-NAP) [9] and GMM-SFA [4] systems. The results show that our Large Margin compensated GMM outperform the state-of-the-art approaches GMM-SFA and GSL-NAP.

The paper is organized as follows. After an overview on Large-Margin GMM training with diagonal covariances in section 2, we describe our new fast training algorithm in section 3. The compensation technique SFA and the GSL-NAP system are then described in sections 4 and 5, respectively. Experimental results are reported in section 6.

2 Overview on Large Margin GMM with diagonal covariances (LM-dGMM)

In this section we start by recalling the original Large Margin GMM training algorithm developed in [11,14]. We then recall the simplified version of this algorithm that we introduced in [12].

In Large Margin GMM [11,14], each class c is modeled by a mixture of ellipsoids in the D- dimensional input space. The mth ellipsoid of the class c is parametrized by a centroid vector μ_{cm} (mean vector), a positive semidefinite (orientation) matrix Ψ_{cm} and a nonnegative scalar offset $\theta_{cm} \geq 0$. These parameters are then collected into a single enlarged matrix Φ_{cm}:

$$\Phi_{cm} = \begin{pmatrix} \Psi_{cm} & -\Psi_{cm}\mu_{cm} \\ -\mu_{cm}^T\Psi_{cm} & \mu_{cm}^T\Psi_{cm}\mu_{cm} + \theta_{cm} \end{pmatrix}. \tag{1}$$

A GMM is first fit to each class using maximum likelihood estimation. Let $\{o_{nt}\}_{t=1}^{T_n}$ ($o_{nt} \in \mathcal{R}^D$) be the T_n feature vectors of the nth segment (i.e. n^{th} speaker training data). Then, for each o_{nt} belonging to the class y_n, $y_n \in \{1, 2, ..., C\}$ where C is the total number of classes, we determine the index m_{nt} of the Gaussian component of the GMM modeling the class y_n which has the highest posterior probability. This index is called *proxy label*.

The training algorithm aims to find matrices Φ_{cm} such that "all" examples are correctly classified by at least one margin unit, leading to the LM-GMM criterion:

$$\forall c \neq y_n,\ \forall m, \qquad z_{nt}^T\Phi_{cm}z_{nt} \ \geq 1 + z_{nt}^T\Phi_{y_n m_{nt}}z_{nt}, \tag{2}$$

where $z_{nt} = \begin{bmatrix} o_{nt} \\ 1 \end{bmatrix}$. Eq. (2) states that for each competing class $c \neq y_n$ the match (in term of Mahalanobis distance) of any centroid in class c is worse than the target centroid by a margin of at least one unit.

In speaker recognition, most of state-of-the art systems use diagonal covariances GMM. In these GMM based speaker recognition systems, a speaker-independent *world model* or *Universal Background Model* (UBM) is first trained with the EM algorithm [15] from tens or hundreds of hours of speech data gathered from a large number of speakers. The background model represents

speaker-independent distribution of the feature vectors. When enrolling a new speaker to the system, the parameters of the UBM are adapted to the feature distribution of the new speaker. It is possible to adapt all the parameters, or only some of them from the background model. Traditionally, in the GMM-UBM approach, the target speaker GMM is derived from the UBM model by updating only the mean parameters using a *maximum a posteriori* (MAP) algorithm [1], while the (diagonal) covariances and the weights remain unchanged.

Following the same philosophy of traditional GMM, we proposed in [12] to neglect the orientation of the covariance matrices in training. We showed in [12] that the resulting simplified algorithm has the advantage of being more efficient than the original one while it still yielded similar or better performances on a speaker identification task. In our Large Margin diagonal GMM (LM-dGMM), each class (speaker) c is initially modeled by a GMM with M diagonal mixtures trained by MAP adaptation of a world model. For each class c, the mth Gaussian is parametrized by a mean vector μ_{cm}, a diagonal covariance matrix $\Sigma_m = diag(\sigma^2_{m1}, ..., \sigma^2_{mD})$, and the scalar factor θ_m which corresponds to the weight of the Gaussian.

With this relaxation on the covariance matrices, for each example o_{nt}, the goal of the training algorithm is now to force the log-likelihood of its proxy label Gaussian m_{nt} to be at least one unit greater than the log-likelihood of each Gaussian component of all competing classes. That is, given the training examples $\{(o_{nt}, y_n, m_{nt})\}_{n=1}^N$, we seek mean vectors μ_{cm} which satisfy the LM-dGMM criterion:

$$\forall c \neq y_n,\ \forall m, \qquad d(o_{nt}, \mu_{cm}) + \theta_m \geq 1 + d(o_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}}, \quad (3)$$

where $d(o_{nt}, \mu_{cm}) = \sum_{i=1}^{D} \frac{(o_{nti} - \mu_{cmi})^2}{2\sigma^2_{mi}}$.

Afterward, these M constraints are fold into a single one using the soft-max inequality $min_m\ a_m \geq -log \sum_m e^{-a_m}$. The segment-based LM-dGMM criterion becomes thus:

$$\begin{aligned} &\forall c \neq y_n, \\ &\tfrac{1}{T_n} \sum_{t=1}^{T_n} \left(-\, log \sum_{m=1}^{M} exp(-d(o_{nt}, \mu_{cm}) - \theta_m) \right) \\ &\geq\ 1 + \tfrac{1}{T_n} \sum_{t=1}^{T_n} d(o_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}}. \end{aligned} \quad (4)$$

The loss function to minimize for LM-dGMM is then given by:

$$\begin{aligned} Ł = \sum_{n=1}^{N} \sum_{c \neq y_n} max\Bigg(0\ ,\ 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} \Bigg(d(o_{nt}, \mu_{y_n m_{nt}}) \\ + \theta_{m_{nt}} + log \sum_{m=1}^{M} exp(-d(o_{nt}, \mu_{cm}) - \theta_m) \Bigg) \Bigg). \end{aligned} \tag{5}$$

3 LM-*d*GMM training with *k*-best Gaussians

3.1 Description of the new LM-dGMM training algorithm

Despite the fact that our LM-dGMM is computationally much faster than the original LM-GMM of [11,14], we still encountered efficiency problems when dealing with high number of Gaussian mixtures. Indeed, even for an easy 50 speakers identification task as the one presented in [12], we could not run the training in a relatively short time with our current implementation. This would imply that large scale applications such as NIST-SRE, where hundreds or thousands of target speakers are available, would be infeasible in reasonable time.

In order to develop a fast training algorithm which could be used in large scale applications, we propose to drastically reduce the number of constraints to satisfy in Eq. (4). By doing so, we would drastically reduce the computational complexity of the loss function and its gradient, which are the quantities responsible for most of the computational time. To achieve this goal we propose to use another property of state-of-the-art GMM systems, that is, decision is not made upon all mixture components but only using the k-best scoring Gaussians.

In other words, for each o_n and each class c, instead of summing over the M mixtures in the left side of equation Eq. (4), we would sum only over the k Gaussians with the highest posterior probabilities selected using the GMM of class c. In order to further improve efficiency and reduce memory requirement, we exploit the property reported in [1] about correspondence between MAP adapted GMM mixtures and UBM mixtures. We use the UBM to select one unique set S_{nt} of k-best Gaussian components per frame o_{nt}, instead of $(C-1)$ sets. This leads to a $(C-1)$ times faster and less memory consuming selection. Thus, the higher the number of target speakers is, the greater computation and memory saving is. More precisely, we now seek mean vectors μ_{cm} that satisfy the large margin constraints in Eq. (6):

$$\begin{aligned} &\forall c \neq y_n, \\ &\tfrac{1}{T_n} \sum_{t=1}^{T_n} \Bigg(- log \sum_{m \in S_{nt}} exp(-d(o_{nt}, \mu_{cm}) - \theta_m) \Bigg) \\ &\geq 1 + \tfrac{1}{T_n} \sum_{t=1}^{T_n} d(o_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}}. \end{aligned} \tag{6}$$

The loss function becomes:

$$Ł = \sum_{n=1}^{N} \sum_{c \neq y_n} max\Bigg(0\ ,\ 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} \Bigg(d(o_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}} + log \sum_{m \in S_{nt}} exp(-d(o_{nt}, \mu_{cm}) - \theta_m) \Bigg)\Bigg). \tag{7}$$

This loss function remains convex and can still be solved using dynamic programming.

During test, we compute a match score depending on both the target model $\{\mu_{cm}, \Sigma_m, \theta_m\}$ and the UBM $\{\mu_{Um}, \Sigma_m, \theta_m\}$ for each test hypothesis. We use again the same principle to achieve fast scoring. Given a test segment of T frames, for each test frame o_t we use the UBM to select the set E_t of k-best scoring proxy labels and compute the average log likelihood ratio using only these k labels:

$$LLR_{avg} = \tfrac{1}{T} \sum_{t=1}^{T} \Bigg(log \sum_{m \in E_t} \exp(-d(o_t, \mu_{cm}) - \theta_m) - log \sum_{m \in E_t} \exp(-d(o_t, \mu_{Um}) - \theta_m) \Bigg). \tag{8}$$

This quantity provides a score for the test segment to be uttered by the target model/speaker c. The higher the score is, the greater the probability that the test segment was uttered by the target speaker is.

3.2 Handling of outliers

We adopt the strategy of [11] to detect outliers and reduce their negative effect on learning. Outliers are detected using the initial GMM models. We compute the accumulated hinge loss incurred by violations of the large margin constraints in Eq. (6):

$$h_n = \sum_{c \neq y_n} max\Bigg(0\ ,\ 1 + \frac{1}{T_n} \sum_{t=1}^{T_n} \Bigg(d(o_{nt}, \mu_{y_n m_{nt}}) + \theta_{m_{nt}} + log \sum_{m \in S_{nt}} exp(-d(o_{nt}, \mu_{cm}) - \theta_m) \Bigg)\Bigg). \tag{9}$$

h_n measures the decrease in the loss function when an initially misclassified segment is corrected during the course of learning. We associate outliers with large values of h_n. We then re-weight the hinge loss terms in Eq. (7) by using segment weights $s_n = min(1, \frac{1}{h_n})$:

$$Ł = \sum_{n=1}^{N} s_n h_n. \tag{10}$$

We solve this unconstrained non-linear optimization problem using the second order optimizer LBFGS [16].

In summary, our new and fast training algorithm of LM-dGMM is the following:

```
- For each class (speaker), initialize with the GMM trained
by MAP adaptation of the UBM,
- select Proxy labels using these GMM,
- select the set of k-best UBM Gaussian components for each
training frame,
- compute the segment weights s_n,
- using the LBFGS algorithm, solve the unconstrained
non-linear minimization problem:
```

$$min \quad Ł. \tag{11}$$

4 Symmetrical Factor Analysis (SFA)

In this section we describe the symmetrical variant of the Factor Analysis model (SFA) [4,5] (Factor Analysis was originally proposed in [2,3]).

Given an M-components GMM adapted by MAP from the UBM, one forms a GMM supervector by stacking the D-dimensional mean vectors, leading to an MD supervector. This GMM supervector can be seen as a mapping of variable-length utterances into a fixed-length high-dimensional vector, through GMM modeling:

$$\phi(\mathrm{x}) = \begin{bmatrix} \mu_{x1} \\ \vdots \\ \mu_{xM} \end{bmatrix}, \tag{12}$$

where the GMM $\{\mu_{xm}, \Sigma_m, w_m\}$ is trained on the utterance x.

In the mean supervector space, a speaker model can be decomposed into three different components:

- a session-speaker independent component (the UBM model),
- a speaker dependent component,
- a session dependent component.

In the following, (h, s) will indicate the *session* h of the *speaker* s. The session-speaker model, can be written as [4]:

$$\mathbf{M}_{(h,s)} \;=\; \mathbf{M} \;+\; \mathbf{D}\mathbf{y}_s \;+\; \mathbf{U}\mathbf{x}_{(h,s)}, \tag{13}$$

where

- $\mathbf{M}_{(h,s)}$ is the session-speaker dependent supervector mean (an MD vector),
- $\mathbf{M}$ is the UBM supervector mean (an MD vector),

- $\mathbf{D}$ is a $MD \times MD$ diagonal matrix, where $\mathbf{DD}^T$ represents the a priori covariance matrix of $\mathbf{y}_s$,
- $\mathbf{y}_s$ is the speaker vector (speaker offset), an MD vector assumed to follow a standard normal distribution $\mathcal{N}(0, \mathrm{I})$,
- $\mathbf{U}$ is the session variability matrix of low rank R (an $MD \times R$ matrix),
- $\mathbf{x}_{(h,s)}$ are the channel factors (session offset), an R vector (theoretically, not dependent on s) assumed to follow a standard normal distribution $\mathcal{N}(0, \mathrm{I})$.

$\mathbf{Dy}_s$ and $\mathbf{Ux}_{(h,s)}$ represent respectively the speaker dependent component and the session dependent component [5].

The factor analysis modeling starts by estimating the $\mathbf{U}$ matrix, using several recordings (sessions) of various speakers. Given the fixed parameters $(\mathbf{M}, \mathbf{D}, \mathbf{U})$, the target models are then compensated by eliminating the session mismatch directly in the model domain. Whereas, the compensation in the test is performed at the frame level (feature domain).

5 The GSL-NAP system

In this section we briefly describe the GMM supervector linear kernel SVM system (GSL)[8] and its associated channel compensation technique, the Nuisance attribute projection (NAP) [10].

5.1 SVM-GMM supervector

The GMM supervector concept (a mapping from the feature space to a high-dimensional space) fits well with the idea of an SVM sequence kernel. The development of distance metrics in the GMM space led to the use of the Kullback-Leibler divergence to propose a linear kernel in the GMM supervector space. For two utterances x and y, the Kullback-Leibler divergence kernel is defined as:

$$\mathrm{K}(\mathrm{x}, \mathrm{y}) = \sum_{m=1}^{M} \left(\sqrt{w_m}\Sigma_m^{-(1/2)}\mu_{xm}\right)^T \left(\sqrt{w_m}\Sigma_m^{-(1/2)}\mu_{ym}\right). \tag{14}$$

The UBM weight and variance parameters are used to normalize the Gaussian means before feeding them into a linear kernel SVM training. This system is referred to as GSL in the rest of the paper.

5.2 Nuisance attribute projection (NAP)

NAP is a pre-processing method that aims to compensate the supervectors by removing the directions of undesired sessions variability, before the SVM training [10]. NAP transforms a supervector ϕ to a compensated supervector $\hat{\phi}$:

$$\hat{\phi} = \phi \; - \; \mathbf{S}(\mathbf{S}^T\phi), \tag{15}$$

using the eigenchannel matrix $\mathbf{S}$, which is trained using different recordings per speakers.

Given a set of expanded recordings:

$$\{\phi(1, s_1) \cdots \phi(h_1, s_1) \cdots \phi(1, s_N) \cdots \phi(h_N, s_N)\}, \tag{16}$$

of N different speakers, with h_i different sessions for each speaker s_i, one first removes the speakers variability by subtracting the mean of the supervectors within each speaker $\{\overline{\phi(s_i)}\}$:

$$\forall s_i, \forall h, \qquad \phi(h, s_i) = \phi(h, s_i) \; - \; \overline{\phi(s_i)}. \tag{17}$$

The resulting supervectors are then pooled into a single matrix:

$$\mathbf{C} = \left[\phi(1, s_1) \cdots \phi(h_1, s_1) \cdots \phi(1, s_N) \cdots \phi(h_N, s_N)\right], \tag{18}$$

representing the intersession variations. One identifies finally the subspace of dimension R where the variations are the largest by solving the eigenvalue problem on the covariance matrix $\mathbf{C}\mathbf{C}^T$, getting thus the projection matrix $\mathbf{S}$ of a size $MD \times R$. Theoretically, the matrix S is similar to the channel matrix U of SFA. This system is referred to as GSL-NAP in the rest of the paper.

6 Experimental results

We perform experiments on the NIST-SRE'2006 [17] speaker verification task and compare the performances of the baseline GMM, the LM-dGMM and the SVM systems, with and without using channel compensation techniques. The comparisons are made on the male part of the NIST-SRE'2006 core condition (1conv4w-1conv4w). Performances are assessed using Detection Error Tradeoff (DET) plots and measured in terms of equal error rate (EER) and minimum of detection cost function (minDCF). The latter is calculated following NIST criteria [18].

The feature extraction is carried out by the filter-bank based cepstral analysis tool Spro [19]. Bandwidth is limited to the 300-3400Hz range. 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate and transformed into Linear Frequency Cepstral Coefficients (LFCC) [20]. Consequently, the feature vector is composed of 50 coefficients including 19 LFCC, their first derivatives, their 11 first second derivatives and the delta-energy. The LFCCs are preprocessed by Cepstral Mean Subtraction and variance normalization [21]. We applied an energy-based voice activity detection to remove silence frames, hence keeping only the most informative frames. Finally, the remaining parameter vectors are normalized to fit a zero mean and unit variance distribution.

We use the state-of-the-art open source software ALIZE/Spkdet [5,22] for GMM, SFA, GSL and GSL-NAP modeling. A male-dependent UBM is trained using all the telephone data from the NIST-SRE'2004. Then we train a MAP adapted GMM for the 349 target speakers belonging to the primary task. The

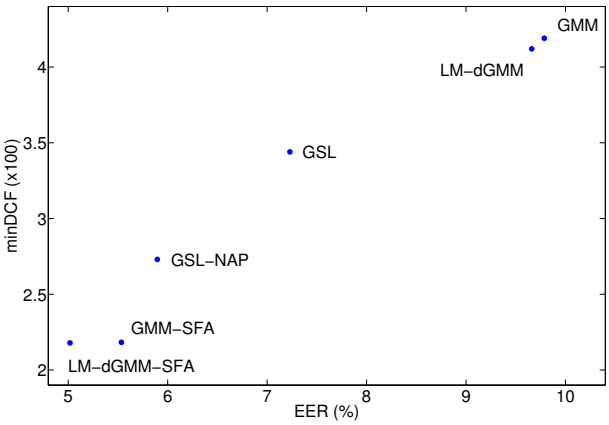

Fig. 1 EER and minDCF performances for GMM, Large Margin diagonal GMM and GSL systems with and without channel compensation.

Table 1 EERs(%) and minDCFs(x100) of GMM, Large Margin diagonal GMM and GSL systems with and without channel compensation.

System	EER	minDCF
GMM	9.79	4.19
LM-dGMM	9.66	4.12
GSL	7.23	3.44
GSL-NAP	5.90	2.73
GMM-SFA	5.53	2.18
LM-dGMM-SFA	5.02	2.18

corresponding list of 22123 trials (involving 1601 test segments) are used for test. Score normalization techniques are not used in our experiments. The so MAP adapted GMM define the baseline GMM system, and are used as initialization for the LM-dGMM one. The GSL system uses a list of 200 impostor speakers from the NIST-SRE'2004, on the SVM training. The LM-dGMM-SFA system is initialized by model domain compensated GMM, which are then discriminated using feature domain compensated data. The session variability matrix $\mathbf{U}$ of SFA and the channel matrix $\mathbf{S}$ of NAP, both of rank $R = 40$, are estimated on NIST-SRE'2004 data using 2934 utterances of 124 different male speakers.

Figure 1 and table 1 provide the EER and minDCF performances of all systems, with and without channel compensation, for models with 512 Gaussian components ($M = 512$). All the LM-dGMM systems scores are obtained with the 10 best Gaussian components selected using the UBM, $k = 10$.

The results show that, without SFA channel compensation, the LM-dGMM system outperforms the classical generative GMM one, however it does yield worse performances than the discriminative approach GSL. Nonetheless, when applying channel compensation techniques, compensated models outperform

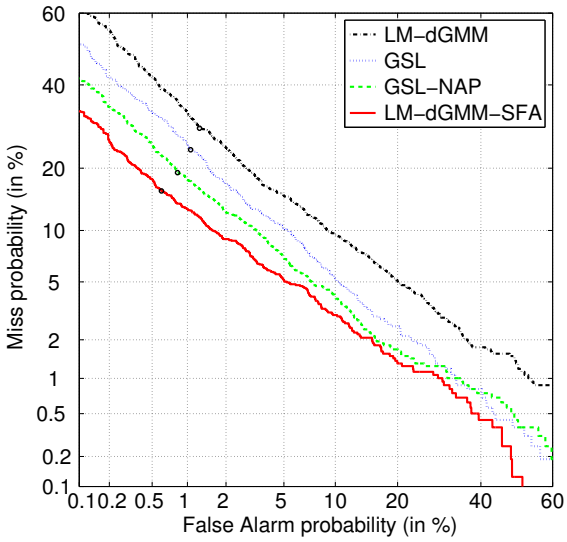

Fig. 2 DET plots for Large Margin diagonal GMM and GSL systems with and without channel compensation.

Table 2 EER(%) performances for LM-dGMM systems with different margins.

Margin	1	1.125	1.5	2	3	5	9
EER	5.02	4.85	5.53	5.53	5.52	5.53	5.53

the non-compensated ones as expected, but the LM-dGMM-SFA system significantly outperforms the GSL-NAP and GMM-SFA ones. Our (compensated) system achieves 5.02% equal error rate, while the GSL-NAP and GMM-SFA achieve respectively 5.90% and 5.53%. This leads to relative reductions of EER of about 14.88% and 9.33% over these systems. Figure 2 shows DET plots for LM-dGMM and GSL systems with and without channel compensation.

Table 2 shows the EER scores of LM-dGMM-SFA models with different minimal margin values to satisfy. Like in SVM, one can see that an improved margin selection can improve the LM-dGMM (+SFA) performance (an ERR of 4.82 here instead of 5.02 with a unit margin). All these results show that our fast Large Margin GMM discriminative learning algorithm not only allows efficient training but also achieves better performances than the state-of-the-art techniques GSL-NAP and GMM-SFA.

7 Conclusion

We presented a new fast algorithm for discriminative training of Large-Margin diagonal GMM by using the k-best scoring Gaussians selected form the UBM. This algorithm is highly efficient which makes it well suited to process large

scale databases such as in NIST-SRE. We carried out experiments on a full speaker verification task under the NIST-SRE'2006 core condition. Combined with the SFA channel compensation technique, the resulting algorithm significantly outperforms the state-of-the-art speaker recognition approaches GSL-NAP and GMM-SFA. Our future work will consist in improving margin selection and outliers handling. This should indeed significantly improve the performances.

References

1. Reynolds D.A., Quatieri T.F., Dunn R.B. (2000) Speaker verification using adapted Gaussian mixture models. Digital signal processing, vol. 10, no. 1-3, pp. 19-41.
2. Kenny P., Boulianne G., Dumouchel P. (2005) Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing, vol. 13, no. 3, pp. 345-354.
3. Kenny P., Boulianne G., Ouellet P., Dumouchel P. (2007) Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1448-1460.
4. Matrouf D., Scheffer N., Fauve B., Bonastre J.-F. (2007) A straightforward and efficient implementation of the factor analysis model for speaker verification. In: Interspeech, pp. 1242-1245.
5. Fauve B.G.B., Matrouf D., Scheffer N., Bonastre J.-F., Mason J.S.D. (2007) State-of-the-Art Performance in Text-Independent Speaker Verification through Open-Source Software. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, Issue 7, pp. 1960-1968.
6. Keshet J., Bengio S. (2009) Automatic speech and speaker recognition: Large margin and kernel methods. Wiley.
7. Louradour J., Daoudi K., Bach F. (2007) Feature space mahalanobis sequence kernels: Application to svm speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2465-2475.
8. Campbell W.M., Sturim D.E., Reynolds D.A. (2006) Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, vol. 13, no. 5, pp. 308-311.
9. Campbell W.M., Sturim D.E., Reynolds D.A., Solomonoff A. (2006) SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: ICASSP, vol. 1, IEEE, pp. I-97-I-100.
10. Solomonoff A., Campbell W.M., Quillen C. (2007) Nuisance Attribute Projection. Speech Communication.
11. Sha F., Saul L.K. (2006) Large margin Gaussian mixture modeling for phonetic classification and recognition. In: ICASSP, vol. 1, IEEE, pp. 265-268.
12. Jourani R., Daoudi K., André-Obrecht R., Aboutajdine D. (2010) Large Margin Gaussian mixture models for speaker identification. In: Interspeech, pp. 1441-1444.
13. http://www.itl.nist.gov/iad/mig//tests/sre/
14. Sha F. (2007) Large margin training of acoustic models for speech recognition. Ph.D. dissertation, University of Pennsylvania.
15. Bishop C.M. (2006) Pattern recognition and machine learning. Springer Science+Business Media, LLC, New York.
16. Nocedal J., Wright S.J. (1999) Numerical optimization. Springer verlag.
17. The NIST Year 2006 Speaker Recognition Evaluation Plan. (2006) `http://www.itl.nist.gov/iad/mig/tests/spk/2006/sre-06\_evalplan-v9.pdf`.
18. Przybocki M., Martin A. (2004) NIST Speaker Recognition Evaluation Chronicles. In: Odyssey-The Speaker, Language Recognition Workshop, pp. 15-22.
19. Gravier G. (2003) SPro: "Speech Signal Processing Toolkit". `https://gforge.inria.fr/projects/spro`.
20. Davis S. B., Mermelstein P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, Issue 4, pp. 357-366.

21. Viikki O., Laurila K. (1998) Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, vol. 25, no. 1-3, pp. 133-147.
22. Bonastre J.-F., Scheffer N., Matrouf D., Fredouille C., Larcher A., Preti A., Pouchoulin G., Evans N., Fauve B., Mason J. (2008) ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In: Odyssey-The Speaker and Language Recognition Workshop.