LIMSI/VOCAPIA SPEAKER VERIFICATION SYSTEM FOR NIST SRE 2012

A. K. Sarkar\textsuperscript{1}, V. B. Le\textsuperscript{2}, C.-T. Do\textsuperscript{1}, A. Roy\textsuperscript{1}, C. Barras\textsuperscript{1}, L. Lamel\textsuperscript{1} and J.-L. Gauvain\textsuperscript{1}

\textsuperscript{1}LIMSI-CNRS, Université Paris-Sud, BP 133, 91403 Orsay, France
\textsuperscript{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 \cite{1} from adapted models on a vector space as,

\[ SCV_r = (GSV_r)\cdot P \] (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, \( P \).

The vector space, \( P \), 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 \cite{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 \( P \) 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 the recently proposed Eigen Factor Radial (EFR) algorithm in \cite{3} for conditioning the SCVs. EFR iteratively normalizes the length of the SCV (i.e. \( w \)) to handle the session variability compensation as in Eq.(2).

\[ \hat{w} \leftarrow \frac{V^{-\frac{1}{2}}(w-m)}{\sqrt{(w-m)^tV^{-1}(w-m)}} \] (2)

where \( \hat{w} \) represents the normalized SCV, \( V \) and \( m \) 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,

\[ \text{score}(\hat{w}_1, \hat{w}_2) = (\hat{w}_1 - \hat{w}_2)^t \Omega^{-1}(\hat{w}_1 - \hat{w}_2) \] (3)

where \( \Omega \) 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) \cite{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,

\[ \hat{\mu}_s = A\mu_s + b; \quad \hat{\Sigma}_s = \Sigma_s \] (4)

where \( \mu_s \) and \( \Sigma_s \) are the Gaussian mean and covariance matrix of the state \( s \) in SI model, respectively. It is well known that MLLR transformation (i.e. \( (A, b) \)) contains speaker related information and is commonly used in super-vector \cite{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 \cite{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.](image)

Recently, in \cite{8, 9} has been proposed a different way of speakers characterization by their MLLR super-vectors than conventional approach. The proposed method in \cite{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,
**3. EXPERIMENTAL SETUP**

### 3.1. GSV-PCA system

47 dimensional PLP feature vectors (15 static with their $\Delta$, $\Delta\Delta$, $\Delta E$ 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 UBM models are used for the GSV-PCA system. 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 $42 \times 42$ dimensional MLLR super-vector for each phonetic class (the bias $b$ is discarded since it dose 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, 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.

### 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 $42 \times 42$ dimensional MLLR super-vector 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.

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 the $m$-vectors. 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.
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.

<table>
<thead>
<tr>
<th>System</th>
<th>m-vector extraction method</th>
<th>Optimal LDA dim.</th>
<th>DET task (%)</th>
<th>EER</th>
<th>MinDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>UBM (Baseline)</td>
<td>(A1) Disjoint- full</td>
<td>1764</td>
<td>50</td>
<td>15.00</td>
<td>0.0614</td>
</tr>
<tr>
<td></td>
<td>(A2) Overlapped</td>
<td>500</td>
<td>50</td>
<td>14.92</td>
<td>0.0588</td>
</tr>
<tr>
<td></td>
<td>Fusion (A1,A2)</td>
<td>-</td>
<td>-</td>
<td>13.95</td>
<td>0.0557</td>
</tr>
<tr>
<td>ASR 1 best</td>
<td>(B1) Disjoint- full</td>
<td>3528</td>
<td>50</td>
<td>12.86</td>
<td>0.0492</td>
</tr>
<tr>
<td></td>
<td>(B2) Overlapped</td>
<td>500</td>
<td>50</td>
<td>12.51</td>
<td>0.0480</td>
</tr>
<tr>
<td></td>
<td>Fusion (B1,B2)</td>
<td>-</td>
<td>-</td>
<td>11.82</td>
<td>0.0445</td>
</tr>
<tr>
<td>ASR Lattice</td>
<td>(C1) Disjoint- full</td>
<td>3528</td>
<td>50</td>
<td>11.97</td>
<td>0.0470</td>
</tr>
<tr>
<td></td>
<td>(C2) Overlapped</td>
<td>500</td>
<td>50</td>
<td>11.92</td>
<td>0.0455</td>
</tr>
<tr>
<td></td>
<td>Fusion (C1,C2)</td>
<td>-</td>
<td>-</td>
<td>11.21</td>
<td>0.0426</td>
</tr>
</tbody>
</table>

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, 30000 and 50004 test trials (approximately 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.

<table>
<thead>
<tr>
<th>System</th>
<th>$\epsilon_{\text{norm}_{a1}}$</th>
<th>$\epsilon_{\text{norm}_{a2}}$</th>
<th>$\epsilon_{\text{primary}}$</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male ASR-Lattice: m-vector</td>
<td>0.0952</td>
<td>0.1045</td>
<td>0.1298</td>
<td>0.99</td>
</tr>
<tr>
<td>Male GSV-PCA</td>
<td>0.0771</td>
<td>0.1755</td>
<td>0.1263</td>
<td>0.91</td>
</tr>
<tr>
<td>Female ASR-lattice: m-vector</td>
<td>0.2190</td>
<td>0.4045</td>
<td>0.3118</td>
<td>2.10</td>
</tr>
<tr>
<td>Female GSV-PCA</td>
<td>0.1420</td>
<td>0.2785</td>
<td>0.2102</td>
<td>1.44</td>
</tr>
<tr>
<td>Male + Female ASR-Lattice-m-vector</td>
<td>0.1702</td>
<td>0.3102</td>
<td>0.2402</td>
<td>1.58</td>
</tr>
<tr>
<td>Male + Female GSV-PCA</td>
<td>0.1159</td>
<td>0.2372</td>
<td>0.1766</td>
<td>1.16</td>
</tr>
<tr>
<td>Fusion(GSV,m-vector)</td>
<td>0.1092</td>
<td>0.2169</td>
<td>0.1630</td>
<td>0.87</td>
</tr>
</tbody>
</table>

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


