LIA SYSTEM DESCRIPTION FOR NIST SRE 2016

Mickael Rouvier, Pierre-Michel Bousquet, Moez Ajili, Waad Ben Kheder
Driss Matrouf, Jean-François Bonastre

LIA, Université d'Avignon, France

ABSTRACT

This paper describes the LIA speaker recognition system developed for the Speaker Recognition Evaluation (SRE) campaign. Eight sub-systems are developed, all based on a state-of-the-art approach: i-vector/PLDA which represents the mainstream technique in text-independent speaker recognition. These sub-systems differ: on the acoustic feature extraction front-end (MFCC, PLP), at the i-vector extraction stage (UBM, DNN or two-feats posteriors) and finally on the data-shifting (IDVC, mean-shifting). The submitted system is a fusion at the score-level of these eight sub-systems.

1. INTRODUCTION

This paper describes the systems developed by the LIA for the 2016 National Institute of Standard and Technology (NIST) Speaker Recognition Evaluation (SRE).

LIA developed eight sub-systems which are described in this paper. The eight sub-systems are based on i-vector/PLDA paradigm. I-vector/PLDA paradigm is the state-of-the-art approach in speaker recognition. The i-vector approach provides an elegant way of reducing a large-dimensional input vector (representing the speaker data) to a small-dimensional feature vector [1]. I-vectors are extracted on total variability space, no distinction is made between speaker and channel variation. Probabilistic Linear Discriminant Analysis (PLDA) is used to disentangle speaker effects from other sources of undesired variability [2, 3].

The sub-systems are constructed by combining different front-end and back-end. The sub-systems differ from acoustic features (MFCC or PLP), i-vector extraction (UBM, DNN or two-feats posteriors) and data-shifting (IDVC or mean-shifting). The submitted system is a fusion at the score-level of these eight sub-systems. All the sub-systems are mainly based on the open-source ALIZE toolkit, freely available [1].

This paper is organized as follows: In Section 2, we present the different datasets used in our systems. We briefly describe our system in Section 3. Details of the submitted systems and the components are described in Section 4. The results for the individual sub-systems and the fused system are presented in Section 5. Finally in Section 6 we show the CPU time and memory requirements for computing the score of one verification trial.

[1] http://alize.univ-avignon.fr

2. TRAIN AND DEVELOPMENT DATASETS

The Table 1 lists the different datasets that we used for training the system.

Dataset(s)	Task
LDC2016E46_SRE16_Cal_My_Net	Development Set
LDC2016E46 (unlabeled)	i-vector normalization
SRE'04, 05, 06, 08 Switchboard-2 Phase II Switchboard-2 Phase III Switchboard Cellular Part 1 Switchboard Cellular Part 2	UBM, T matrix, PLDA, IDVC
Switchboard-1 Release 2	DNN

Table 1. Datasets used for training the system.

3. OVERVIEW OF THE SYSTEM

The different steps of our i-vector systems can be summarized as follows:

- **Feature extraction** : Two acoustics features are extracted : Mel-Frequency Ceptral Coefficients (MFCC) and Perceptual Linear Prediction (PLP).
- **Voice activity detection (VAD)** : remove silence and low energy speech based on the C0 component.
- **I-vector extraction** : three different kind of i-vectors are extracted (GMM/i-vector, DNN/i-vector and two-feats/i-vector). These i-vectors differ from the manner in which the statistics are collected.
- **Pre-normalization** : The raw i-vectors extracted are first whitened and length-normalized (**LW**-normalization).
- **Data-shifting** : considering the language mismatch between training and development corpus. Two data-shifting methods are used in order to compensate this mismatch : Inter Dataset Variability Compensation (IDVC) and mean-shifting.

- **PLDA learning** : a non-classical PLDA is learned on the i-vectors. This PLDA is fully described in Section 4.6.

- **Post-normalization** : The between- and within-class covariance matrix etimated by the PLDA are diagonalized and a new **LW**-normalization is applied on the i-vectors.

- **PLDA scoring** : a verification score is calculated.

4. SYSTEM COMPONENTS

4.1. Acoustic features

Before feature extraction, all waveforms are first downsampled to 8 kHz, and blocked into 25 ms frames with a 10 ms skip-rate. Two acoustic features are extracted using Kaldi toolkit [4] : MFCC and PLP. All features use 20 cepstral coefficients and log-energy, appended with the first and second order time derivatives, thus providing 60 dimensional feature vectors. A cepstral mean normalization is applied with a window size of 3 seconds.

4.2. Voice Activity Detection

VAD removes silence and low energy speech segments. A simple energy-based VAD is used based on the C0 component of the acoustic feature. The algorithm is based on thresholding the log-energy and taking the consensus of threshold decisions within a window of 11 frames centred on the current frame.

4.3. i-vector

An i-vector extractor is a data-driven front-end that maps temporal sequences of feature vectors (e.g., MFCC or PLP) into a single point in a low-dimensional vector space. This is accomplished by collecting sufficient statistics. In these evaluation the statistics are obtained from : Universal Background Model (UBM), Deep Neural Network (DNN) and two-feats. All the extracted i-vectors are 400-dimension.

4.3.1. UBM

The generation of i-vectors requires use of UBM which modelizes the generation of the acoustic features (cepstrum + first and second derivatives). The UBM used here is a GMM (Gaussian Mixture Model) of 4096 Gaussians, where each Gaussian is characterized by its mean and its full-covariance matrix. The UBM is trained on SRE'04-08 and Switchboard corpus, using the standard EM algorithm.

4.3.2. DNN

In [5], authors propose to collect statistic by using a DNN that are trained to classify phoneme states. DNN is trained with 4 hidden layers. The input layer takes 60 dimensional MFCC features with 7-frame temporal context and cepstral mean subtraction (CMS) performed over a window of 6 seconds. Each hidden layer has 1024 nodes. The ouput dimension is 4096 senone. The forced alignment between the state-level transcripts and the corresponding speech signals by the GMM/HMM triphon system is used to generate labels for DNN training.

4.3.3. two-feats

It is well known that the sucess key of the i-vector paradigm is the robustness of the a-posteriori probabilities estimation against the UBM. To increase the robustness of this estimation, we propose to use two acoustic features streams instead of only one: PLP based stream and MFCC based stream. To do so, we firstly estimate a PLP based UBM and for each frame, we generate the a posteriori probabilities. Then, we use the MFCC frames and these last a posteriori probabilities to estimate the parameters of the MCCF based UBM (one EM iterartion). At the end of this process, we obtain two UBMs having the same topology: same number of Gaussians with correspondance between pairs of Gaussians having the same index in the two UBMs (PLP and MFCC). From now on, the a posteriori probablities for a given frame is obtained by combining the ones coming from the PLP-UBM and MFCC-UBM. Lets, $P_{mfcc} = [p_1^{mfcc}, ..., p_{4096}^{mfcc}]$ and $P_{plp} = [p_1^{plp}, ..., p_{4096}^{plp}]$, the final a posteriori porobabilities are given by:

$$p_i = \frac{p_i^{plp} * p_i^{mfcc}}{\sum_j p_j^{plp} * p_j^{mfcc}}$$

4.4. Pre-normalization

I-vectors are whitened and length-normalized in order to make them more Gaussian, and also to reduce the shift between training and test data, as remarked in [6]. The whitening technique we use for NIST SRE 2016 evaluation is a standardization according to the within-class covariance matrix **W**, as proposed in [1, 7]. We denote by **LW** this transformation (Length-normalization of **W**-standardized vectors).

4.5. Data-shifting

4.5.1. Inter dataset variability compensation

In order to reduce the shifts of language and gender, we include in our system the Inter Dataset Variability Compensation (IDVC) technique, as described in [8]. This technique seeks to compensate eventual mismatches, between training

Table 2. Details of the subsets used for IDVC method

subset	language	native language	gender	#segments	#speakers
1	english	english	F	13934	774
2	english	english	M	10379	504
3	english	non english	F	6087	784
4	english	non english	M	9326	517
5	non english	all	F	4576	663
6	non english	all	M	2868	413
additional	—	—	all	2272	—

and test data, by removing unexpected variability of model parameters. We apply this technique to limit the uncertainty of mean and within-speaker covariance matrix involved by gender and language mismatches. The IDVC method is trained on 6 subsets of segments from NIST SRE 2004, 2005, 2006, 2008 and the additional subset of development data from the evaluation major language provided by NIST (the latter only for mean-subspace removal, as this subset is unlabeled). Table 2 details the content of these subsets.

4.5.2. Mean-shifting

Mean-shifting calculates the mean of the Call-My-Net development data and subtract it to the test i-vectors.

4.6. PLDA

4.6.1. Learning

LIA systems use the PLDA learning proposed in the Kaldi toolkit [9, 4]. Given a set of n_s vectors from a training speaker s, this model assumes that the centered mean vector m_s of speaker s can be decomposed as:

$$m_s = x_s + y_s$$

where

$$\begin{aligned} x_s &\sim \mathcal{N}(0, \mathbf{B}) \\ y_s &\sim \mathcal{N}\left(0, \frac{1}{n_s}\mathbf{W}\right) \end{aligned} \tag{1}$$

$\mathbf{B}$ (resp. $\mathbf{W}$) denoting the between (resp. within)-class covariance matrix and $\mathcal{N}(.)$ the Gaussian pdf. Thus, this model takes into account some uncertainty about the speaker mean position, depending on the size n_s of its training set.

Starting from deterministic estimations of $\mathbf{B}$ and $\mathbf{W}$, an EM-like iterative algorithm is applied to optimize these matrices. It can be shown that the distributions of x_s and y_s a posteriori of m_s are Gaussian:

$$x_s|m_s \sim \mathcal{N}(w_s, \mathbf{M}_s) \tag{2}$$
$$y_s|m_s \sim \mathcal{N}(m_s - w_s, \mathbf{M}_s) \tag{3}$$

where

$$\begin{aligned} \mathbf{M}_s &= \left(\mathbf{B}^{-1} + n_s\mathbf{W}^{-1}\right)^{-1} \\ w_s &= n_s\mathbf{M}\mathbf{W}^{-1}m_s \end{aligned} \tag{4}$$

It can also be shown that the contributions of this speaker-class to the between and within-class covariance are respectively equal to:

$$E\left[x_s x_s^t\right] = w_s w_s^t + \mathbf{M}_s \tag{5}$$
$$E\left[y_s y_s^t\right] = (m_s - w_s)(m_s - w_s)^t + \mathbf{M}_s \tag{6}$$

Thus, $\mathbf{B}$ and $\mathbf{W}$ can be updated as follows:

$$\begin{aligned} \mathcal{B} &= \frac{1}{S}\sum\nolimits_s \left(w_s w_s^t + \mathbf{M}_s\right) \\ \mathcal{W} &= \frac{1}{N}\sum\nolimits_s n_s \left((m_k - w_k)(m_k - w_k)^t + \mathbf{M}_k\right) \end{aligned}$$

where S is the number of training speakers and $N = \sum_s n_s$ is the total amount of observations. The $(\mathbf{B}, \mathbf{W})$ learning algorithm is described below:

4.6.2. Algorithm

S = number of classes; n_s = number of observations for speaker s
$N = \sum_s n_s$ total amount of observations
Compute initial $\mathbf{B}$ and $\mathbf{W}$
$m_{all} = \frac{1}{S}\sum_s m_s = \frac{1}{S}\sum_s \left(\frac{1}{n_s}\sum_{x \in s} \mathbf{w}\right)$
for $iter = 1$ to $nb\_iterations$
  $\mathcal{B} = 0; \mathcal{W} = 0$
  for each speaker s
    $m_s \leftarrow m_s - m_{all}$ (centered mean)
    $\mathbf{M}_s = \left(\mathbf{B}^{-1} + n_s\mathbf{W}^{-1}\right)^{-1}$
    $w_s = n_s\mathbf{M}_s\mathbf{W}^{-1}m_s$
    $\mathcal{B} \leftarrow \mathcal{B} + w_s w_s^t + \mathbf{M}_s$
    $\mathcal{W} \leftarrow \mathcal{W} + n_s(m_s - w_s)(m_s - w_s)^t + n_s\mathbf{M}_s$
  $\mathbf{B} \leftarrow \mathcal{B}/S$; $\mathbf{W} \leftarrow \mathcal{W}/N$

Feat/I-vec/Shifting	Mandarin (CMN)				Cebuano (CEB)				Equalized		Unequalized	
	Male		Female		Male		Female		Male+Female		Male+Female	
	EER(%)	minC	EER(%)	minC	EER(%)	minC	EER(%)	minC	EER(%)	minC	EER(%)	minC
1 - MFCC/UBM/IDVC	6.19	0.323	14.56	0.644	22.92	0.820	23.16	0.840	16.71	0.657	16.67	0.721
2 - MFCC/UBM/Mean	7.76	0.34	18.64	0.689	25.08	0.849	22.85	0.812	18.58	0.672	19.03	0.699
3 - PLP/UBM/IDVC	5.04	0.223	15.49	0.654	23.17	0.855	25.19	0.881	17.22	0.65	17.36	0.779
4 - PLP/UBM/Mean	7.07	0.282	16.79	0.657	24.10	0.892	23.29	0.803	17.81	0.658	18.33	0.732
5 - MFCC/DNN/IDVC	5.47	0.292	16.75	0.710	27.50	0.937	25.71	0.878	18.86	0.704	18.58	0.783
6 - MFCC/DNN/Mean	6.75	0.310	20.88	0.727	27.56	0.960	25.17	0.865	20.09	0.716	20.97	0.768
7 - MFCC/two-feats/IDVC	6.15	0.267	15.32	0.607	24.08	0.844	23.20	0.832	17.19	0.638	17.44	0.729
8 - MFCC/two-feats/Mean	7.56	0.296	18.60	0.641	25.58	0.858	23.49	0.784	18.81	0.644	19.02	0.687

Table 3. Performance of our 8 single systems: MFCC/PLP/DNN/two-feats with two techniques of data-shifting: IDVC or mean-shifting (substraction of mean).

System	Mandarin (CMN)				Cebuano (CEB)				Equalized		Unequalized	
	Male		Female		Male		Female		Male+Female		Male+Female	
	EER(%)	minC	EER(%)	minC	EER(%)	minC	EER(%)	minC	EER(%)	minC	EER(%)	minC
Fusion A	5.95	0.235	15.24	0.546	22.77	0.841	22.43	0.788	16.60	0.602	16.71	0.688
Fusion B	5.39	0.224	15.64	0.574	22.83	0.845	22.39	0.794	16.56	0.609	16.90	0.669

Table 4. Performance of fusion by mean of scores from single systems 1 to 4 (Fusion A) and 1 to 8 (Fusion B).

4.7. Post-normalization

In [2], a post-PLDA normalization procedure is proposed that simultaneously diagonalizes between- and within-class covariance provided by the PLDA learning described above. Given $\mathbf{B}$ and $\mathbf{W}$ matrices estimated by PLDA learning, the following transformation is applied:

- Transform data: $\mathbf{w} \rightarrow \mathbf{W}^{-\frac{1}{2}}\mathbf{w}$ (or, equivalently, $\mathbf{w} \rightarrow \mathbf{L}^{-1}\mathbf{w}$ where $\mathbf{W} = \mathbf{L}\mathbf{L}^t$ is the Cholesky decomposition of $\mathbf{W}$),
- Compute SVD of $\mathbf{W}^{-\frac{1}{2}}\mathbf{B}\mathbf{W}^{-\frac{1}{2}} = \mathbf{P}\mathbf{\Psi}\mathbf{P}^t$, where $\mathbf{P}$ is the eigenvector matrix and $\mathbf{\Psi}$ is the diagonal matrix of eigenvalues,
- Rotate data: $\mathbf{w} \rightarrow \mathbf{P}^t\mathbf{w}$.

By this way, matrices $\mathbf{B}$ and $\mathbf{W}$ become diagonal matrices $\mathbf{\Psi}$ and $\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

As observed in [10], we assume that length-normalizing the test data after this procedure will contribute to further reduce the shift between training and test data. Thus, our post-normalization procedure turns out to be equivalent to the $\mathbf{LW}$-normalization described above.

It can be noticed that, after this post-normalization which diagonalizes covariance matrices, it is shown in [2] that reducing the dimensionality of the between-class variability (*eigenvoice* subspace rank) can be easily achieved by keeping the largest elements of $\mathbf{\Psi}$ and setting the rest to zero. Thus, this transformation allows fast estimation of PLDA parameters with various eigenvoice ranks.

5. LIA SUBMISSION RESULTS

Total 8 sub-systems are constructed by various front-end and back-end combinations as summarized in Table 3. The Equal Error Rate (EER) and minC$_{primary}$ (minC) cost functions obtained from these systems are shown, also detailed performance by gender and language (Cebuano and Mandarin).

We note that the principal metric is minC$_{primary}$ and therefore the systems and sub-systems are optimized on this metric.

Table 4 shows the performance gain obtained by fusion of single system scores. "Fusion A" is the score obtained by the mean of scores (equal weights) from single systems 1 to 4 presented in Table 3. We note that the system is the primary system of NIST SRE'16. "Fusion B" is the score obtained by the mean of scores from the 8 single systems. We note that the system is one of the secondary systems of NIST SRE'16.

6. COMPUTATION TIME AND MEMORY

Table 5 shows the CPU time and memory requirements for computing the score of one verification trial (for the subsystem 2). All tasks were implemented in C++.

7. CONCLUSION

We have described the LIA site speaker recognition system submitted to the 2016 NIST SRE. The systems developed were a fusion of i-vector based sub-systems using different front-end and back-end.

Table 5. Computation time and memory consumption of various part of the system to produce the score of one verification trial. All tasks were performed on a 64-bit Linux server with 512G RAM and an AMD Opteron Processor 6238 running at 2.6GHz. All CPU times are based on one core of the processor.

Task	Task Name	CPU Time (sec.) per Utt.	Memory consumption (MB)
1	MFCC Extraction	0.91	3.6
2	Voice Activity Detection	0.75	7.3
3	Computing statistics	1.33	810
4	I-vector Extraction	0.66	810
5	PLDA Scoring	0.01	54
	Overall	3.66	–

8. REFERENCES

[1] Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Sergey Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision*, pp. 531–542, 2006.

[3] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[5] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1695–1699.

[6] Sandro Cumani and Pietro Laface, "I-vector transformation and scaling for plda based speaker recognition," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, pp. 39–46.

[7] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldřich Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012.

[8] Hagai Aronowitz, "Compensating inter-dataset variability in plda hyper-parameters for robust speaker recognition," in *Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

[9] "Kaldi toolkit," http://kaldi-asr.org/.

[10] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.