The Idiap Speaker Recognition Evaluation System at NIST SRE 2012

Elie Khoury, Laurent El Shafey, Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland
{elie.khoury, laurent.el-shafey, sebastien.marcel}@idiap.ch

***Abstract*—In this paper, we present the Idiap Research Institute submission to the 2012 NIST Speaker Recognition Evaluation. Our system is based on the Inter-Session Variability (ISV) modelling technique. The implementation of the system relies on Bob, a free signal processing and machine learning toolbox developed at Idiap. The NIST official results show the effectiveness of the proposed approach, especially on added noise data.**

I. Introduction

This paper is dedicated to the participation of the Biometric group at Idiap Research Institute in the 2012 NIST Speaker Recognition Evaluation. The proposed system is based on the Inter-Session Variability modelling that is an easy solution to handle channel variations. Figure 1 briefly illustrates the three main components of our system: the feature extraction, the modelling, and the score normalization and calibration. Section III presents the feature extraction module. Section IV details the inter-session variability modelling technique. Score normalization and calibration are presented in Section V. The organization of the development set used by the I4U coalition is detailed in section VI. Section VII shows the results obtained on the development set and on SRE'12 set. Section VIII briefly describes the Bob Toolbox used in our implementation. Section IX concludes the paper.

II. Feature Extraction and Preprocessing

The audio segments are first denoised using Qualcomm-ICSI-OGI front end [1]. Then, 19 Mel Frequency Cepstrum Coefficient (MFCC) and log energy, together with their first (Δ) and second derivatives ($\Delta\Delta$) are computed every 10ms. Thus, each acoustic vector is composed of 60 features. Coefficients are obtained by computing 24 filter bank coefficients over 20ms Hamming windowed frames every 10ms.

The log energy values of each audio segment are normalized and then used to train a 2 Gaussians Classifier where the Gaussian distribution with the higher mean value corresponds to the speech part. In this way, non-speech part is discarded. Finally, a feature normalization step based on Cepstral Mean and Variance Normalization (CMVN) is applied on the remaining speech.

This work was supported by the Swiss National Science Foundation under the LOBI project, contract no. SNSF-235.

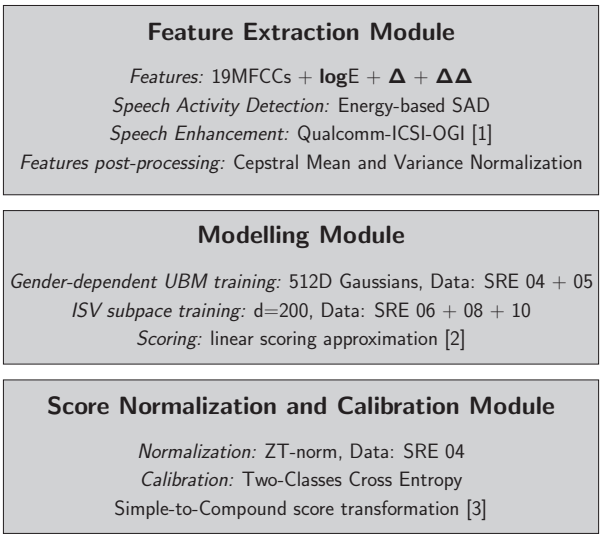

Fig. 1. Illustration of the different components of the proposed system

III. Session Variability Modelling

A. World Model Training

Universal Background Models (UBM) in the proposed system are gender dependent. Both male and female UBM models are mixtures of 512 Gaussian components (GMM). They are trained on NIST SRE'04 and SRE'05 using classical EM-ML algorithm (25 iterations) after a first step of k-means clustering (25 iterations). Covariance matrices are diagonal. Practically, the implementation of the training process is parallelized in order to cope with the huge amount of data needed. It is worth noting that the amount of time needed to train the 2 UBM models is roughly equal to 1200 hours (50 days) on a single core CPU.

B. Inter-Session Variability (ISV)

A popular way to enroll client models is achieved by using mean-only relevance MAP (maximum *a posteriori*) adaptation [2]. In terms of GMM mean supervectors, client enrollment can then be expressed as

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i,$$

where $\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i$, $\boldsymbol{D}$ is diagonal and $\boldsymbol{z}_i$ is a latent variable assumed to be normally distributed $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Ideally, the resulting client model should be robust to any variations within the client's enrollment data. However, with

the previous formulation, there is no explicit modelling of session variability. ISV [6] aims to estimate and exclude the effects of within-client variation, in order to create more reliable client models. The main assumption is that the particular conditions of an audio segment $\boldsymbol{O}_{i,j}$ result in an offset, $\boldsymbol{u}_{i,j}$, to the target's GMM mean supervector

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{d}_i + \boldsymbol{u}_{i,j}.$$

In addition, the offset is assumed to lie in a low-dimensional subspace, such that

$$\boldsymbol{u}_{i,j} = \boldsymbol{U}\boldsymbol{x}_{i,j}.$$

JFA [7] can be seen as an extension of ISV, the difference being in the client-dependent offset.

Inter-session variability modelling (ISV) [6]: $\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i$

Joint factor analysis (JFA) [7]: $\boldsymbol{d}_i = \boldsymbol{V}\boldsymbol{y}_i + \hat{\boldsymbol{D}}\boldsymbol{z}_i$

Finally, a description of the three stages involved when using these session variability modelling techniques is provided below.

Enrollment: Find the true session-independent target model, $\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i$, by jointly estimating $\{\boldsymbol{d}_i, \boldsymbol{u}_{i,1}, \boldsymbol{u}_{i,2}, ..., \boldsymbol{u}_{i,J}\}$.

Scoring: For a test segment $\boldsymbol{O}_t$, estimate the offset $\boldsymbol{u}_t$, then score as usual:

$$h\left(\boldsymbol{O}_t, \boldsymbol{s}_i\right) = \frac{1}{K}\sum_{k=1}^{K}\left[\log\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{s}_i + \boldsymbol{u}_t\right)\right) - \log\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{m} + \boldsymbol{u}_t\right)\right)\right]$$

In practice, this log-likelihood ratio is estimated using a first order approximation known as linear scoring [3]. In addition, the previous equation assumes that the session offsets from both the client model and the UBM are identical (this is known as the *LPT assumption* in [3]).

Training: Find the maximum likelihood subspaces on background data ($\boldsymbol{U}$ for ISV, plus $\boldsymbol{V}$ and $\hat{\boldsymbol{D}}$ for JFA); Estimate the latent variables $\boldsymbol{x}_{i,j}$, $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ using a standard normal prior and MAP estimation.

The gender dependent subspaces U (d=200) of the ISV models were trained on a collection of data from NIST 2006-2008-2010 SRE.

IV. Score Normalization and Calibration

In the previous section, we presented how the ISV scores are computed using the linear scoring approximation. Afterwards, score normalization based on ZT-norm [8] is applied. The cohort set is composed of 121 male speakers and 183 female speakers selected from NIST SRE'04. It is used for both T-norm and Z-norm.

The calibration is done using the BOSARIS toolbox [4], and is composed of two steps: 1) the first step is based on a two-classes cross entropy calibration; 2) the second step is a transformation of the simple likelihood ratios into compound likelihood ratios. This last step is specific to SRE'12. It takes into account *a piori* probabilities about known and unknown targets and non-targets as detailed by NIST SRE'12 Evaluation plan[1].

V. Development Set

The development set results from a joint effort within the I4U consortium [9], [10]. Its design is based on the SRE'06, SRE'08 and SRE'10. It consists of two sets: DEV and EVAL. It is worth noting that the EVAL set was designed in a way that ensures the same target speakers as in the SRE'12 set, and that DEV-Test and EVAL-Test are completely disjoint.

Additive noise. Two noisy versions (SNR-levels 6dB and 15dB) of each audio segment were included by randomly adding one of 10 noise segments: 9 HVAC (heating, ventilation, and air-conditioning) and 1 crowd noise. The procedure of adding noise was done using FaNT [11]. More details about this set can be found in [9], [10].

VI. Results

Table I details the performance obtained on I4U development set for both male and female speakers. It is good to note that those results do not consider the last simple-to-compound transformation. The results show an overall EER of less than 2%, and a minimum Cprimary of less than 0.2.

TABLE I
System performance based on Equal Error Rate (EER) and minimum Cprimary (minCp) on I4U Development set for both male and female speakers.

DEV				**EVAL**			
male		**female**		**male**		**female**	
EER	minCp	EER	minCp	EER	minCp	EER	minCp
1.74%	0.1637	2.03%	0.1904	1.19%	0.1755	1.39%	0.1655

Table II shows the official numbers provided by NIST on five different conditions: *det1* for clean-interview condition, *det2* for clean-telephone condition, *det3* for added- noise-interview condition, *det4* for added-noise-telephone condition, and *det5* for real-noise-telephone condition. The official ranking provided by NIST shows that our results are very competitive especially on added-noise conditions (det3, det4). The results depicted in Table II suggest that the SRE'12 dataset is more challenging than the development set.

TABLE II
System performance based on minimum and actual Cprimary on I4U DEV and EVAL sets, and on NIST SRE12 sub-conditions.

	DEV	**EVAL**	**NIST SRE12**				
			det1	*det2*	***det3***	***det4***	*det5*
minCp	0.181	0.168	0.410	0.422	**0.353**	**0.285**	0.534
actCp	0.186	0.312	0.536	0.464	1.125	0.311	0.638

[1]http://www.nist.gov/itl/iad/mig/sre12.cfm.

VII. Bob Toolbox

The proposed system is implemented using Bob[2] [12] toolbox. Bob is a free signal processing and machine learning toolbox originally developed by the Biometric group at Idiap Research Institute. It gathers several speech and image processing routines (*e.g.* MFCC, LFCC, LBP, SIFT, *etc.*), together with a bunch of machine learning algorithms (*e.g.* PCA, LDA, MLP, SVM, ISV, JFA, GMM, *etc.*). It also supports protocols for several audio and image biometric databases (*e.g.* MOBIO, BANCA, NIST SRE'12, at&t *etc.*). The toolbox provides instructions for organizing and distributing extensions as satellite packages. It is worth noting that the source code for the proposed system will be available very soon as one of Bob's satellite packages.

VIII. Conclusions

This paper describes the Idiap speaker recognition system for the 2012 NIST SRE. The system is based on inter-session variability modelling. Official ranking show that the proposed system is a competitive system especially for added-noise conditions. Future work will be carried out on other state-of-the-art techniques such as Joint Factor Analysis (JFA), and total variability modelling (i-vectors). We will also evaluate the local binary features proposed by [13]. Another future direction is the improvement of the speech activity detection (SAD) technique by exploring the 4Hz modulation energy. Further, future work will evaluate recent methods on mobile environment[3] using MOBIO[4] database [14].

References

[1] http://www1.icsi.berkeley.edu/Speech/papers/qio/

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 2000.

[3] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of scoring methods used in speaker recognition with joint factor analysis. *ICASSP*, 2009.

[4] N. Brümmer and E. de Villiers, The Bosaris Toolkit. Available: https://sites.google.com/site/bosaristoolkit/.

[5] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Intersession variability modelling and joint factor analysis for face authentication. *International Joint Conference on Biometrics*, 2011.

[6] R. Vogt and S. Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 2008.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. Digital Signal Process., 2000.

[9] R. Saeidi et *al.*. I4U Submission to NIST SRE 2012: A Large-Scale Collaborative Effort for Noise-Robust Speaker Verification. *submitted to ICASSP 2013*, 2012.

[10] k. A. Lee et *al.*.The I4U Submission to the 2012 NIST Speaker Recognition Evaluation

[11] http://dnt.kr.hsnr.de/download.html

[12] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. *ACM Multimedia*, 2012.

[13] A. Roy, M. Magimai.-Doss and S. Marcel. A Fast Parts-based Approach to Speaker Verification using Boosted Slice Classifiers, *IEEE Transactions on Information Forensics and Security*, 2011.

[14] E. Khoury, M. Günther, and S. Marcel. ICB 2013 - Competition on speaker recognition in mobile environment using the MOBIO database: the evaluation plan. 2012.

[2] http://www.idiap.ch/software/bob

[3] http://www.beat-eu.org/evaluations/icb-2013-speaker-recognition-mobio

[4] http://www.idiap.ch/dataset/mobio