# Anchor and UBM-based Multi-Class MLLR M-Vector System for Speaker Verification

Achintya K Sarkar, Claude Barras

# Anchor and UBM-based Multi-Class MLLR *M-Vector* System for Speaker Verification

*A. K. Sarkar and C. Barras*

LIMSI-CNRS, Université Paris-Sud, BP 133, 91403 Orsay, France
{sarkar,barras}@limsi.fr

## Abstract

In this paper, we propose two techniques to extend the recently introduced global Maximum Likelihood Linear Regression (MLLR) transformation (i.e. super-vector) based m-vector system for speaker verification into a multi-class MLLR m-vector system in the Universal Background Model (UBM) framework. In the first method, Gaussian mean vectors of the UBM are first grouped into several classes using conventional K-means and a proposed clustering algorithm based on Expectation Maximization (EM) and Maximum Likelihood (ML) concepts. Then, MLLR transformations are calculated for a given speech data with respect to each class, which are used in the form of super-vector for speaker representation by their m-vectors. In the second approach, several MLLR transformations are estimated with respect to pre-defined models called anchors. The proposed systems show better performance than the conventional system. Furthermore, the proposed UBM-based system does not require additional alignment of speech data with respect to the UBM for estimation of multiple MLLR transformations. We also further show that the proposed EM & ML clustering algorithm is robust to random initialization and provides equal or comparable system performance compared to K-means. The experimental results are shown on NIST 2008 SRE core condition over various tasks.

**Index Terms**: m-Vector, Multi-Class MLLR, Anchor Model, EM Clustering, Speaker Verification

## 1. Introduction

The recently introduced m-vector technique [1] based on Universal Background Modeling (UBM) uses a *global* Maximum Likelihood Linear Regression (MLLR) transformation in the form of a super-vector for speaker characterization by their *m-vectors*. As per [1], the global MLLR transformation is estimated with respect to UBM for a given speaker data/speech segment without any phonetic/speech transcription knowledge, and is then *uniformly* segmented using a sliding overlapped window; each segment is called an *m-vector*. During test, m-vectors of the test utterance and claimant are post-processed for session variability compensation before scoring. It is shown in [1, 2] that the m-vector system is able to retrieve more speaker relevant information from the MLLR super-vector than the conventional way of speaker representation by their *full* MLLR super-vectors and yields promising Speaker Verification (SV) performance compared to the classical i-vector based SV system. Later, the effectiveness of the m-vector technique is also revealed in a Automatic Speech Recognition (ASR) based system [2] with phonetic class wise MLLR transformation and shows performance better than UBM based.

However, one of the major drawback of the ASR based systems is that it is computationally very expensive for estimation of the MLLR transformation. Since, it uses the Hidden Markov Modeling (HMM) concept to capture the temporal information of phones and requires huge modeling parameters. On the other-hand, UBM based system uses 512-2048 Gaussian components for modeling. Therefore, UBM based systems are more suitable in real time applications for SV than ASR.

Motivation of this paper is to extend the conventional global/single class MLLR transformation i.e. super-vector based m-vector system in UBM framework into multi-class wise MLLR transformation based m-vector system to incorporate the advantage of class specific MLLR transformations. Our proposed techniques are broadly divided into two categories: first case, the Gaussian components of the UBM are clustered into different groups and then an MLLR transformation is estimated with respect to each class using the sufficient statistics accumulated from the Gaussian components of the particular class. Two clustering algorithms are considered: one is conventional K-means and the other is a proposed algorithm based on the concept of Expectation Maximization (EM) and Maximum Likelihood (ML). It develops two proposed systems called, *K-means* and EM multi-class MLLR m-vector systems, respectively. The salient feature of these proposed systems is that it does not require additional alignment of data even for estimation of multiple MLLR transformations with respect to UBM compared to the conventional UBM based m-vector system. Further, the proposed clustering technique is robust to random initialization, unlike K-means, and provides equal or comparable system performance which is best obtained with K-means over several pass run of experiments.

In the second case, MLLR transformations are estimated with respect to pre-defined models called *anchors* for the m-vector system. Anchor models are built by clustering either *non-target* or *target training* speaker data. It yields two proposed multi-class MLLR m-vector systems: one is *non-target* and the other is *target anchor based*, respectively. Several recent studies of speaker identification using anchor modeling can be found in [3, 4]. We show that the proposed system provides better speaker verification performance than the conventional m-vector system. Experimental results are presented on various tasks of the NIST 2008 SRE core condition.

The paper is organized as follows: Section 2 describes MLLR super-vector. Section 3 describes m-vector technique. Section 4 describes proposed systems. Section 5 describes post-processing and scoring. Baseline system and experimental setup are described in Section 6. Results and discussions are presented in Section 7 before the conclusion in Section 8.

This work was partly realized as part of the Quaero Program funded by OSEO (French State agency for innovation).

## 2. MLLR Super-Vector

Estimation of a MLLR [5] transformation $W$ for a given speaker/speech data $X = \{x_1, x_2, \ldots, x_T\}$ with respect to UBM involve the following steps:

***Initial*:** Load UBM, feature vectors, $X$ and calculate the probabilistic alignment, $\gamma_j(t)$ for the $j^{th}$ Gaussian of UBM as:

$$\gamma_j(t) = p(j|x_t) = \frac{\omega_j b_j(x_t)}{\sum_{k=1}^{c} \omega_k b_k(x_t)} \quad (1)$$

where $c$ and $b_k$ indicate the number of Gaussians and the density function of the $k^{th}$ Gaussian of the UBM, respectively.

**Step 1:** Calculate the following two sufficient statistics for the $i^{th}$ *components (dimension) of the feature vectors*,

$$K^{(i)} = \sum_{j=1}^{c} \sum_{t=1}^{T} \gamma_j(t) \frac{1}{\sigma_{ji}^2} x_i(t) \mu_j^{'} \quad (2)$$

$$G^{(i)} = \sum_{j=1}^{c} \frac{1}{\sigma_{ji}^2} \mu_j \mu_j^{'} \sum_{t=1}^{T} \gamma_j(t) \quad (3)$$

$\mu_j$ and $\sigma_{ji}^2$ are $j^{th}$ mean, and $i^{th}$ component of $j^{th}$ covariance matrix of UBM, respectively. The symbol $(.)'$ indicates matrix transpose.

**Step 2:** $i^{th}$ *row of the MLLR transformation* $W$ is obtained as,

$$W_i = K^{(i)} {G^{(i)}}^{-1} \quad (4)$$

**Step 3:** Repeat Step 1 to 2 upto feature vector dimension

Afterward, the rows of the MLLR transform are stacked [6] to form a super-vector. We use 47 dimensional feature vectors which gives $47 * 47 = 2209$ dimensional MLLR super-vectors.

## 3. m-Vector Technique

In this technique [1, 2], speakers are characterized by *m-vectors*, which are obtained by *uniform* segmentation of their MLLR super-vectors using an overlapped window as illustrated in Fig. 1. Following eqn.(4), each row of the MLLR transformation is associated to a particular component of the feature vectors. It gives several m-vectors per speaker and constitutes several sub-systems. In the test phase, m-vectors of the test utterance are extracted in a similar manner and scored against the corresponding m-vectors of the claimant. Before scoring, m-vectors are post-processed for session variability compensation. It is observed in [1, 2] that *full* system (which represents speakers conventionally by their full super-vector) also contains complementary information for m-vector system and fusion of both systems further reduce the speaker verification error rate. Hence, all system performances are presented in this paper with fusion of their *full* system with m-vector system.

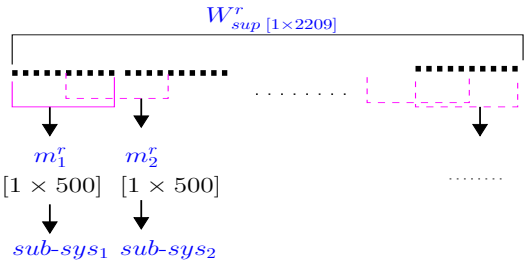


Figure 1: *m-vector extraction of* $r^{th}$ *speaker from his/her MLLR super-vector using an overlapped window of* 500 *elements.*

## 4. Proposed Systems

### 4.1. EM multi-class MLLR m-vector system

Here, UBM's Gaussian mean vectors are first clustered into different groups using the proposed clustering algorithm based on the concept of Expectation Maximization (EM) and Maximum Likelihood (ML) as described in *Algorithm 1*.

Algorithm 1: *Proposed* clustering algorithm using EM and ML

***Initial:*** Load UBM and chose number of clusters $L$

**Step 1:** Use Gaussian mean vectors of the UBM as feature vectors, $Y = \{\mu_1, \mu_2, \ldots \mu_c\}$

**Step 2:** Train a $L$ components Gaussian Mixture Model (GMM) $\sim \mathcal{N}(\tilde{\omega}_i, \tilde{\mu}_i, \tilde{\Sigma}_i)$, $i = 1 \ldots L$, using the feature vectors $Y$ with EM algorithm of *random initialization*

**Step 3:** Iterate EM algorithm in *Step 2* several times

**Step 4:** Separate *each Gaussian component* of the GMM obtained in *Step 2* as a *single Gaussian model* and discard the weights $\tilde{\omega}_i$ to give *equal importance* to all the models:

$$\lambda_i \sim \mathcal{N}(\tilde{\mu}_i, \tilde{\Sigma}_i) \quad (5)$$

**Step 5:** Assign the $c^{th}$ Gaussian mean vector of the UBM, i.e. $\mu_c$ to cluster $k$ in the ML sense as,

$$k = \arg \max_{1 \le j \le L} p(\mu_c|\lambda_j) \quad (6)$$

1000 iterations are used in *Step* 3 of *Algorithm 1* (with constraints on initial and final variance ceiling, flooring of global data). The parameters of the models $\lambda_1, \ldots \lambda_L$ are slightly different each run for a particular cluster, however it yields the same final clustering output, showing that this clustering algorithm is not affected by the random initialization.

After that, a MLLR transformation is estimated for a given speech data $X = \{x_1, x_2, \ldots, x_T\}$ with respect to each class using the *sufficient statistics* accumulated from the *Gaussian components* for the *respective class* described in *Algorithm 2*.

Algorithm 2: Estimation of cluster-wise MLLR transformation

**Step 1:** Estimate $\gamma_j(t)$ for the feature vector $X$ with respect to the UBM as in Eqn.(1)

**Step 2:** For the $L^{th}$ class, compute the sufficient statistics using *Gaussian components* $\epsilon$ $L$ as in Eqn.(2-3),

$$K_L^{(i)} = \sum_{j \epsilon L} \sum_{t=1}^{T} \gamma_j(t) \frac{1}{\sigma_{ji}^2} x_i(t) \mu_j^{'} \quad (7)$$

$$G_L^{(i)} = \sum_{j \epsilon L} \frac{1}{\sigma_{ji}^2} \mu_j \mu_j^{'} \sum_{t=1}^{T} \gamma_j(t) \quad (8)$$

**Step 3:** $i^{th}$ row of the MLLR transformation for $L^{th}$ class is obtained,

$$W_i^L = K_L^{(i)} {G^{(i)}}_L^{-1} \quad (9)$$

**Step 4:** Repeat *Step 2 to 3* upto the number of classes

*It can be observed from* Algorithm 2 *in Step* 1 *that alignment of data is required* only once *with respect to UBM, even with estimation of multiple class-wise MLLR transformations*. Finally, these MLLR transformations are used for speaker verification with m-vector technique described in Sec.3. Fig.2 illustrates the above procedure.

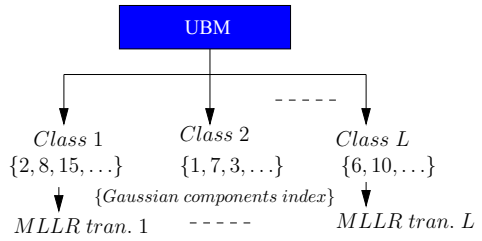


Figure 2: *Clustering of UBM Gaussian components and estimation of an MLLR transformation with respect to each cluster.*

### 4.2. K-means multi-class MLLR m-vector system

This system is similar to the *EM multi-class MLLR m-vector* system except that a conventional K-means algorithm with random initialization and an euclidean distance measure is used for clustering. Clustering stops when the clusters are stable.

### 4.3. Anchor based multi-class MLLR m-vector systems

*Non-target anchor multi-class MLLR m-vector system:* UBM training non-target speaker data are first clustered into different groups using their MLLR super-vectors and then cluster-wise model is derived from UBM with Maximum a posteriori (MAP) adaptation using data from the respective group. Then, cluster-specific models are used to iteratively recluster the data in the ML sense similarly to Multiple Background Model (MBM) formation in [7]. It generates new cluster-wise models after each iteration. We follow 20 such iterations and observe that clustered associated data are not altered. Finally, a Gaussian Mixture Model (GMM) with 512 components called *anchor model/anchor* is estimated with respect to each cluster using data belonging to the particular cluster from scratch. During training/testing, MLLR transformations of a given speech data are estimated with respect to *anchor models* for m-vectors.

*Target anchor case* is similar to the *non-target anchor* case with the only difference that target speaker training data is used for the clustering and anchor model formation.

## 5. Post-processing and Scoring

Different session variability compensation techniques can be found in literature, e.g., LDA followed by Within Class Covariance Normalization (WCCN) or Probabilistic (P)-LDA [8, 9]. In our setup, Linear Discriminant Analysis (LDA) projected m-vectors are conditioned using the Eigen Factor Radial (EFR) [10] algorithm recently introduced in i-vector environment to handle the session variability compensation as in Eqn.(10).

$$\hat{m} \leftarrow \frac{V^{-\frac{1}{2}}(m - \overline{m})}{\sqrt{(m - \overline{m})' V^{-1} (m - \overline{m})}} \qquad (10)$$

where $\bar{m}$ and $V$ are respectively, the mean and covariance matrix of m-vectors for non-target speakers in the development set and $\hat{m}$ represents the conditioned m-vector.

During test phase, the score between the two LDA-EFR processed m-vectors is calculated using a Mahalanobis based scoring function [10]. LDA and EFR are implemented separately for each sub-system. Finally, m-vector scores for the respective sub-systems are fused with equal weights across a particular LDA dimension. All results presented in the paper were computed with two iterations of EFR.

## 6. Experimental Setup

Following [1], the baseline system considers a global MLLR transformation derived from the UBM for a given speech utterance and processed by the m-vector technique to characterize the speaker as described in Sec.3. All experiments are carried on NIST 2008 SRE male speakers as per NIST evaluation plan [11]. There are 1270 utterances for training 1270 target models. Each utterance is around 5 minutes long with 2.5 minutes of speech in average.

47 dimensional PLP features (15 static with their $\Delta$, $\Delta\Delta$, $\Delta$E and $\Delta\Delta$E) are extracted from the speech signal each 10 ms with a Hamming window over the 0-3800 Hz bandwidth. An energy-based voice activity detection is applied on the feature vectors to discard less energetic or silent frames. Then, selected frames are normalized to zero mean and unit variance at the utterance level. A male gender dependent UBM of 512 mixture with diagonal covariance matrices, is trained using *non-target* speaker data from NIST 2004 SRE; unless mentioned, all reported experiment are shown for a UBM with 512 Gaussians. LDA and EFR are estimated using 12399 utterances from 890 non-target speakers over NIST 2004-05, Switchboard 1, 2, 3 and Switchboard cell 1 & 2 (about 15 sessions per speaker). All systems use a single iteration of adaptation for MLLR transformation. If the inverse of $G^{(i)}$ matrix for a particular class is singular due to a lack of data, the global MLLR transformation is used instead. Equal Error Rate (EER) and Minimum Detection Cost Function (MinDCF) are used for the evaluation of the system performances as per NIST 2008 plan [11].

## 7. Results and Discussion

For analysis, Speaker Verification (SV) performances in terms of EER are compared on NIST 2008 SRE core condition det 7 task. For simplicity, optimal LDA dimension is not shown in the tables. All m-vector system results are presented in the paper for m-vector size of 500.

### 7.1. Selection of optimal anchor multi-class MLLR m-vector system

Table 1 shows the effect of varying the number of anchor models on SV performance with the proposed anchor-based multi-class MLLR m-vector system. Table 2 compares the performance of the baseline system for different UBM sizes with the optimal anchor-based multi-class m-vector system obtained in Table 1. In the case of two anchors, one cluster contains 45% and the other one 55% of the total training data also used for UBM training. From Tables 1 and 2, it can be observed that the proposed anchor multi-class MLLR m-vector system performs

Table 1: *Performance of the anchor-based multi-class MLLR m-vector system depending on the number of anchor models on NIST 2008 SRE core condition (det 7 task).*

| Anchor based m-vector system | # of anchors [% EER] | | | |
|---|---|---|---|---|
| | 2 | 3 | 4 | 5 |
| Non-target | 3.31 | 3.15 | **2.93** | 3.10 |
| Target | 3.20 | 3.13 | 3.27 | 3.10 |

Table 2: *EER of best non-target anchor-based multi-class MLLR m-vector system and of the baseline system with various UBM sizes on NIST 2008 SRE core condition (det 7 task); each anchor model has* 512 *mixture components.*

| m-vector system | UBM size for baseline/(equ. # of anchors) | | | |
|---|---|---|---|---|
| | 512/(1) | 1024/(2) | 1536/(3) | 2048/(4) |
| Baseline | 3.45 | 3.70 | 3.60 | 3.78 |
| **Anch. non-target** | - | 3.31 | 3.15 | **2.93** |

Table 3: *EER for a number of class-wise MLLR transformations with UBM based multi-class MLLR m-vector system on NIST 2008 SRE core condition (det 7 task). Number of Gaussians in the respective classes are shown in parenthesis.*

| m-vector system | Clustering Algorithm | # of class-wise MLLR trans. | EER (%) |
|---|---|---|---|
| Baseline | - | 1 (global) | 3.45 |
| EM multi-class MLLR | Proposed EM & ML | **2** (358,154) | **3.21** |
| | | 3 (100,100,312) | 3.44 |
| K-means multi-class MLLR | Conventional K-means | 2 (252,260) | **3.22** |
| | | **3** (170,180,162) | **3.08** |

better than the baseline system with a UBM having an equivalent number of Gaussian components.

### 7.2. Selection of optimal UBM multi-class MLLR m-vector system

Table 3 compares the SV performance of the proposed UBM-based multi-class MLLR m-vector systems for various number of class wise MLLR transformations on NIST 2008 SRE core condition det 7 task. For the K-means case, the system performance is given for the experiment pass which showed the best SV performance over 10 runs as in Fig.3.

From Table 3, it can be observed that the proposed multi-class MLLR m-vector systems show better performance than the baseline system. Both proposed systems give lower EER as the number of classes (i.e. MLLR transformations) increases and obtain optimal results for 2 and 3 classes, respectively with proposed and K-means algorithm. However, the performance for 2 and 3 classes are very similar in K-means.

Having few classes leads to a higher acoustic variability within a class, but increasing the number of classes reduces the amount of data for each class and may split an acoustic context across several classes. Hence, further clustering is not performed.

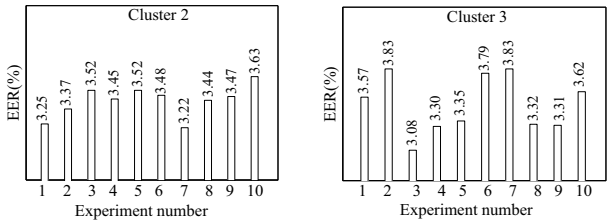


Figure 3: *Effect of random initialization with K-means clustering algorithm in terms of speaker verification EER on det 7 task of NIST 2008 SRE core condition.*

From Table 3 and Figure 3, it can be deduced that the proposed EM clustering method is robust to random initialization, compared to K-means, and gives equal or comparable SV performance than the *best system obtained with K-means. Most of the experiments results obtained with K-means for* 3 *classes are similar to the proposed clustering.* It also reflects that the proposed algorithm provides optimal clustering and does not require many experiments to judge system performance like with K-means.

### 7.3. Performance over different recording conditions

Table 4 compares the SV performance of the proposed optimal multi-class MLLR m-vector systems obtained in Tables 1 & 3 with the baseline system on NIST 2008 SRE over various tasks. From Table 4, it can be observed that the proposed UBM based multi-class MLLR m-vector system shows lower EER and MinDCF in most of the det tasks. The performance of the proposed algorithm having 2 classes also shows very comparable performance to the system which is even obtained with 3 classes in K-means. Moreover, it does not require additional temporal alignment of the data with respect to the UBM for estimation of multi-class wise MLLR transforms compared to the baseline system (see Algorithm 2).

Table 4: *Comparison of speaker verification performance of the baseline system with the proposed optimal multi-class MLLR m-vector systems (in Tables 1 & 3) on NIST 2008 SRE core condition over various tasks.*

| m-vector system | %EER/(MinDCF) det 5 | det 6 | det 7 | det 8 |
|---|---|---|---|---|
| Baseline | 7.11 (0.0351) | 6.46 (0.0392) | 3.45 0.0193 | 2.92 (0.0155) |
| EM multi-class | **5.51** (**0.0298**) | 6.62 (**0.0382**) | **3.21** (**0.0191**) | **2.20** (**0.0121**) |
| K-means multi-class | **5.55** (**0.0300**) | 6.50 (**0.0380**) | **3.08** (**0.0181**) | **2.16** (**0.0101**) |
| Anch. non-target multi-class | 8.00 (0.0361) | 6.57 (**0.0372**) | **2.93** (**0.0186**) | **1.75** (**0.0132**) |

Anchor based non-target multi-class MLLR m-vector system shows *considerably* better performance for det 7 & 8 (related to english data: tel-tel configuration) and slightly degradation in case of det 5 (tel-mic) & 6 (tel-tel with mix of languages). It can be due to the fact that training data of *anchor* associated clusters are not well balanced across language or microphone.

## 8. Conclusion

In this paper, we extended the conventional global MLLR transformation based m-vector system in UBM framework into multi-class wise MLLR m-vector system to account for the advantage of class specific MLLR transformations. We have proposed two techniques: in the first approach, Gaussian mean vectors of the UBM are grouped into several classes using conventional K-means, and a proposed clustering algorithm based on EM and ML concepts. Then, MLLR transformations are calculated with respect to each class for a given speech data using *sufficient statistics accumulated* from the *Gaussians* of the *particular* class, which are used in the form of super-vector for speaker representation by their m-vectors. Hence, it does not require additional alignment of speech data with respect to UBM for multiple MLLR transformations. In the second case, MLLR transformations are estimated with respect to predefined *anchor models*. The proposed systems show better performance than the conventional system. The experimental results are compared on various tasks in core condition of NIST 2008 SRE. We also show that the proposed EM & ML based clustering algorithm is robust to random initialization and provides equal or comparable system performance compared to K-means. Moreover, it does not require many experiments to judge the system performance like K-means based system. Lastly, anchor based system indicates that it is better to use multiple MLLR transformations derived with respect to a number of anchors rather than use a larger UBM based single-class/global MLLR transformation for speaker verification in m-vector framework.

## 9. References


[1] A. K. Sarkar, J. F. Bonastre, and D. Matrouf, "Speaker Verification using m-vector Extracted from MLLR Supervector," in *Proc. of 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 21–25.

[2] A. K. Sarkar, C. Barras, and V. B. Le, "Lattice MLLR based *m-vector* System for Speaker Verification," in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2013, Vancouver, Canada.

[3] A. K. Sarkar and S. Umesh, "Eigen-voice Based Anchor Modeling System for Speaker Identification using MLLR Super-vector," in *Proc. of Interspeech*, 2011, pp. 2357–2360.

[4] A. K. Sarkar, S. Umesh, and J. F. Bonastre, "Computationally Efficient Speaker Identification Using Fast-MLLR Based Anchor Modeling," in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2012, pp. 4357–4360.

[5] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[6] A. Stolcke et al., "MLLR Transforms as Features in Speaker Recognition," in *Proc. of Eur. Conf. Speech Commun. and Tech. (EUROSPEECH)*, 2005, pp. 2425–2428.

[7] A. K. Sarkar and S. Umesh, "Use of VTL-wise Models in Feature-Mapping Framework to Achieve Performance of Multiple-Background Models in Speaker Verification," in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2011.

[8] Simon J.D. Prince, "Computer Vision: Models Learning and Inference," in *Cambridge University Press, 2012, In press*.

[9] M. Senoussaoui et al., "Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition," in *Proc. of Interspeech*, 2011, pp. 25–28.

[10] P. M. Bousquet, D. Matrouf, and J. F. Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition," in *Proc. of Interspeech*, 2011, pp. 485–488.

[11] The NIST Year 2008 Speaker Recognition Evaluation Plan., "http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf," .