# Acoustic Feature Transformation using UBM-based LDA for Speaker Recognition

*Chengzhu Yu, Gang Liu, John H. L. Hansen*

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX, U.S.A.
{chengzhu.yu, gang.liu, john.hansen}@utdallas.edu

10.21437/Interspeech.2014-420

## Abstract

In state-of-the-art speaker recognition system, universal background model (UBM) plays a role of acoustic space division. Each Gaussian mixture of trained UBM represents one distinct acoustic region. The posterior probabilities of features belonging to each region are further used as core components of Baum-Welch statistics. Therefore, the quality of estimated Baum-Welch statistics depends highly on how acoustic regions are separable with each other. In this paper, we propose to transform the front end acoustical features into a space where the separability of mixtures of trained UBM can be optimized. To achieve this, an UBM was first trained from the acoustical features and a transformation matrix is estimated using linear discriminant analysis (LDA) by treating each mixture of trained UBM as independent class. Therefore, the proposed method named as UBM-based LDA (uLDA) does not require any speaker labels or other supervised information. The obtained transformation matrix is then applied to acoustic features for i-Vector extraction. Experimental results on the male part of core conditions of NIST SRE 2010 dataset confirmed the improved performance using proposed method.

**Index Terms**: Speaker recognition, i-Vector, universal background model (UBM), Baum-Welch statistic, LDA.

## 1. Introduction

State-of-the-art speaker recognition systems are based on i-Vector extraction and PLDA classifier [1–9]. In those systems, the role of universal background model (UBM) can be intuitively explained as acoustic space division where each mixture represents one distinct region. The posterior probabilities of features belonging to each mixture are further used for obtaining Baum-Welch statistics. Therefore, the accuracy of UBM posterior probabilities have direct influence on the overall performance of speaker recognition systems. Note that the notion "UBM posterior probabilities" used in this paper refers to mixture-wise posterior probabilities which are main bulding blocks of Baum-Welch statistics (see Sec. 2).

Previous studies have been attempted to improve the accuracy of UBM posterior probabilities by adapting UBM to the features of each utterance [10–13]. In [10], a vector taylor series (VTS) modeling based method was employed to adapt the UBM to noisy features, while in [13] uncertainties associated with noisy or enhanced features were also taken into account for adjusting extracted Baum-Welch statistic. The strategy of UBM adaptation for reliable Baum-Welch extraction have also proven to be successful in a recent study [12] where JFA-based front end was used. Above methods for improving the accuracy of Baum-Welch statistics are all based on UBM adaptation which works well when the mismatch between the data is significant [12] such as those under severe additive noise.

In this study, we propose to improve the UBM posterior probabilities by projecting the acoustic features into a new space where mixtures of trained UBM have optimized separability. The rationale behind this approach is that the UBM posterior probabilities are better predicted when those mixtures are well separated with each other. To find the new space, we first train an UBM using extracted acoustic features and then linear discriminant analysis (LDA) based method is applied by treating each mixture as independence class [14]. Therefore, LDA used in proposed method does not require any manually labeled information such as speaker identity or channel information [15]. The transformation matrix obtained from LDA training is then directly applied to acoustic features followed by a UBM retraining and standard i-Vector extraction system.

In Sec. 2, we present a short overview of the conventional i-Vector extraction framework as well as the role of UBM posterior probabilies. Sec. 3 contains the description of proposed method. In Sec. 4 and 5, we present results to show the effectiveness of proposed framework.

## 2. i-Vector extraction process

In this section, we present a short overview of the role of UBM posterior probabilities in i-Vector extraction framework. In a conventional i-Vector extraction framework, speaker and channel dependent GMM supervector is modeled as follows:

$$M = m + Tw, \tag{1}$$

where $m$ is the supervector obtained from the universal background model (UBM), $T$ is the low rank total variability matrix representing the basis of reduced total variability space, and $w$ is the low rank factor loadings referred to as i-Vectors.

The estimation of the total variability matrix $T$ employs expectation maximization (EM) method as described in [16]. After training the total variability matrix, the i-Vector of given speech utterance can be represented using Baum-Welch zeroth ($N_s$) and centralized first ($F_s$) order statistics

$$w_s^* = (T' N_s \Sigma^{-1} T + I)^{-1} T \Sigma^{-1} F_s, \tag{2}$$

where $\Sigma$ is the covariance matrix obtained from UBM model and $I$ is the identity matrix. Note that the zeroth ($N_s$) and centralized first ($F_s$) order statistics are expressed as

$$N_s = \begin{bmatrix} N_s^{C=1} & 0 & 0 & 0 \\ 0 & N_s^{C=2} & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & N_s^{C=c} \end{bmatrix}, \tag{3}$$

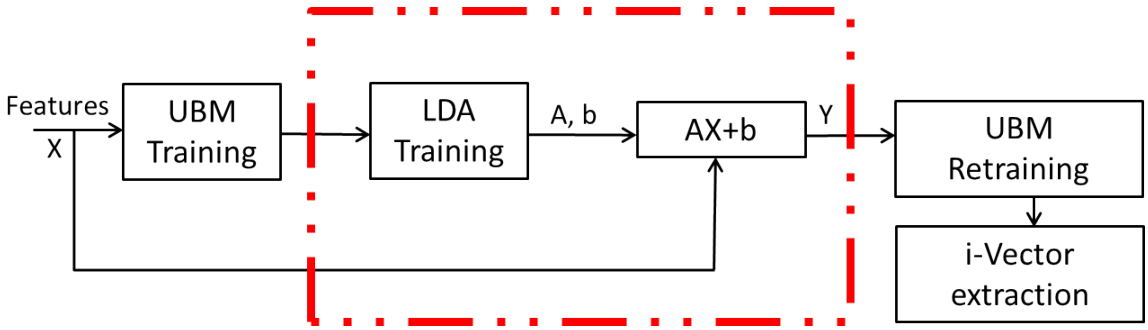


Figure 1: *Flowchart of proposed method.*

$$F_s = \begin{bmatrix} F_s^{C=1} & 0 & 0 & 0 \\ 0 & F_s^{C=2} & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & F_s^{C=c} \end{bmatrix}, \quad (4)$$

where

$$N_s^{C=c} = \sum_t P(c|X_t, \theta_{UBM}), \quad (5)$$

$$F_s^{C=c} = \sum_t P(c|X_t, \theta_{UBM})(X_t - \mu_c). \quad (6)$$

and $c$ is the index of UBM mixture component, $X_t$ is acoustic feature at time $t$, $\mu_c$ is the mean of $c$th Gaussian component. From above equations we could see that the mixture-wise UBM posterior probabilities $P(c|X_t, \theta_{UBM})$ directly determines the final extracted i-Vectors.

## 3. Proposed method

It is obvious from Sec.2 that the role of UBM in i-Vector extraction system is to separate acoustic space into different regions by means of Gaussian mixtures, In accordance, the UBM posterior probabilities are the predictions of features belonging to different regions. Therefore, it can be easily expected that if those regions are more separable with each other then it is easier to predict it. Motivated by this, we project the acoustic features into a space where separability of mixtures of trained UBM can be maximized. The transformation matrix is estimated using LDA from trained UBM by treating each Gaussian component as individual class. Derived transformation matrix is then applied directly to acoustic features followed by a UBM retraining process and standard i-Vector extraction system. The purpose of this acoustic feature transformation is to optimize the separability among Gaussian components of trained UBM. The flowchart of proposed method (uLDA) can be found in Fig.1.

### 3.1. UBM-based LDA (uLDA)

In proposed method, LDA is applied by treating each mixture of trained UBM as a separate class. Therefore, between-class scatter matrix can be obtained from the means of UBM mixtures as follow:

$$S_b = \sum_{C=1}^{c} p(c)(\mu_c - \mu)(\mu_c - \mu)', \quad (7)$$

where $p(c)$ and $\mu_c$ is the prior probability and mean of $c$th Gaussian component, and $\mu$ is the average of all Gaussian means. Similarly, within-class scatter matrix can be obtained from covariances of UBM mixtures

$$S_w = \sum_{C=1}^{c} p(c)\Sigma_c, \quad (8)$$

where $\Sigma_c$ is covariance matrix of $c$th Gaussian component. To optimize the separability of different Gaussian components of UBM, we need to maximize $S_b$ while minimizing $S_w$ at the same time. Therefore, the problem of maximizing separability of UBM mixtures becomes the classical LDA problem of finding transformation matrix $A$ that maximize objective function

$$J(A) = \frac{|S_b|}{|S_w|}. \quad (9)$$

Therefore, the colums of optimal trasformation matrix $A^*$ are eigenvectors corresponding to the largest eigen values of the following equation,

$$S_w^{-1} S_b A = \lambda A, \quad (10)$$

The obtained transformation matrix $A^\star$ which maximizes objective function $J(A)$ is then applied to the front end acoustic features such that

$$Y_t = A^\star X_t + b, \quad (11)$$

where b is the bias vector. The transformed features $Y$ are further used for UBM retraining as well as i-Vector extraction. Fig.2 shows an example of UBM before and after transformation.

## 4. Experiments

We evaluate the proposed method on the male part of core conditions of NIST SRE 2010 dataset. The baseline system is composed of 36-dimension feature vectors (12 MFCC + $\Delta$ + $\Delta\Delta$) extracted using a 25 ms window with 10 ms shift and normalized using a 3-s sliding window. In all experiments, an energy based voice activity detection (VAD) is applied. We used 1024-component gender dependent diagonal covariance universal background models (UBM). The HTK toolkit is employed for UBM training with 15 iterations for each split. The data used for training composed of data from Switch-board II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data. The i-Vectors of dimension 400 are extracted. This dimension is later reduced to 200 by LDA followed by length normalization and PLDA.

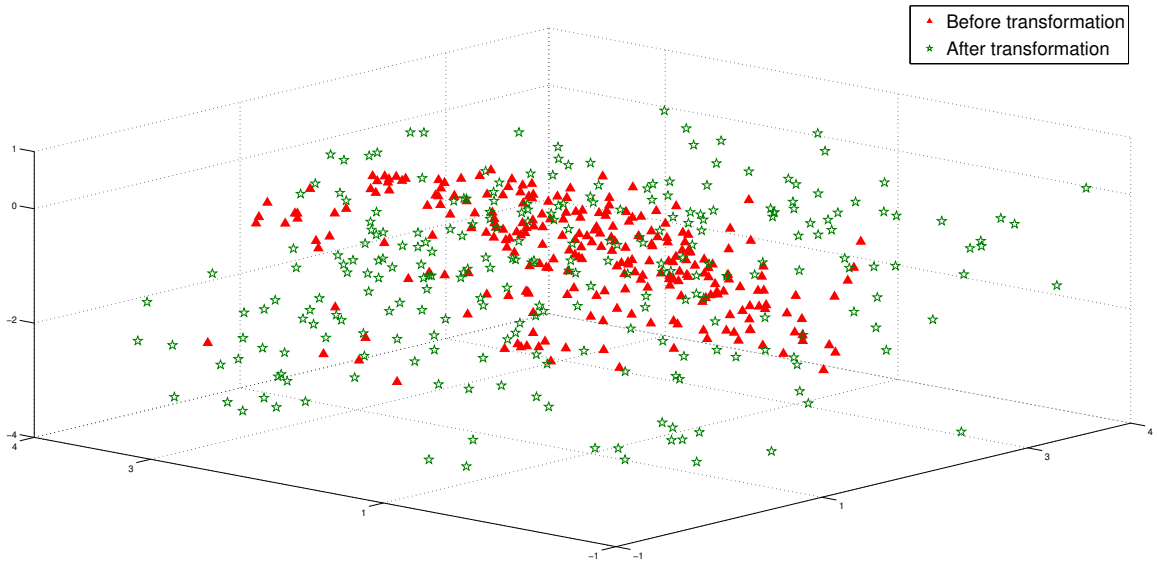


Figure 2: *Spread of UBM mixture means before and after uLDA transformation. Note that only first 3 dimensions of feature and first 256 mixtures of UBM are used as an illustration.*

| | Condition Names | **EER(%)** | | **MinDCF$_{08}$** | | **minDCF$_{10}$** | |
|---|---|---|---|---|---|---|---|
| | | baseline | uLDA | baseline | uLDA | baseline | uLDA |
| 1 | Int SameMic | 3.34282 | **3.01724** | 0.17024 | **0.16330** | 0.47768 | **0.44533** |
| 2 | Int DiffMic | 7.19855 | **6.59409** | 0.39667 | **0.35615** | 0.87779 | **0.85180** |
| 3 | Int Tel | 3.26646 | 3.38639 | 0.16717 | **0.15890** | 0.60918 | **0.55661** |
| 4 | Int Mic | 4.49156 | **4.03295** | 0.21925 | **0.20497** | 0.63019 | **0.54856** |
| 5 | Tel Tel | 2.35682 | **2.00587** | 0.12749 | **0.12738** | 0.39866 | **0.38244** |
| 6 | Tel High Vocal Effort | 2.60384 | 2.74409 | 0.16331 | 0.18359 | 0.56180 | 0.56181 |
| 7 | Mic High Vocal Effort | 3.48358 | 3.83206 | 0.17981 | 0.19094 | 0.39665 | 0.55858 |
| 8 | Tel Low Vocal Effort | 0.45945 | **0.45869** | 0.04201 | **0.03211** | 0.21689 | **0.16807** |
| 9 | Mic Low Vocal Effort | 1.42083 | 1.81505 | 0.05869 | **0.05776** | 0.12821 | 0.14358 |

Table 1: *Results from **male** part of core condions of NIST SRE 2010 dataset. uLDA indicates the proposed method of UBM based LDA.*

For evaluation of proposed method, the same parameter settings were used as baseline system except that the acoustic features were reduced to 32 dimensions with proposed uLDA method.

## 5. Results

The results of proposed method (uLDA) as well as baseline i-Vector system are summarized in Table. 1. The performace metric used are equal error rate (**%EER**), normalized minimum Detection Cost Function (DCF) from NIST SRE 2008 (**MinDCF$_{08}$**), and NIST SRE 2010 (**MinDCF$_{10}$**). The Detection Error Trade-off (DET) curves on condition 1 and 4 are plotted in Fig. 3 as an illustration.

The results show that the proposed method performs consistently better than baseline i-Vector system when evaluated on condition $1-5$ (normal vocal effort). The improvement is also consistent on all three performance metrics used in this study. The proposed method also shows some improvement on the condtion $8-9$ (low vocal effort) especially on condition 8. However, proposed method does not show improvment in performance when evaluated on condition $6-7$ (high vocal effort). We attribute this to the relatively high mismatch between the dataset we used to obtain transformation matrix and those for evaluation.

## 6. Discussion

We propose to improve the accuracy of UBM posterior probabilities by transforming the acoustic features into a new space where separabilities of trained UBM mixtures are maximized. The experiment result from NIST SRE10 dataset confirms the superior performance of proposed method compared to baseline i-Vector system.

The results of proposed uLDA method with only dimension of 32 is reported in this paper. However, simillar perfor-

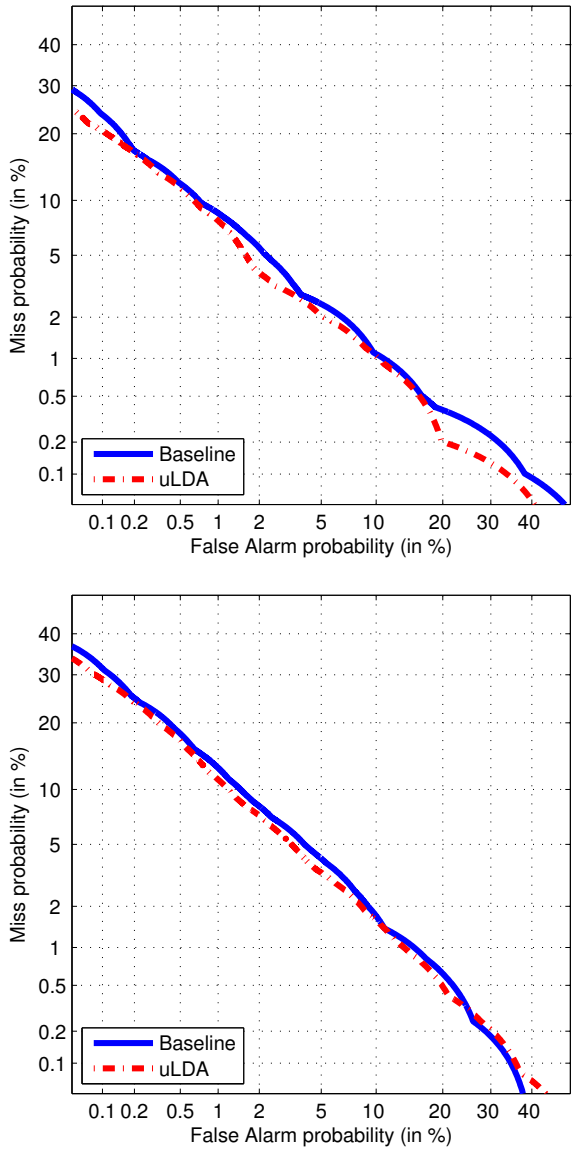


Figure 3: *Detection Error Trade-off (DET) curves on condition 1 and 4.*

mance was also observed when other reduced dimensions were applied. Automatic searching of the optimal dimesion for proposed uLDA method will be considered in our future work.

In this study, only MFCC features from single time frame are used. As proposed method is capable of reducing the dimension of acoustic features in an unsupervised fashion, a natural extention of this paper will be the incorporation of contextual information as well as ohter sources of features.

## 7. Acknowledgments

This research was supported by National Science Foundation (NSF) under Grant 1219130.

## 8. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[2] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *Proc. ICASSP, Florence, Italy*, 2014.

[4] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems.," in *Interspeech*, 2011, pp. 249–252.

[5] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. L. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *Proc. of ICASSP*, 2013, pp. 6784–6786.

[6] G. Liu, C. Yu, A. Misra, N. Shokouhi, and J. H. L. Hansen, "Investigating State-of-the-Art Speaker Verification in the Case of Unlabeled Development Data," *Proc. Odyssey speaker and language recognition workshop, Joensuu, Finland*, 2014.

[7] G. Liu, T. Hasan, H. Boril, and J. H. L. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Proc. of ICASSP*, 2013, pp. 7755–7759.

[8] O. Glembek, L. Burget, N. Brümmer, O. Plchot, and P. Matejka, "Discriminatively trained i-vector extractor for speaker verification.," in *INTERSPEECH*, 2011, pp. 137–140.

[9] G. Liu, Y. Lei, and J. H. L Hansen, "Robust feature front-end for speaker identification," in *Proc. of ICASSP*, 2012, pp. 4233–4236.

[10] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using Vector Taylor Series for speaker recognition," in *Proc. of ICASSP*, 2013, pp. 6788–6791.

[11] X. Zhao and Y. Dong, "Variational bayesian joint factor analysis models for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 1032–1042, 2012.

[12] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recognition," in *Proc. of ICASSP*, Florence, Italy, 2014.

[13] C. Yu, G. Liu, S. Hahm, and J. H. L. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Proc. of ICASSP*, Florence, Italy, 2014.

[14] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, "Feature enhancement with joint use of consecutive corrupted and noise feature vectors with discriminative region weighting," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2172–2181, 2013.

[15] Q. Jin and A. Waibel, "Application of LDA to speaker recognition.," in *INTERSPEECH*, 2000, pp. 250–253.

[16] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.