LIMSI/VOCAPIA SPEAKER VERIFICATION SYSTEM FOR NIST SRE 2012

A. K. Sarkar[1], V. B. Le[2], C.-T. Do[1], A. Roy[1], C. Barras[1], L. Lamel[1] and J.-L. Gauvain[1]

[1]LIMSI-CNRS, Université Paris-Sud, BP 133, 91403 Orsay, France
[2]Vocapia Research, 28 Rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France
[sarkar,ctdo,roy,barras]@limsi.fr, levb@vocapia.com

LIMSI and Vocapia Research developed two main speaker verification systems which were combined for submission to the NIST SRE 2012 core condition: a GSV-PCA system and a Lattice MLLR-based *m-vector* system. Both are super-vector based.

1. GSV-PCA SYSTEM

In this system, target speakers are represented by *Speaker Characterization Vectors (SCVs)* which are obtained by projecting their Gaussian Mixture Model (GMM) super-vectors [1] from adapted models on a vector space as,

$$SCV_r = (\mathbf{GSV_r})^t.\mathbf{P} \tag{1}$$

where SCV_r represents the SCV of the r^{th} target speaker obtained by projecting his/her GMM super-vector, GSV_r, on vector space, $\mathbf{P}$.

The vector space, $\mathbf{P}$, is built by Principal Component Analysis (PCA) of pooled GMM super-vectors from many speakers and can be thought as analogous to the *total variability space* in state-of-the-art speaker verification system using i-vector concept [2]. During test, SCV of the test utterance is scored against the claimant specific SCV obtained during training. Before scoring, SCVs are conditioned for session variability compensation.

GMM super-vector is calculated with respect to a Universal Background Model (UBM) with 3 iterations of MAP adaptation technique for a given speech data. We use a UBM with 512 Gaussian components and 47 dimensional features in this system. It gives $512 \times 47 = 24064$ dimensional GMM-super-vector. The vector space $\mathbf{P}$ is calculated by PCA using 21621 utterances (i.e. 21621 GMM super-vectors) from 1871 speakers over various SRE databases.

For session variability compensation, we use recently proposed Eigen Factor Radial (EFR) algorithm in [3] for conditioning the SCVs. EFR iteratively normalize the length of the SCV (i.e. w) to handle the *session variability compensation* as in Eq.(2).

$$\hat{w} \leftarrow \frac{V^{-\frac{1}{2}}(w - \overline{w})}{\sqrt{(w - \overline{w})^t V^{-1}(w - \overline{w})}} \tag{2}$$

where $\hat{w}$ represents the normalized SCV. V and $\overline{w}$ denote the covariance matrix and mean vector of the training SCVs, respectively, in the successive iterations.

During the test, Mahalanobis distance measure is used for scoring between the two normalized SCVs (i.e. $\hat{w_1}, \hat{w_2}$) as,

$$score(\hat{w_1}, \hat{w_2}) = (\hat{w_1} - \hat{w_2})^t \Omega^{-1} (\hat{w_1} - \hat{w_2}) \tag{3}$$

where Ω is the within-class covariance matrix calculated using development data set.

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

2. LATTICE MLLR M-VECTOR SYSTEM

Maximum Likelihood Linear Regression (MLLR) [4] is commonly used for speaker adaptation in Hidden Markov Model (HMM)-based ASR systems. It estimates an affine transformation (A, b) with respect to a Speaker Independent (SI) HMM in Maximum Likelihood (ML) sense. It can be expressed as,

$$\hat{\mu_s} = A\mu_s + b; \quad \hat{\Sigma_s} = \Sigma_s \tag{4}$$

where μ_s and Σ_s are the Gaussian mean and covariance matrix of the state s in SI model, respectively. It is well known to that MLLR transformation (i.e. (A, b)) contains speaker related information and is commonly used in super-vector [5] form for speaker recognition. Fig.1 graphically illustrates the MLLR super-vector estimation with respect to SI HMM of r^{th} speaker using his/her speech data. Generally, MLLR super-vectors are used for speaker modeling in a Support Vector Machine (SVM) framework, and an Automatic Speech Recognition (ASR) front-end is used for estimating several MLLR transformations for a given (speaker) speech segment with respect to pre-defined phonetic classes. Several variant of speaker recognition system based on MLLR super-vector can be found in literature [6, 7].

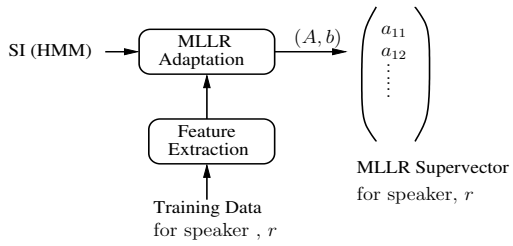

Fig. 1. *MLLR super-vector extraction from MLLR transformation of the r^{th} speaker using his/her training data with respect to a speaker independent HMM.*

Recently, in [8, 9] has been proposed a different way of speakers characterization by their MLLR super-vectors than conventional approach. The proposed method in [9] is called *m-vector*, where speakers are represented by *m-vectors* which are obtained by uniform segmentation of their MLLR super-vectors using an overlapped sliding window. It gives several m-vectors to represent a speaker and each *m-vector* is processed separately which constitutes several subsystems. Fig.2 graphically illustrates the m-vectors extraction of r^{th} speaker from his/her MLLR-super-vector with overlapped sliding window of 500 elements. During test, *m-vectors* of the test utterance are scored against the claimant specific *m-vectors*. Before scoring,

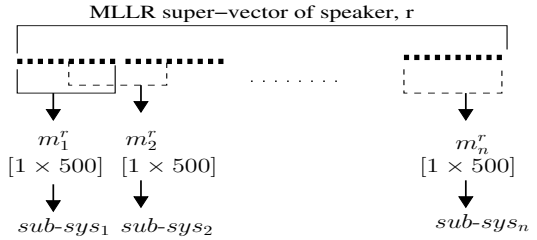

Fig. 2. *m-vector extraction for the r^{th} speaker from his/her MLLR super-vector using an overlapped sliding window of* 500 *elements with* 50% *overlap of its adjacent m-vectors.*

m-vectors are conditioned by EFR algorithm for session variability compensation. It is shown in [9] that *m-vector* system shows promising performance with compared to state-of-the-art i-vector system even though it used UBM for estimating the MLLR transformation without any phonetic knowledge of the speech segment.

For NIST SRE 2012, we enhanced the *m-vector* approach by integrating an ASR system. In order to be more robust to transcription errors, latticed-based MLLR transforms are estimated using the word-level lattice output of the ASR, converted into a phonetic graph. Multi-class MLLR super-vectors are then extracted with respect to phonetic classes (vowels and consonants). In this system, we use 42 dimensional feature vectors which gives $2 \times 42 \times 42 = 3528$ dimensional MLLR super-vector (without bias); more details about the use of lattice MLLR for speaker verification can be found in [10].

3. EXPERIMENTAL SETUP

3.1. GSV-PCA system

47 dimensional PLP feature vectors (15 static with their Δ, $\Delta\Delta$, ΔE and $\Delta\Delta$E) are extracted from the speech signal at 10 ms rate over the 0-3800 Hz bandwidth. Voice activity detection is then applied on the feature vectors to discard the less energized or silence frames. Finally, energized frames are normalized to zero mean and unity variance at utterance level. Two gender dependent UBMs having 512 Gaussian components with diagonal covariance matrices, are trained using data from NIST 2004 SRE.

For noise compensation, we use Feature Mapping (FM) [11] technique. As per feature mapping, we derive different noise level dependent GMM models (from -15 dB to $+15$ dB with increment of 5 dB over babble i.e. crowd and car noise) from noise independent UBM with a single iteration of MAP adaptation using respective level noisy data. These data were obtained by artificially adding [12] noise from the NOISEX-92 to clean speech signals, at different SNR levels. During training and test phase, first best noise model of the utterance is selected using the top-10 scoring, then feature vectors are mapped from the best selected noise dependent model space to noise independent model space i.e. UBM space using top-1 decoding on respective systems (male and female).

For PCA and EFR, 21621 utterances (target training example plus 12399 utterances of 890 non-target speakers) over 1871 speakers (including targets who have more than 5 examples in training) are used. Non-target data are collected from NIST 2004-2005, Switchboard II part 1, 2 & 3, Switchboard cellular part 1 & 2, with about 15 sessions per speaker. In MAP adaptation, the value of relevance factor 10 is considered for all systems.

3.2. m-vector system

For spectral analysis, 42 dimensional feature vectors including 12 Mel-PLP feature, log-energy and F_0 along with their first- and second-order derivatives are extracted from the speech signal each 10 ms using a 30 seconds Hamming window over bandwidth 0-3800Hz. Voice activity detection is applied as a pre-processing step to discard less energized or silent frames. Finally, detected speech segments are normalized to zero mean and unit variance at the utterance level.

The Large Vocabulary Continuous Speech Recognition (LVCSR) system used MLLR transforms estimation is similar to the LIMSI RT'04 LVCSR system [13]. The acoustic models are trained on about 2000 hours of manually transcribed Conversational Telephone Speech (CTS) data using the PLP+F_0 features concatenated with additional MLP features [14]. The model sets cover about $48k$ phone contexts, with $11.5k$ tied states and 32 Gaussians per state. Silence is modeled by a single state with 1024 Gaussians. Two manually derived phonetic classes: vowels and consonants are used for MLLR transformations, resulting in a 42×42 dimensional MLLR transformation for each phonetic class (the bias b is discarded since it does not provide significant gain in our setup). Totally, we get a $(2 \times 42 \times 42) = 3528$ dimensional MLLR super-vector.

In this system, Linear Discriminant Analysis (LDA) is applied on the m-vectors to discriminant the speakers before conditioning. Each *m-vector* sub-system has its own LDA projection matrix. For session variability compensation, two iterations of the EFR are applied on the *m-vectors*, associated to the Mahalanobis distance measure for scoring. Scores of the different sub-systems are fused for a particular LDA dimension across all sub-systems, with equal weights given to all sub-systems. The data set for LDA and EFR algorithm are used as same as in GSV-PCA system.

4. RESULTS ON DEVELOPMENT DATA SET

Table 1 shows the comparison of speaker verification performance using *m-vector* technique for different approach of MLLR super-vector estimation on NIST 2008 SRE core condition (male speakers) over various tasks. The performance of the overlapped *m-vector* systems are shown for *m-vector* size of 500 elements which correspond to size of the sliding window. *Disjoint -full* case speakers are represented by their full MLLR super-vectors [8]. Here, only 890 non-target speakers data (i.e. 12399 utterances) are used for LDA and EFR, which are totally disjoint from NIST 2008 SRE. The system performances are shown using Equal Error rate (EER) and Minimum Detection Cost Function (MinDCF) as per 2008 SRE evaluation [16]. For fusion, equal weights are given to all systems. In case of UBM system, MLLR super-vector for a given utterance is derived from a global MLLR transformation which is estimated with respect to UBM (without any transcriptions). It gives 1764 dimensional MLLR super-vector (without bias).

Table 1. *Comparison of performance of speaker verification with* m-vector *technique using different approach of MLLR super-vector estimation on NIST 2008 SRE core condition (male speakers) over various tasks.*

System	m-vector extraction method	m-vector size	Optimal LDA dim.	DET task: (%) EER (MinDCF) 1	3	4	5	6	7
UBM (Baseline)	(A1) Disjoint- full	1764	50	15.00 (0.0614)	15.54 (0.0641)	15.55 (0.0612)	10.53 (0.0467)	9.32 (0.0485)	6.67 (0.0362)
	(A2) Overlapped	500	50	14.92 (0.0588)	15.34 (0.0611)	13.00 (0.0514)	9.55 (0.0380)	7.70 (0.0378)	5.74 (0.0271)
	Fusion (A1,A2)	-	-	**13.95 (0.0557)**	**14.37 (0.0579)**	**12.63 (0.0491)**	**8.55 (0.0353)**	**7.70 (0.0382)**	**5.51 (0.0259)**
ASR 1 best	(B1) Disjoint- full	3528	50	12.86 (0.0492)	13.30 (0.0510)	10.36 (0.0451)	8.45 (0.0337)	6.34 (0.0395)	3.89 (0.0214)
	(B2) Overlapped	500	50	12.51 (0.0480)	12.88 (0.0498)	9.61 (0.0417)	7.69 (0.0293)	5.89 (0.0354)	3.18 (0.0160)
	Fusion(B1,B2)	-	-	**11.82 (0.0445)**	**12.09 (0.0464)**	**8.77 (0.0407)**	**7.34 (0.0274)**	**5.89 (0.0346)**	**3.13 (0.0144)**
ASR Lattice	(C1) Disjoint- full	3528	50	11.98 (0.0470)	12.46 (0.0487)	9.39 (0.0410)	8.20 (0.0322)	7.05 (0.0379)	3.42 (0.0180)
	(C2) Overlapped	500	50	11.92 (0.0455)	12.25 (0.0471)	8.52 (0.0392)	7.09 (0.0260)	5.74 (0.0351)	2.97 (0.0162)
	Fusion(C1,C2)	-	-	**11.21 (0.0426)**	**11.52 (0.0439)**	**8.07 (0.0382)**	**6.75 (0.0259)**	**5.54 (0.0345)**	**2.88 (0.0147)**

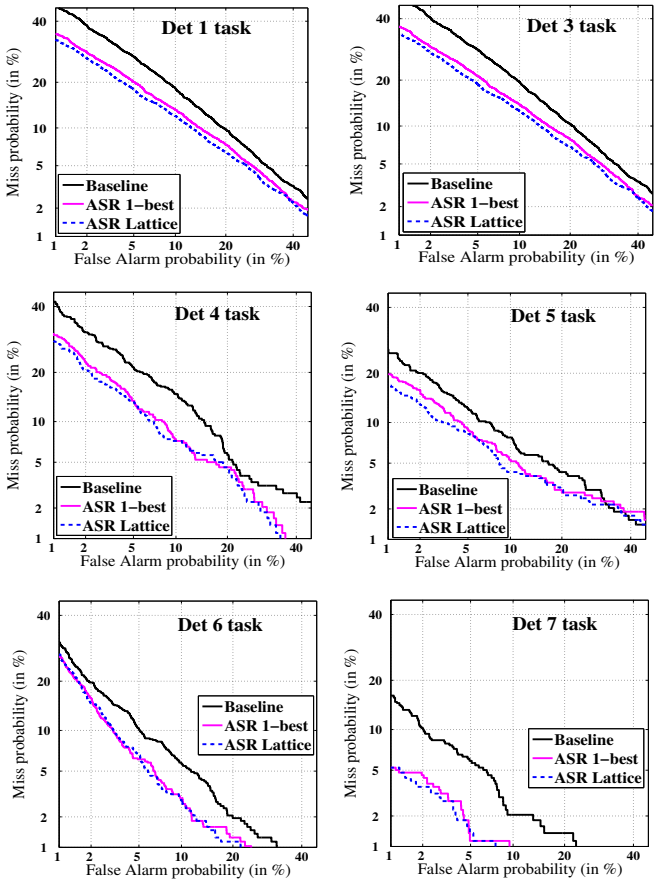

Fig. 3. *Comparison of performance of speaker verification of* m-vector *systems (fusion) for different approach of MLLR super-vector extraction on NIST 2008 SRE core condition over different tasks.*

Fig.3 compares the Detection Error Tradeoff (DET) plots of the respective *m-vector* systems (fusion) on NIST 2008 SRE core condition over various tasks. From Table 1 and Fig.3, it is observed that *m-vector* system with phonological knowledge i.e. ASR shows significantly better performance than conventional UBM based system. And lattice based system also reflects the accountability of erroneous in speech transcription for MLLR transformation by reducing the error rate over the conventional 1-best hypothesis method [4]. More details on this system can be found in [15].

Table 2 shows the performance of speaker verification with GSV-PCA and *m-vector* techniques on NIST 2012 SRE development data set. The training and testing dataset are developed from the target speakers (who have multiple sessions for training) training example by dividing them into two disjoint parts: three randomly chosen sessions per speaker are taken for the test set, three other for a validation set, and the remaining sessions are kept for training the models. The validation set allowed to verify that the performance is stable according to the selection of the sessions in the test set. It results respectively, 36000 and 52944 test trials (approximatively 5% true trials, the remaining being impostor trials) for male and female experiments. In this case, PCA and EFR are built using the data mentioned in experimental setup (Sec.3) of the respective systems. The performance of GSV-PCA system and *m-vector* systems are shown for respectively, PCA 800 (i.e. SCV size) and LDA 50 dimensions. System performances are measured as per NIST 2012 SRE cost function [17] and EER.

5. SUBMITTED SYSTEMS

Scores of both systems were converted to log-likelihoods using piece-wise linear transformation estimated on the development set.

The primary system submitted by LIMSI and Vocapia Research to NIST SRE 2012 for the Core test condition is a score-level linear fusion between both systems with combination weights optimized on the development set.

Table 2. *Comparison of performance of speaker verification of GSV-PCA and m-vector systems on NIST 2012 SRE development dataset.*

	System	$c_{norm_{a1}}$	$c_{norm_{a2}}$	$c_{primary}$	EER (%)
Male	ASR-Lattice: m-vector	0.0952	0.1645	0.1298	0.99
	GSV-PCA	0.0771	0.1755	0.1263	0.91
Female	ASR-lattice: m-vector	0.2190	0.4045	0.3118	2.10
	GSV-PCA	0.1420	0.2785	0.2102	1.44
Male + Female	ASR-Lattice-m-vector	0.1702	0.3102	0.2402	1.58
	GSV-PCA	0.1159	0.2372	0.1766	1.16
	Fusion(GSV,m-vector)	**0.1092**	**0.2169**	**0.1630**	**0.87**

The contrastive system consists in the Lattice MLLR *m-vector* system; for some trials, no speech was detected in the audio may be due to the low SNR and the system could not provide a score; in this case, scores computed using a generic UBM-based MLLR *m-vector* system were used instead.

6. REFERENCES

[1] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Process. Lett.*, vol. 13, pp. 308–311, 2006.

[2] N. Dehak et al., "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[3] 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.

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

[5] A. Stolcke et al., "MLLR Transforms as Features in Speaker Recognition," in *Proc. of EUROSPEECH*, 2005, pp. 2425–2428.

[6] Z. N. Karam and W. M. Campbell, "A Multi-class MLLR Kernel for SVM Speaker Recognition," in *Proc. of ICASSP*, 2008, pp. 4117–4120.

[7] M. Ferras et al., "Constrained MLLR for Speaker Recognition," in *Proc. of ICASSP*, 2007, pp. 53–56.

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

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

[10] M. Ferras, C. Barras, and J. L. Gauvain, "Lattice-based MLLR for Speaker Recognition," in *Proc. of ICASSP*, 2009, pp. 4537–4540.

[11] D.A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2003, pp. 6–10.

[12] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton:FLN, 2007.

[13] R. Prasad et al., "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System," in *Proc. of INTERSPEECH*, 2005, pp. 1645–1648.

[14] P. Fousek, L. Lamel, and J. L. Gauvain, "Transcribing Broadcast Data using MLP Features," in *Proc. of INTERSPEECH*, 2008, pp. 1433–1436.

[15] A. K. Sarkar, C. Barras, and V. B. Le, "Lattice MLLR based m-vector system for speaker verification," *ICASSP*, 2013 (submitted).

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

[17] The NIST Year 2012 Speaker Recognition Evaluation Plan, "http://www.nist.gov/itl/iad/mig/upload/nist_sre12_evalplan-v17r1.pdf," .