Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Investigation of Speaker-Clustered UBMs based on Vocal Tract Lengths and MLLR matrices for Speaker Verification

A. K. Sarkar and S. Umesh

Department of Electrical Engineering
Indian Institute of Technology Madras, India
sarkar.achintya@gmail.com, umeshs@iitm.ac.in

Abstract

It is common to use a single speaker independent large Gaussian Mixture Model based Universal Background Model (GMM-UBM) as the alternative hypothesis for speaker verification tasks. The speaker models are themselves derived from the UBM using Maximum a Posteriori (MAP) adaptation technique. During verification, log likelihood ratio is calculated between the target model and the GMM-UBM to accept or reject the claimant. The use of a single UBM for different groups of population may not be appropriate especially when the impostors are close to the target speaker. In this paper, we investigate the use of Speaker Cluster-wise UBM (SC-UBM) for a *group of target speakers* based on two different similarity measures. In the first approach, speakers are grouped into different clusters depending on their Vocal Tract Lengths (VTLs). The group of speakers having same VTL parameter indicates similarity in vocal-tract geometry and constitutes a speaker-dependent characteristic. In the second approach, we use Maximum Likelihood Linear Regression (MLLR) matrices of target speakers to create MLLR super-vectors and use them to cluster speakers into different groups. The SC-UBMs are derived from GMM-UBM using MLLR adaptation using data from the corresponding group of target speakers. Finally, speaker dependent models are adapted from their respective SC-UBM using MAP. In the proposed method, log likelihood ratio is calculated between target model and its corresponding SC-UBM. We compare performance of the above method with the single UBM method for varying number of clusters. The experiments are performed on the NIST 2004 SRE core condition and we show that the proposed method with a slight increase in the number of UBMs always outperforms the conventional single GMM-UBM system.

1. Introduction

Speaker verification is a binary decision problem. The claimant is accepted or rejected based on Log Likelihood Ratio (LLR), $\Lambda(X)$ calculated between claimant model (λ_c) and alternative hypothesis ($\lambda_{alt-hyp}$) for a given test feature (X). Mathematically,

$$\Lambda(X) = log\ Pr(X|\lambda_c) - log\ Pr(X|\lambda_{alt-hyp}) \quad (1)$$

If Log Likelihood Ratio is greater than the predefined threshold, then claimant is accepted, otherwise the claim is rejected.

There are several techniques available in the literature for selecting the alternative hypothesis. These can be broadly divided into two categories: one is cohort based [1, 2, 3] speaker-dependent technique and the other is speaker-independent Universal Background Modeling (UBM) [4]. In cohort based approach, either each speaker maintains a set of models (other closest speakers) called *cohorts* for alternative hypothesis or a single model obtained from the cohorts called *Individual Background Model (IBM)* [5]. During verification, the log likelihoods of the test utterance are calculated against the claimant model and the cohort set corresponding to the claimant model. In case of cohort, the likelihood values from cohort set are combined into a single value before LLR calculation. Several studies related to the combination of the likelihood values can be found in [1, 3, 6, 7]. In case of IBM, LLR is calculated between the claimant and the corresponding IBM.

Reynolds et al. [4] proposed the commonly used UBM technique, where a large Gaussian Mixtures Model (e.g. 2048 components) UBM (GMM-UBM) is trained using data from many speakers. Hence, GMM-UBM represents a speaker independent model in the feature space. The speaker models are then adapted from GMM-UBM using his/her training data by Maximum a Posteriori (MAP) adaptation technique. During verification stage, log likelihood ratio is calculated between the GMM-UBM and claimant model. The GMM-UBM is considered as the alternative hypothesis for all speakers in the database.

In [1, 7], it was argued that it is more logical for the closest speaker to the target to be in the cohort set to protect the system from false acceptance of close imposters. Hence the conventional GMM-UBM system, which maintains *a single UBM for all target speakers*, may not be appropriate to reject close imposters to the target in all cases.

In this paper, we propose to use a separate UBM for a group of target speakers (instead of a single UBM for all speakers) and refer to this approach as Speaker Cluster-wise UBM (SC-UBM). Since the SC-UBM models will be more specific to corresponding target speaker groups for which they are trained, they will be able to reject imposters close to the target. We propose to use the idea of Vocal Tract Lengths (VTLs) [8] to group the speakers into different clusters. VTL can vary from approximately 13cm for adult female speakers to over 18cm for adult male speakers. Speakers with same VTL factor have similar vocal-tract geometry and therefore the corresponding speech signals have similar spectral characteristics. VTL factor can therefore be used as a speaker-dependent characteristic.

In another approach to grouping of speakers, we cluster the target speakers based on the Maximum Likelihood Linear Regression (MLLR) matrices which are estimated for each speaker using the training data and the *speaker-independent GMM-UBM*. The columns of the MLLR matrices are stacked to form super-vectors. The speaker specific MLLR super-vector represents the speakers in high dimensional space. Then speakers are grouped into cluster using their super-vector by K-means

algorithm.

In our proposed method, the speakers are clustered into groups based on either of the similarity measures described above. Then separate UBM, called Speaker Clustered wise UBM (SC-UBM), are obtained for every speaker group from the *speaker-independent* GMM-UBM model. The target speaker dependent model for a speaker is then derived from his/her corresponding group dependent SC-UBM by MAP adaptation using his/her training data. The SC-UBMs will be close to the respective group of target speakers, and hence will be able to reject close imposters as proposed in [1, 7]. During verification, log likelihood ratio is calculated between claimant model and its corresponding SC-UBM.

The paper is organized as follows. In Section 2 & 3, we describe the method for target speaker clustering using VTLN factor and MLLR super-vector respectively. In Section 4, we describe how SC-UBMs are formed. In Section 5, the experimental setup is described. Section 6 describes the baseline system. The experimental results for our proposed method is compared with the baseline system in Section 7. Finally, in Section 8, we conclude the paper.

2. Vocal Tract Length

A major source of inter-speaker variability is due to differences in Vocal Tract Lengths (VTLs) among the speakers. If the vocal-tract is modeled as a uniform tube, then differences in VTL lead to scaling of the resonant frequencies. Therefore, the spectra of two speakers having different VTLs are related as follows:

$$S_A(f) = S_B(\alpha f) \tag{2}$$

where, α is the ratio of VTLs of speakers A and B. The α is also called VTLN (or VTL) factor or warp factor. Fig. 1 shows

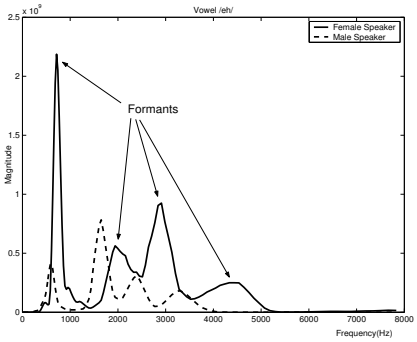

Figure 1: *The spectra of vowel /eh/ for male and female speaker.*

the smoothed spectra of vowel /eh/ enunciated by a male and a female speaker. To match the formants of male speaker to that of the female speaker (as reference), it is required to scale frequency axis, i.e. expand the spectra of the male speaker.

In practice, there exists no reference speaker with respect to whom α can be estimated. Therefore, a Maximum Likelihood (ML) based grid search is used to find the best warp factor for each speaker with respect to the speaker-independent (SI) model. We consider GMM-UBM as the SI model for α estimation,i.e.

$$\hat{\alpha} = \arg\max_{\alpha} Pr(X^{\alpha}|\lambda_{GMM-UBM}) \tag{3}$$

where, X^{α} is the warped feature obtained by scaling the spectra with α. Due to physiological constraints on the geometry of the vocal-tract, the ratio of VTLs are usually in the range of 0.80 to 1.20. In this paper, we use steps of 0.02 for grid-search in this range.

2.1. Speaker Clustering using VTLN factors

In this case, target speakers are grouped into clusters based on their VTLN factor. *Algorithm 1* describes the steps followed to form the speaker clusters using VTLN factors. Some VTLN related work can be found in [9, 10, 11].

Algorithm 1: Target speaker clustering using VTLN factor

Initial Step: Generate warped features of all utterances, $\{X_r^{\alpha}\}$, for a target speaker (r) for $\alpha \in [0.80, 1.20]$

Step 1: Estimate best $\hat{\alpha}$ for target speaker, r using,

$$\hat{\alpha}_r = \arg\max_{0.80 \le \alpha \le 1.20} Pr(X_r^{\alpha}|\lambda_{GMM-UBM})$$

Step 2: Repeat Step 1 for all target speakers.

Step 3: Group the target speakers into different clusters based on their VTLN factor α, i.e. cluster C_{α_j} contains speaker r if ML estimate of warping factor for speaker r is α_j. For example, if speaker r has VTLN factor 1.04, then it will belong to cluster $C_{1.04}$.

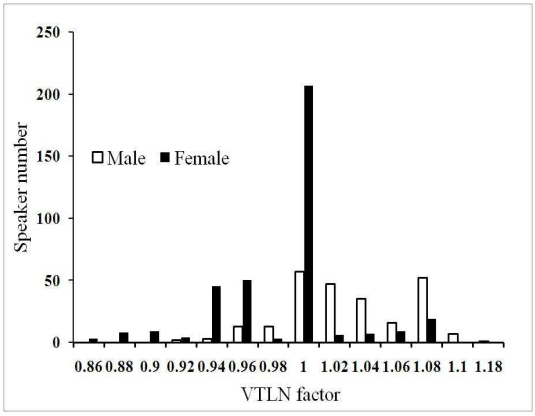

Figure 2: *Histogram of VTLN factor for male and female target speakers.*

Fig. 2 shows the histogram of VTL factor, α, for male and female target speakers. As expected, since male speakers generally have larger VTL than female speakers, most of the male speakers have α larger than 1.0 and most of the female speakers have α less than 1.0. Each target speaker cluster is formed by pooling the speakers who belong to the the same α class as shown in Fig. 3.

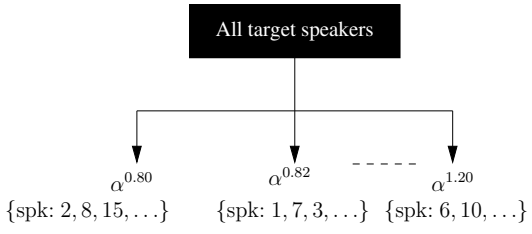

Figure 3: *Illustration of target speaker grouping/clustering based on their VTLN factor (α).*

In our experiment, we get 14 speaker groups based on VTLN factor on NIST 2004 SRE in core condition.

3. MLLR for Speaker Adaptation

Maximum Likelihood Linear Regression (MLLR) [12] is a commonly used technique in Automatic Speech Recognition (ASR) for speaker adaptation. In MLLR adaptation the mean vectors of the speaker independent (SI) model are linearly transformed and the transform is estimated using speaker adaptation data in a Maximum Likelihood framework. Mathematically,

$$\hat{\mu} = W\mu + b, \quad \hat{\Sigma} = \Sigma \tag{4}$$

where μ and Σ represent the mean and co-variance matrix of the SI model, and (W, b) are the MLLR transformation parameters. The transformed model parameters are $\hat{\mu}$ and $\hat{\Sigma}$.

3.1. Speaker clustering using MLLR super-vector

During training, MLLR transforms (W, b) are estimated for every target speaker in the training set with respect to the speaker independent GMM-UBM. The MLLR super-vectors are formed by concatenating the columns of their MLLR transformation similar to [13, 14]. This is illustrated in Fig. 4.

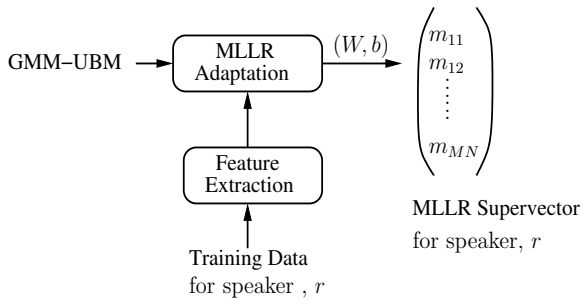

Figure 4: *MLLR super-vector concept.*

In [13, 14], it is shown that speaker specific super-vector obtained from MLLR or constrained-MLLR matrix contain speaker related information. Then, they use the speaker wise super-vector in Support Vector Machines (SVMs) environment for speaker recognition task. In our case, we just use the speaker wise super-vector (obtained from MLLR transformation) containing speaker specific information for grouping the target speakers in GMM-UBM framework. Then speaker clusters are formed from the MLLR super-vector using K-means algorithm. Euclidean distance measure is considered for similarity measure in K-means algorithm. The convergence of the K-means algorithm is achieved when cluster membership does not change. The steps are described in Algorithm 2.

Algorithm 2: Target speaker clustering using MLLR super-vector

Initial Step: Store MLLR super-vector, $\vec{M}_i$ from all speakers, (say, $1 \leq i \leq R$).

Step 1: Randomly select N super-vectors as cluster centroids, Ω_N

Step 2: Compute Euclidean distance from $\vec{M}_i$ to Ω_N, for $1 \leq i \leq R$.

Step 3: Assign $\vec{M}_i$ and corresponding speaker, i to the cluster with minimum distance.

Step 4. Update the cluster centroid

Step 5. Repeat the Step 2 to 4 until cluster membership does not alter.

In our experiments, we have studied the performance with different number of clusters, N, and have found that 5 clusters give the best performance for MLLR based clustering. A more detailed analysis is presented later in the paper.

4. Building SC-UBM

Irrespective of the clustering scheme (i.e. MLLR super-vector or VTLN factor), the SC-UBMs are formed as described in Algorithm 3.

Algorithm 3: SC-UBMs formation

Step 1: Load the feature vectors of all the training utterances, X_{C_j}, from all the target speakers in cluster C_j.

Step 2: Build a $SC - UBM^{C_j}$ for speaker cluster C_j with single iteration of MLLR adaptation from speaker-independent GMM-UBM using Eqn. (4),

$$\hat{\mu_{C_j}} = W_{C_j}\mu + b, \quad \hat{\Sigma_{C_j}} = \Sigma \tag{5}$$

$\hat{\mu_{C_j}}$ and $\hat{\Sigma_{C_j}}$ are the model parameters of $SC - UBM^{C_j}$.

Step 3: Repeat Step 1 to 2 for all clusters.

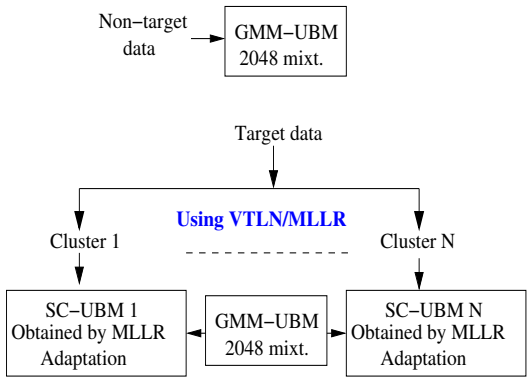

Figure 5: *Illustrates how SC-UBMs are derived for respective speaker group.*

Fig. 5 illustrates how cluster dependent SC-UBMs are built after clustering the target speaker using VTLN factor or MLLR super-vector. Speaker dependent models are then obtained from the corresponding SC-UBM using 2-iterations of MAP adaptation. Only mean adaptation is performed.

5. Experimental setup

All experiments are performed on NIST 2004 SRE core condition, i.e., single-side training and single-side test as per NIST evaluation plan [15]. The data set consists of 616 single-sided, single conversation for training 616 target speaker models. Each conversation is approximately 5 minutes long with 2.5 minutes of speech collected over various channels, handsets and languages. Details of the database can be found in [15].

The speaker independent GMM-UBM with 2048 mixture components with diagonal covariance matrices, is trained using data from non-target speakers in NIST 2002 SRE and Switchboard-1 Release -2. A 39 dimensional MFCC feature vector (C_1 to C_{13} with Δ and $\Delta\Delta$ coefficients, excluding C_0) is extracted at a frame rate of 10ms with 20ms Hamming windowed signal.

Two different frame removal techniques [16] are used to remove silence and frames with less energy. One is a bi-gaussian modeling of the energy component of the frames (for NIST 2002 SRE & Switchboard-1 Release-2 corpora) and the other is a tri-gaussian modeling of 0-mean and 1-variance normalized energy component of the frames (for NIST 2004 SRE). Finally, after removal of silence, features are normalized to fit a zero-mean and unit-variance distribution at utterance level.

6. Baseline System

The baseline system used in our experiment is the conventional GMM-UBM system proposed in [4]. The following three different systems are considered (as baselines) to compare the performance with our proposed method.

Baseline-I: Speaker independent GMM-UBM with 2048 Gaussian components is trained by pooling data from different population of non-target speakers [4].

Baseline-II: Gender dependent (Male and Female) GMM-UBMs of 1024 components are trained pooling data from gender-specific population of non-target speakers. The speaker independent GMM-UBM with 2048 mixtures is then derived from the gender dependent models by agglomerating the Gaussian components and renormalizing the mixture weights [4].

Baseline-III: Gender dependent GMM-UBMs with 2048 mixtures are derived from speaker independent GMM-UBM using single iteration of MLLR adaptation using data from respective gender of non-target speakers. Note that in this case, the speakers used to train the GMM-UBM are clustered according to gender, and the corresponding gender dependent UBM is built. However, the target speakers are *not clustered* and they are trained and tested using the corresponding gender-dependent UBM model.

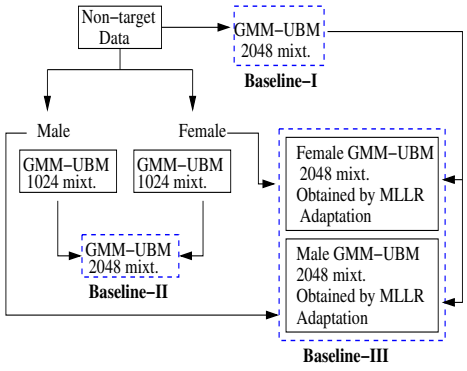

Figure 6: *Illustrates the Baseline systems.*

A schematic diagram of baseline systems are shown in Fig. 6. In case of Baseline I & II, speaker adapted models are derived from speaker independent GMM-UBM using two iterations of MAP. And in the case of Baseline III, speaker adapted models are derived from gender dependent GMM-UBM using two iterations of MAP. In all cases only means of the UBMs are adapted [4]. The value of relevance factor used for MAP is 16 in all experiments in this paper.

During verification, log likelihood ratio between claimant model and corresponding UBM (GMM-UBM or SC-UBM) were calculated using fast scoring technique described in [4]. Since, LLR calculation in speaker verification task is associated with target model and corresponding UBM , hence the proposed method requires same computation time as single GMM-UBM based speaker verification system. In our experiment, we considered top 15 best scoring mixture components for every feature vector.

7. Results and discussion

The different verification systems are evaluated using Equal Error rate (EER) and Minimum Detection Cost Function (MinDCF) as performance measures. EER value is calculated from the Detection Error Tradeoff (DET) curves [17]. The Detection Cost Function is defined as

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &\quad + C_{FA} \times P_{FA|NonTarget} \times (1 - P_{Target}) \end{aligned}$$

where $C_{Miss} = 10$, $C_{FA} = 1$ and $P_{Target} = 0.01$.

7.1. Effect of Number of Speaker-Clusters

Table 1: *Variation of EER and MinDCF values for different number of speaker clusters using VTLN factor.*

System	No. of clusters	EER(%)	MinDCF
SC-UBM using VTLN	4	14.32	0.0595
	6	14.05	0.0591
	8	14.05	0.0593
	10	14.10	0.0592
	14	13.96	0.0593

Table 2: *Variation of EER and MinDCF values for different number of speaker clusters using MLLR super-vector.*

System	No. of clusters	EER(%)	MinDCF
SC-UBM using MLLR	2	13.83	0.0578
	4	13.54	0.0566
	5	13.37	0.0565
	10	13.62	0.0564
	20	14.29	0.0558
	30	13.98	0.0569
	40	13.75	0.0575

Table. 1 & Table. 2 show the effect on EER and MinDCF as the number of target speaker clusters are varied for the two clustering methods. In the case of VTL factor based clustering, we get 14 clusters on NIST 2004 SRE core condition (i.e. only 14 values of α have non-empty clusters). Then number of clusters are reduced from 14 to 4 by re-estimating the VTLN factor (α) with a new set of α values by *iteratively* removing that α value which has the least number of speakers. In the case of MLLR super-vector clustering, different number of clusters are formed

Table 3: Comparison of EER and MinDCF for different Systems on NIST 2004 SRE core condition.

System	EER (%)	(%) improv. by		MinDCF	(%) improv. by	
		(a)	(b)		(a)	(b)
Baseline-I	15.42	9.47	13.29	0.0597	0.67	5.36
Baseline-II	15.28	8.64	12.50	0.0589	-0.67	4.07
Baseline-III	15.07	7.37	11.28	0.0597	0.67	5.36
SC-UBM using VTLN (a)	13.96	-	-	0.0593	-	-
SC-UBM using MLLR (b)	13.37	-	-	0.0565	-	-

by initializing the cluster centroids with the desired number of clusters and building the clusters from scratch.

From Table. 1 & 2, the following observation can be made:

- In the case of VTLN factor based clustering, the best results are obtained using 14 clusters, and the trend is improving performance as the number of clusters increase. The degraded performance for lesser number of speaker groups could be due to different VTLs being assigned to the same class. This leads to speaker variability within the same clusters leading to an increase in EER.
- Clustering using MLLR super-vector shows the best result for 5 speaker-cluster. The EER value increases as the number of clusters are increased beyond 5. This may be due to clusters being split even for speakers with similar characteristic.
- MLLR super-vector based clustering gives better performance than the VTLN based clustering, although both methods provide significant improvement over the conventional single-UBM based method as shown in the DET curves and tables below.

Fig. 7 shows the DET curves of the two proposed methods and the three baseline systems described in Sec. 6. We have used 14 clusters for VTL method and 5 clusters for MLLR-super-vector method since they provide the best performance. In Table. 3 the EER and MinDCF performances of our proposed methods and

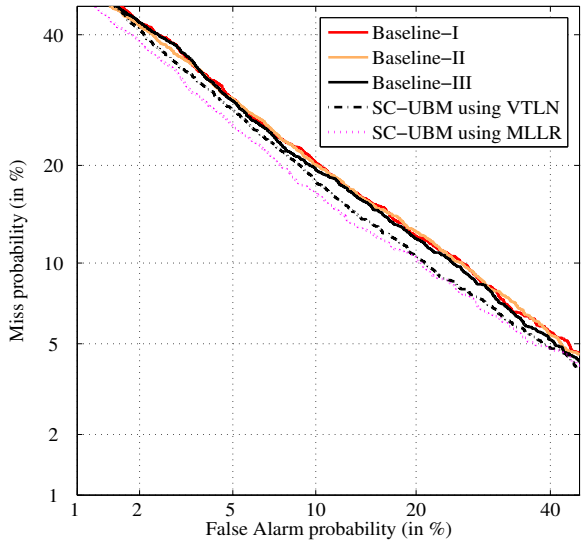

Figure 7: *Comparison of Baseline systems with SC-UBM based systems on NIST 2004 SRE core condition.*

From Table. 3, the following observations can be made:

- The proposed method always gives lower EER value compared to all baseline systems. VTLN-wise SC-UBM system shows MinDCF which is comparable to baseline systems. But SC-UBM system using MLLR super-vector shows significant relative improvement of minDCF value over the baseline.
- SC-UBM system using MLLR super-vector performs better than VTLN factor based SC-UBM system.

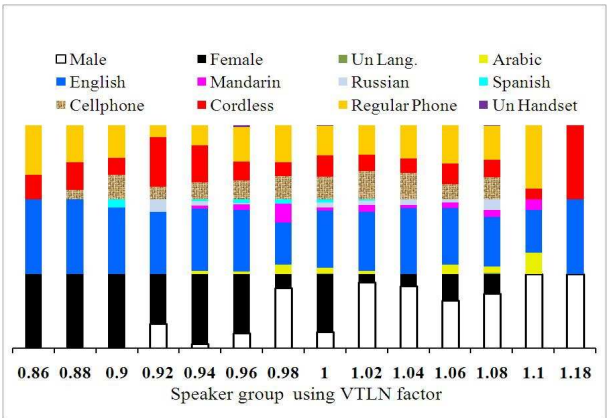

Figure 8: *Distribution of data in each speaker group using VTLN factor.*

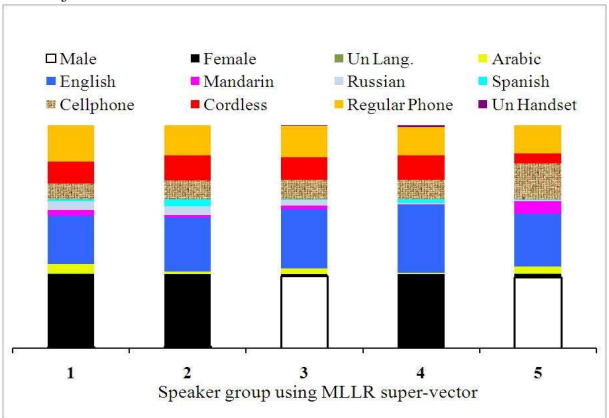

Figure 9: *Distribution of data in each speaker group using MLLR super-vector.*

7.2. Gender-wise Speaker Clustering

To investigate why MLLR-super-vector based system performs better than VTLN-based target speaker clustering, we plot the distribution of gender, language, handset etc. within each

speaker cluster in Fig. 8 & 9. The most striking observation from the figures is that the MLLR super-vector wise clustering separates the male and female target speakers into distinct clusters. On the other hand, in many of the VTLN based speaker clusters there is a mix of the both male and female target speakers for the same VTLN factor. This is possible, since male and female speakers may have same physical geometry especially in the neighborhood of α being unity which corresponds to an "average" VTL.

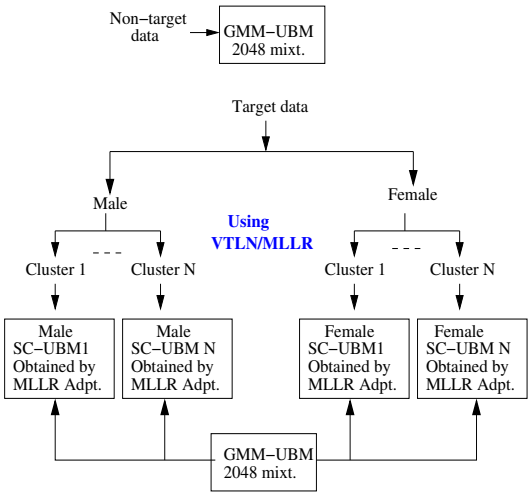

Figure 10: *Illustrates SC-UBMs are formed after clustering the target speaker in gender-wise using VTLN factor/MLLR super-vector.*

Therefore, we perform another set of experiments where we cluster the target speakers *within each gender* using VTLN factor and MLLR super-vector as shown in Fig. 10. In this case, VTLN factor, α and MLLR super-vector are estimated with respect to the *gender-dependent GMM-UBM* Baseline-III system. Please note that the Baseline-III system corresponds to *partitioning the (non-target) "train" speakers according to gender*. In the proposed SC-UBM methods, we partition the target speakers into different clusters. We now describe the gender-dependent speaker-clustering of target speakers below.

Table 4: *Variation of EER and MinDCF values for gender wise speaker clusters using VTLN factor.*

System	No. of clusters	EER(%)	MinDCF
SC-UBM using VTLN	2M+2F	13.50	0.0575
	4M+4F	13.54	0.0575
	6M+6F	13.61	0.0577

Table 5: *Variation of EER and MinDCF values for gender wise speaker clusters using MLLR super-vector.*

System	No. of clusters	EER(%)	MinDCF
SC-UBM using MLLR	2M+2F	13.14	0.0562
	4M+4F	13.33	0.0557
	6M+6F	13.13	0.0566

7.2.1. Effect of Number of cluster

Table. 4 & 5 show the effect of the number of clusters within each gender class (M-male, F-female). As seen from the tables, there is an improvement in performance in both methods of speaker-clustering. However, MLLR super-vector is still marginally better than VTLN based method. Since splitting each gender into two clusters seem to give the best performance in both methods, we will use two cluster per gender in all subsequent experiments.

Table 6: *Performance of gender wise speaker verification for 2M and 2F SC-UBM per gender.*

System	No. of clusters	EER(%)	MinDCF
SC-UBM using VTLN	2M	13.08	0.0520
	2F	13.79	0.0618
SC-UBM using MLLR	2M	12.62	0.0494
	2F	13.45	0.0612

From Table. 6, it is observed that in both cases male speakers performance is better than female speakers.

7.2.2. Performance of gender-wise clustering

We now compare the performance of our proposed method with the conventional single-UBM method. Further, since we are using gender-wise clustering, it is also useful to show the performance of our proposed method with a simple gender-wise splitting of target speakers. We consider the following systems in our experiments:

(i) **Two cluster gender-dependent SC-UBM: SC-UBM (M, F)** Target speakers are grouped into male and female clusters. Two gender dependent SC-UBMs are derived from speaker-independent GMM-UBM using data from respective speakers in the gender cluster. The SC-UBMs and speaker-dependent models are built using the method described before.

(ii) **Four cluster gender-dependent SC-UBM using VTLN factor: VTLN (2M, 2F)** Male and female speakers are separately clustered into two groups using VTLN factor, α. Hence, we get four SC-UBMs – 2 VTLN-based cluster for each gender.

(iii) **Four cluster gender-dependent SC-UBM using MLLR super-vector: MLLR (2M, 2F)** It is similar to VTLN system, the only difference is that MLLR super-vector is used for grouping the speakers.

Fig. 11 shows the DET curves of the systems (i)-(iii) with baseline systems. Table. 7 shows the performance of the systems in term of the EER and MinDCF value. From Fig. 11 and Table. 7 following observations can be noted:

- System (i), (ii) and (iii) performs better than all the baseline systems.
- System (ii) & (iii) are better than (i).
- Gender dependent, VTLN factor wise speaker grouping shows similar performance to using MLLR super-vector for speaker clustering.
- All SC-UBM system perform consistently better than the earlier case i.e. without gender wise target speaker clustering.

Table 7: Comparison of EER and MinDCF for different Systems on NIST 2004 SRE core condition.

System	EER (%)	(%) improv. by			MinDCF	(%) improv. by		
		(i)	(ii)	(iii)		(i)	(ii)	(iii)
Baseline-I	15.42	9.66	12.45	14.79	0.0597	3.18	3.69	5.86
Baseline-II	15.28	8.84	11.65	14.01	0.0589	1.87	2.38	4.58
Baseline-III	15.07	7.56	10.42	12.81	0.0597	3.18	3.69	5.86
SC-UBM (M, F) (i)	13.93	-	-	-	0.0578	-	-	-
SC-UBM using VTLN (2M, 2F) (ii)	13.50	-	-	-	0.0575	-	-	-
SC-UBM using MLLR (2M, 2F) (iii)	13.14	-	-	-	0.0562	-	-	-

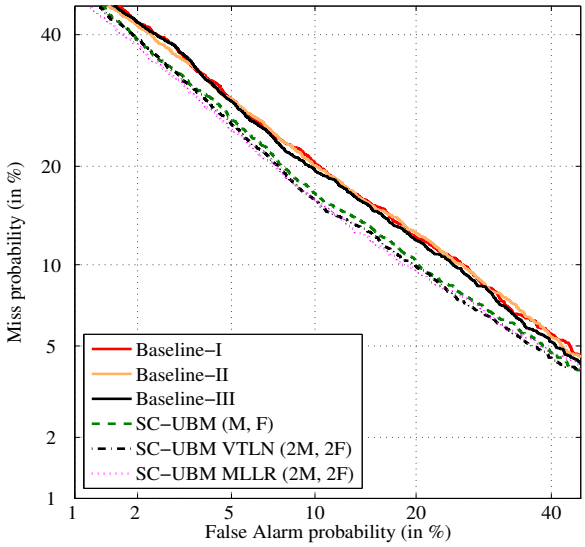

Figure 11: *Comparison of Baseline systems with gender wise SC-UBM systems on NIST 2004 SRE core condition.*

It is important to note that speaker clustering using VTLN factor or MLLR super-vector is done in *training phase* to form the SC-UBM and corresponding speaker models from their respective SC-UBM. The LLR calculation during verification is associated with target model and corresponding SC-UBM. Hence, the proposed method requires same amount of computation as single GMM-UBM based speaker verification system.

8. Conclusion

In this paper, we have proposed the use of a separate background model for each groups of speakers, i.e. speaker cluster wise UBM (SC-UBM). We have investigated clustering of speakers using their Vocal Tract Length factor as well as MLLR super-vectors. The SC-UBMs are derived from speaker independent GMM-UBM using data from the respective speaker cluster. Finally, speaker dependent model are adapted from their respective SC-UBM using MAP. The experiments are performed on NIST 2004 SRE core condition. Experimental results show that SC-UBM systems achieve lower EER and MinDCF over conventional single-UBM baseline systems. Further, we show that using gender-wise speaker-clustering provides additional gain in performance. Therefore, we conclude, that for a small increase in the number of background models, we get a significant improvement in speaker-verification performance.

9. Acknowledgment

A part of this work is supported by SERC project funding `SR/S3/EECE/058/2008` from the Department of Science & Technology, Ministry of Science & Technology, India.

10. References

[1] A. E. Rosenberg, J. DeLong, CH Lee, BH Jaung, and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," *ICSLP*, 1992.

[2] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker Verification," *ICASSP*, 1996.

[3] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[4] D.A. Reynolds, T.F. Quateri, and R.B. Dunn, "Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, Jan2000.

[5] Yossi Bar-Yosef and Yuval Bistritz, "Adaptive Individual Background Model for Speaker Verification," *Interspeech*, 2009.

[6] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *DSP*, vol. 1, pp. 89–106, 1991.

[7] D. Tran and M. Wagner, "A Proposed Likelihood Transformation for Speaker Verification," *ICASSP, Turkey*, 2000.

[8] Li Lee and Richard C. Rose, "Speaker Normalization using Efficient Erequency Warping Procedures," *ICASSP*, vol. 1, pp. 353–356, 1996.

[9] D. R. Sanand, D. Dinesh Kumar, and S. Umesh, "Linear Transformation Approach to VTLN using Dynamic Frequency Warping," *Interspeech 2007, Belgium*, 2007.

[10] P. T. Akhil, S. P. Rath, S. Umesh, and D. R. sanand, "A Computationally Efficient Approach to Warp Factor Estimation in VTLN Using EM Algorithm and Sufficient Statistics," in *Interspeech 2008, Australia*, 2008.

[11] D. R. Sanand and S. Umesh, "Study of Jacobian Compensation Using Linear Transformation of Conventional MFCC for VTLN," in *Interspeech 2008, Australia*, 2008.

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

[13] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "Mllr Transforms as Features in Speaker Recognition," *Eurospeech*, pp. 2425–2428, 2005.

[14] M Ferras, Cheung Chi Leung, C Barras, and J L Gauvain, "Constrained MLLR for Speaker Recognition," *ICASSP-2007*, vol. 4, pp. 53–56, 2007.

[15] The Evaluation Plan of NIST 2004 Speaker Recognition Campaign. http://www.itl.nist.gov/iad/mig//tests/sre/2004/SRE04_evalplan-v1a.pdf, ," .

[16] Jean Francois Bonastre, Nicolas Scheffer, Corinne Fredouille, and Driss Matrouf, "Nist'04 Speaker Recognition Evaluation Campaign: New LIA Speaker Detection Plateform based on ALIZE Toolkit," *in NIST SRE'04 Workshop, Toledo, Spain*, Jun. 2004.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowskiand, and M. Przybocki, "The Det Curve in Assessment of Detection Task Performance," *In Proceedings of the European Conference on Speech Communication and Technology*, pp. 1895–1898, 1997.