The 2013 Speaker Recognition Evaluation in Mobile Environment

E. Khoury*,[1] B. Vesnicer,[2] J. Franco-Pedroso,[3] R. Violato,[4] Z. Boulkenafet,[5] L.M. Mazaira Fernández,[6] M. Diez,[7]
J. Kosmala,[8] H. Khemiri,[9] T. Cipr,[10] R. Saeidi,[11] M. Günther,[1] J. Žganec-Gros,[2] R. Zazo Candil,[3] F. Simões,[4] M. Bengherabi,[5]
A. Álvarez Marquina,[6] M. Penagarikano,[7] A. Abad,[8] M. Boulayemen,[9] P. Schwarz,[10,12] D. Van Leeuwen,[11] J. González-Domínguez,[3]
M. Uliani Neto,[4] E. Boutellaa,[5] P. Gómez Vilda,[6] A. Varona,[7] D. Petrovska-Delacrétaz,[9] P. Matějka,[10,12] J. González-Rodríguez,[3]
T. Pereira,[4] F. Harizi,[5] L. J. Rodriguez-Fuentes,[7] L. El Shafey,[1] M. Angeloni,[4] G. Bordel,[7] G. Chollet,[9] S. Marcel[1]

[1]Idiap Research Institute (CH), [2]Alpineon Ltd. (SLO), [3]Universidad Autónoma de Madrid (ES),
[4]CPqD (BR), [5]Centre de Développement des Technologies Avancées (DZ), [6]Universidad Politécnica de Madrid (ES),
[7]University of the Basque Country (ES), [8]L2F/INESC-ID (PT), [9]Institut Mines-Télécom (FR), [10]Phonexia s.r.o. (CZ),
[11]Radboud University Nijmegen (NL), [12]Brno University of Technology (CZ)

Abstract

This paper evaluates the performance of the twelve primary systems submitted to the evaluation on speaker verification in the context of a mobile environment using the MOBIO database. The mobile environment provides a challenging and realistic test-bed for current state-of-the-art speaker verification techniques. Results in terms of equal error rate (EER), half total error rate (HTER) and detection error trade-off (DET) confirm that the best performing systems are based on total variability modeling, and are the fusion of several sub-systems. Nevertheless, the good old UBM-GMM based systems are still competitive. The results also show that the use of additional data for training as well as gender-dependent features can be helpful.

1. Introduction

Automatic speaker verification is the use of a machine to verify a person's claimed identity from his voice. This topic is investigated since 1970^{th} [1], and is regularly evaluated by the National Institute of Standards and Technology (NIST)[1] since 1996. The NIST speaker recognition evaluation (SRE) series aims to contribute to the direction of research efforts of *text independent* speaker recognition. During this series, many techniques have been proposed. One common thread with current successful techniques is their ability to cope with inter-session variability that can come from acoustic environments, communication channels, languages, and states of the speaker (stress, etc.).

*Corresponding author: elie.khoury@idiap.ch

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

Following the same spirit as the NIST SRE, the Biometric Group at the Idiap Research Institute organized the evaluation on *text independent* speaker recognition. This evaluation is the second in an ongoing series of speaker and face recognition evaluations conducted in a mobile environment. It is carried out on the MOBIO database [2], which contains videos of talking faces that were captured with mobile devices. The MOBIO database provides a unique opportunity to analyze two mature biometrics (speaker and face) side by side in a mobile environment. The average speech duration of MOBIO segments is around 8 s, far lower than the one of NIST SRE (around 70 s in SRE 2012).

In total, 12 institutions participated in the speaker verification evaluation, and provided 21 valid submissions (12 primary and 9 secondary submissions). To assure a fair evaluation and comparable results, all participants of the evaluation had to strictly follow an unbiased evaluation protocol. In the first phase of the evaluation, the *training* and the *development* set was made available to the participants. In these sets, each audio file was labeled with the *client ID*, and the participants were allowed to optimize their system parameters according to these data. In the second phase, an *evaluation* set was given to the participants. In the evaluation set, the file names were anonymized, so that client ID could not be read out from them. Participants were asked to compute the scores according to the protocols and send the final score files to the organizers to evaluate them.

The structure of the paper is as follows: Section 2 introduces the MOBIO database. Section 3 describes the employed evaluation metrics. Section 4 presents the participants and their systems. Section 5 evaluates the verification performances of the systems and discusses some further characteristics of them. Section 6 concludes the paper.

2. The MOBIO Database

The MOBIO database is a bi-modal (face/speaker) database recorded from 152 people. The database has a female-male ratio of nearly 1:2 (100 males and 52 females) and was collected from August 2008 to July 2010 in six different sites from five different countries. In total 12 sessions were captured for each individual.

The database was recorded using two types of mobile devices: mobile phones (NOKIA N93i) and laptop computers (standard 2008 MacBook). In this evaluation we will only use the mobile phone data with a sampling rate of 16kHz. The MOBIO database is a challenging database since the data is acquired on Mobile devices possibly with real noise, and the speech segments can be very short (less than $2sec$). More technical details about the MOBIO database can be found in [2] and on its official web page[2], which also contains instructions on how to obtain the data.

Based on the gender of the clients, two different evaluation protocols *male* and *female* were generated. In order to have an unbiased evaluation, the clients of the database are split up into three different sets: training, development and evaluation set.

Training set. The data of this set are used to learn the background parameters of the algorithm (UBM, subspaces, *etc.*). They can also be used for score normalization (cohort, *etc.*). It is worth noting that participants can use external data in their background training, however they should explicitly precise it in their system description.

Development set. The data of this set are used to tune meta-parameters of the algorithm (*e. g.* number of Gaussians, dimension of the subspaces, *etc.*). For the enrollment of a client model, 5 audio files of the client are provided, and **it is forbidden to use the information of other clients of the development set**. The remaining audio files of the clients serve as probe files, and likelihood scores have to be computed between all probe files and all client models. In systems that require score calibration these scores can be used to train the calibration parameters.

Evaluation set. The data of this set are used for computing the final evaluation performance. It has a structure similar to the development set. The only difference is that the file names are anonymized in order to prevent participants to optimize their system on the evaluation set.

Table 1 statistically details each of the sets described above. It specifies the number of files, the number of targets, and the number of trials.

[2] https://www.idiap.ch/dataset/mobio

3. Evaluation Method

To evaluate the speaker verification performance, the used metrics are based on the *false acceptance rate* (FAR) and the *false rejection rate* (FRR). The definition of these rates is dependent on a certain *threshold* θ:

$$\begin{aligned} \mathrm{FAR}(\theta) &= \frac{|\{s_{\mathrm{imp}} \mid s_{\mathrm{imp}} \geq \theta\}|}{|\{s_{\mathrm{imp}}\}|} \\ \mathrm{FRR}(\theta) &= \frac{|\{s_{\mathrm{cli}} \mid s_{\mathrm{cli}} < \theta\}|}{|\{s_{\mathrm{cli}}\}|} \end{aligned} \tag{1}$$

where s_{cli} are client scores, while s_{imp} are imposter scores. We evaluate the FAR and the FRR for both the development and the evaluation set independently.

The first evaluation metric we will use is based on the *equal error rate* (EER) and the *half total error rate* (HTER). Particularly, it defines a score threshold θ_{dev}, based on the EER of the development set, and compute the HTER[3] using this threshold:

$$\begin{aligned} \theta_{\mathrm{dev}} &= \arg\min_{\theta} |\mathrm{FAR}_{\mathrm{dev}}(\theta) - \mathrm{FRR}_{\mathrm{dev}}(\theta)| \\ \mathrm{EER} &= \frac{\mathrm{FAR}_{\mathrm{dev}}(\theta_{\mathrm{dev}}) + \mathrm{FRR}_{\mathrm{dev}}(\theta_{\mathrm{dev}})}{2} \\ \mathrm{HTER} &= \frac{\mathrm{FAR}_{\mathrm{eval}}(\theta_{\mathrm{dev}}) + \mathrm{FRR}_{\mathrm{eval}}(\theta_{\mathrm{dev}})}{2} \end{aligned} \tag{2}$$

The second metric is the *detection error trade-off* (DET). In this metric, the FRR is plotted against the FAR. The DET curves serve to evaluate the calibration of the verification systems.

4. Participants

12 sites have fulfilled the protocol rules of the evaluation. Their names and their identifiers are reported in Table 2. In this section, we briefly review the techniques used by the participants.

Table 3 summarizes the feature extraction setup of the systems. It shows the different techniques for feature extraction (MFCC, LFCC, etc.), voice activity detection (energy-based, phoneme-based, etc.), speech enhancement (spectral subtraction, Wiener filtering, etc.), and feature post-processing (feature warping, cepstral mean subtractions, etc.).

Table 4 summarizes the classifier approaches used in the submitted systems. This table shows the techniques used for: modeling (*Total variability modeling*, *Gaussian mixture models*, etc.), post-processing (LDA, WCCN, etc.), scoring (*log likelihood ratio*, *linear scoring*, etc.), and score normalization and calibration (t-norm, zt-norm, etc.).

[3] In speaker recognition terminology, this corresponds to C_{DET} with equal priors and costs. However, HTER is adopted in this evaluation in order to use the same terminology as for face recognition.

Table 1. PARTITIONING OF THE MOBIO DATABASE. *This table details the number of targets and audio files of the training set, as well as the number of targets and enrollment audio files, and the number of test segments and trials for the development and the evaluation set.*

	Background		Development					Evaluation				
			Enrollment		Test			Enrollment		Test		
	Speakers	*Files*	*Targets*	*Files*	*Speakers*	*Files*	*Trials*	*Targets*	*Files*	*Speakers*	*Files*	*Trials*
MALE	37	7104	24	120	24	2520	60480	38	190	38	3990	151620
FEMALE	13	2496	18	90	18	1890	34020	20	100	20	2100	42000

Table 2. PARTICIPANTS. *The institutions and the identifiers of their submitted primary system (by alphabetic order)*

Institution	System Identifier
Alpineon Ltd., Slovenia	Alpineon
ATVS Biometric Recognition Group - Universidad Autónoma de Madrid, Spain	ATVS
Centre de Développement des Technologies Avancées, Algeria	CDTA
CPqD, Brazil	CPqD
GIAPSI, Universidad Politécnica de Madrid, Spain	GIAPSI
GTTS - University of the Basque Country (UPV/EHU), Spain	EHU
Idiap Research Institute, Switzerland	IDIAP
L2F/INESC-ID, Portugal	L2F
Joint submission of L2F/INESC-ID and UPV/EHU	L2F-EHU
Institut Mines-Télécom (Télécom ParisTech-Télécom SudParis), France	Mines-Telecom
Phonexia s.r.o., Czech Republic	Phonexia
Radboud University Nijmegen, The Netherlands	RUN

Alpineon. The Alpineon KC OpComm system is the fusion of 9 different *total variability modeling* [11] (also known as i-Vector) based sub-systems. All sub-systems are identical, but use different acoustic features. 3 different cepstral-based features (MFCC, LFCC, PLP) are extracted over 3 different frequency regions (0-8000 Hz, 0-4000 Hz and 300-3400 Hz). The 9 sub-systems are exclusively trained on the MOBIO training set. Since the training dataset is relatively small, a gender-independent training (with 512 Gaussians) is done, and the dimensionality of the eigen-voices is relatively low (dim=49).

ATVS. The ATVS system consists of a standard i-Vector configuration with PLDA modeling [12]. The gender-dependent UBMs with 2048 components are trained using data from Switchboard[6] (SWB-I, SWB-II phase 2 and 3), and Mixer (from NIST SRE 04, 05, 06, 08, 10) with around 1300 speakers per gender. PLDA is trained with a subset of the same data in addition to MOBIO training dataset (in total, around 600 speakers per gender). i-Vector length normalization is applied to palliate duration variability of utterances and is trained on the same subset as PLDA. At the scoring level, symmetric score normalization (s-norm) is applied using gender-dependent cohorts that are extracted from the MOBIO database with a total number of 300 speakers each.

[6] `http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html`

CDTA. The CDTA speaker recognition system is also based on an i-Vector framework [11]. The main novelty of the system is the use of the *conformal embedding analysis* (CEA) [9]. CEA uses the cosine similarity distance and local graph modeling to map data into a low dimensional representation with a higher discrimination between classes. Unlike the *linear discriminant analysis* (LDA), CEA has no assumption about the distribution of data, and the use of the cosine distance gives more robustness to channel variation. Gender-dependent UBM models consisting of 128 Gaussians are trained using only the MOBIO training database.

CPqD. The CPqD system is the fusion of two sub-systems. The first sub-system is based on the UBM-GMM modeling, while the second is based on the i-Vector modeling. In the UBM-GMM approach, a gender-dependent UBM model with 512 components is adapted for each target model using the *maximum a posteriori* (MAP) algorithm. The relevance factor is tuned on the development set and is found to be relatively high (r=512). In the i-Vector approach, a gender-independent UBM model with 256 components is used, and simple cosine distance is directly computed on the extracted i-vectors (dim=400). The fusion of the two sub-systems is done at the score level using the *linear logistic regression* implemented in the Bob toolbox.

GIAPSI. The GIAPSI system consists of a UBM-GMM configuration followed by a scoring based on log likelihood

Table 3. FEATURE EXTRACTION SETUP FOR THE DIFFERENT SYSTEMS. *ZCR: zero cross rate, ME-4Hz: modulation of the energy near 4Hz, CMVN: cepstral mean and variance normalization, CMS: Cepstral Mean Subtraction.*

System	Sampling rate	Features	Voice activity detection	Speech enhancement	Features post-processing
Alpineon	16kHz	MFCC, LFCC, PLP	Energy-based [3]	-	CMVN
ATVS	16kHz	MFCC	Energy-based + SOX[4]	Qualcomm-ICSI-OGI[5]	CMN + RASTA + Feature warping
CDTA	16kHz	MFCC	Energy-based + ZCR [4]	-	-
CPqD	16kHz	MFCC	ITU-T G.729b	-	CMVN + Feature warping
GIAPSI	16kHz	Gender-dependent MFCC, F0, F3	Energy-based	Channel noise removal	CMS + RASTA + Feature warping
EHU	8kHz	MFCC	Energy-based	-	RASTA + Feature warping
IDIAP	16kHz	MFCC	Energy-based + ME-4Hz	Qualcomm-ICSI-OGI	CMVN
L2F	16kHz	MFCC	Energy-based	-	CMVN + Feature warping
Mines-Telecom	16kHz	MFCC	Energy-based [5]	-	Feature warping
Phonexia	8kHz	MFCC	{Energy + F0 + phoneme}-based	-	Feature warping
RUN	8kHz	MFCC	Energy-based [6]	Wiener filtering [7, 8]	Feature warping

Table 4. MODELING AND SCORING TECHNIQUES USED IN THE DIFFERENT SYSTEMS. *PLDA: probabilistic linear discriminant analysis, NAP: nuisance attribute projection, SVM: super vector machine, LDA: linear discriminant analysis, CEA: conformal embedding analysis, WCCN: within class covariance normalization, LNorm: Length normalization.*

System	Modeling technique	Post-processing	Scoring technique	Score normalization and calibration
Alpineon	Total variability modeling (i-Vector)	LDA + WCCN + LNorm	PLDA	linear logistic regression score fusion
ATVS	i-Vector	LNorm	PLDA	s-norm
CDTA	i-Vector	CEA [9] + WCCN	cosine distance	-
CPqD (sub-I)	Guassian Mixture Modeling (UBM-GMM)	-	log likelihood ratio	-
CPqD (sub-II)	i-Vector	-	cosine distance	-
CPqD (Fusion)	-	-	-	linear logistic regression score fusion
EHU	i-Vector	-	PLDA	linear logistic regression score calibration
GIAPSI	UBM-GMM	-	log likelihood ratio	-
IDIAP	Intersession variability modeling (ISV)	-	linear scoring [10]	zt-norm
L2F (sub-G)	UBM-GMM	-	log likelihood ratio	-
L2F (sub-S)	Gaussian supervector (GSV)	NAP	SVM	z-norm
L2F (sub-I)	i-Vector	WCCN	cosine distance	t-norm
L2F (Fusion)	-	-	-	linear logistic regression score fusion
L2F-EHU	-	-	-	linear logistic regression score fusion
Mines-Telecom	UBM-GMM	-	log likelihood ratio	-
Phonexia	i-Vector	LDA + WCCN + LNorm	PLDA	length dependent linear logistic regression
RUN	i-Vector	LDA + WCCN	PLDA	linear calibration

ratio. The particularity of the system is the use of gender-dependent features. The normal MFCC parameters (extracted from the power spectral density) are augmented with the MFCC coefficients extracted from the glottal and vocal-tract estimates [13] separation algorithm over the speech signal. Thus, the female features vector is composed of: 24 MFCC + their first derivatives Δ (34 mel-spaced filter bank), 2 MFCC extracted from the glottal estimate, 2 MFCC + Δ extracted from the vocal tract estimate, F0 estimate, and F3 estimate. The male features vector is composed of: 28 MFCC + Δ (38 mel-spaced filter bank), 8 MFCC extracted from the glottal estimate, $\Delta Energy$, F0 estimate, and F3 estimate. The gender-dependent UBMs are trained exclusively using the MOBIO training dataset.

EHU. The EHU system is also based on the *total variability* modeling [11]. The feature extraction and the voice activity detection (VAD) are done using the SAUTRELA toolkit[7] [14]. PLDA [12] is applied directly on the extracted 500 dimensional i-Vector space. Gender independent 1024 component UBM and i-Vector extractor, and gender dependent PLDA systems are trained on the background set of the MOBIO database. The development data set is used only for the calibration estimation.

IDIAP. The IDIAP system is based on a single classification framework employing the *inter-session variability* (ISV) modeling technique [15]. This system was used for NIST SRE 2012 [16]. The implementation of the system relies on Bob[8] [17], an open source signal-processing and machine learning toolbox originally developed by the Biometrics Group at Idiap. ISV belongs to the same family of *inter-session variability modeling* techniques as *Joint Factor Analysis* (JFA). The only difference is that eigen-voice and eigen-channel spaces are merged. ISV scores are computed using *linear scoring* approximation [10]. Finally, scores are normalized using zt-norm. The cohort speakers are chosen from the MOBIO training dataset.

[7] http://gtts.ehu.es/TWiki/bin/view/Sautrela

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

L2F. The L2F (INESC-ID) primary system results from the fusion of three sub-systems. The first sub-system consists of a standard UBM-GMM configuration (1024 Gaussians, MAP adaptation, relevance factor=16) with gender-dependent UBMs. The second sub-system is a Gaussian super-vector (GSV) based system with a gender-dependent UBM (256 Gaussians) that are adapted to the clients using MAP adaptation. Gaussian super-vectors are composed by concatenating the Gaussian means. Channel normalization is accomplished using *nuisance attribute projection* (NAP) [18]. Those super-vectors are used to train a *support vector machine* (SVM) using *Kullback-Leibler* (KL) linear distance. Scores are computed as the distance to the hyperplane. The third sub-system uses the i-Vector based approach with gender-dependent UBMs (256 Gaussians) and 400 dimensional total variability (TV) matrix. One i-Vector is computed per speaker using all the enrollment data. Scoring uses a simple cosine distance. Prior to scoring, i-Vectors are normalized using the WCCN technique. The training of the various UBM models is done exclusively using MOBIO training dataset. Finally, the fusion of the three sub-systems uses *linear logistic regression* implemented in BOSARIS[9].

L2F-EHU. The joint submission of the EHU and L2F participants is based on the fusion of their two primary systems that are described above. The fusion is done using *linear logistic regression*.

Mines-Telecom. The Mines-Telecom primary system uses the UBM-GMM approach. The system is based on the reproducible BioSecure Speaker baseline system[10] described in [5]. The initial configuration of this system is as follows: 16 LFCC coefficients + deltas + delta energy, 300-3400Hz frequency range, energy-based voice activity detection, and models with 512 Gaussians. The gender-dependent UBM models are trained on NIST SRE 03-04. The score normalization uses t-norm.

This baseline system was adapted to the MOBIO database. First, the feature vector is composed of 20 MFCC coefficients (32 Mel filter bank) together with their first derivatives and the delta energy. This is intended to better exploit the 16KHz range. Second, *Feature warping* is added to the original tool chain. In contrast, a score normalization step is not applied. Another particularity of the primary system is the use of the MOBIO training dataset and the Voxforge[11] dataset.

In addition, two secondary systems were submitted. The first system exclusively uses the MOBIO training dataset, the second system uses both NIST SRE (03-04) and MOBIO dataset to train an UBM model with 1024 Gaussians (in this case, the original signals were down sampled to 8kHz) and 16 MFCC coefficients (24 Mel filter bank). For both of them, the performances are lower but close to the primary system.

Phonexia. The Phonexia system is based on the i-Vector framework. A production system was used and the adaptation took less than a day. First, the speech detection consists of three steps: 1) an energy-based VAD with a fixed threshold, 2) a VAD based on F0 detection and smoothing, and 3) a VAD based on neural networks (phoneme posterior estimation + Viterbi realignment). Second, the gender-independent UBM model is trained on several telephone datasets (LDC, NIST, and internal data). The training of the i-Vectors uses a low dimensional (dim=400) matrix that defines both the speaker and channel subspaces. Then LDA is applied to reduce the voice-print size to 250 dimensions. The post-processing of i-Vectors include mean normalization, WCCN, and normalization to unit vector length. The scoring is done using PLDA[12] [12], which compares the voice-prints with full rank matrices for both within-speaker and across-speaker variability. Finally, a length-dependent score calibration using piece-wise linear logistic regression is applied using a cohort external to the MOBIO database.

RUN. The RUN system consists of a standard i-Vector configuration with PLDA modeling. This system was developed for the NIST SRE 2012 evaluation [19]. Speech enhancement is applied for both speech activity detection and feature extraction. For noise estimation, the *improved minima controlled recursive averaging* (IMCRA) approach [8] is used. Gender-dependent UBM with 2048 components is trained using NIST-SRE 2004-2006, Switchboard, and Fisher dataset [20]. i-Vectors are trained using a low dimensional (dim=400) TV matrix. LDA projection is applied in order to reduce the i-Vectors dimension to 200. Prior to the PLDA modeling, the i-Vectors are processed by i-Vector centering, WCCN, and length normalization.

5. Performance Results

The score files sent by the participants are evaluated using the two different verification metrics described in section 3. Table 5 shows the equal error rates on the development set and the half total error rates on the evaluation set for both genders. Clearly, the error rates on *Female* are higher than on *Male*. This might be caused by the fact that the database contains more men than women.

[9] https://sites.google.com/site/bosaristoolkit/

[10] svnext.it-sudparis.eu/svnview2-eph/ref_syst/Speech_Alize/doc/howTo.pdf

[11] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit

[12] The voiceprint comparison module is adapted to the MOBIO dataset using a ready-made tool offered to Phonexia's clients. Phonexia offers SID systems and these tools to any research/education institution for free.

In table 5, the fusion systems are marked with *, and the ones that use external training data are marked with $^+$. The best overall rates are highlighted in bold font, while the best results of the simple systems are in italics. Among the fusion systems, clearly Alpineon gets the best scores, but, unlike other systems, the performance differs between development and evaluation set. This might be due to an over-tuning of the parameters on the development set.

L2F-EHU system is a good example that shows that the fusion of different systems can improve the results: In all cases, the fusion of the L2F and EHU systems is better than their simple systems[13].

Among the simple systems, Phonexia seems to be well tuned on the development (best simple system on the development set for both Female and Male). On the evaluation set, Mines-Telecom obtains the best simple system performance on Female. Obviously, the use of additional suitable data (Voxforge database) for training the UBM is helpful. The GIAPSI system gets the best simple system performance on Male. This is probably due to the use of gender-dependent features (see section 4). Unlike the Mine-Telecom system, the use of non-suitable external data such as NIST SRE data without any adaptation to the MOBIO database decreases the performance[14]. This is the case of RUN [15] that got good results at NIST SRE 2012.

Fig. 1 shows the DET curves of the different systems on the development and evaluation sets for both Female and Male. Those plots confirm the same conclusions as above. They also show that RUN, although the high error rates, is a well calibrated system (curves with an angle close to $45°$).

Fusion of all primary systems. An additional experiment that combines the 12 primary systems[16] using *linear logistic regression* [21] is made, and Table 6 shows the performance of the fusion system obtained on both DEV and EVAL sets. By comparing the results with the best system, all the error rates dropped significantly (more than 30% in the worst case). This implies that there is still a good margin for improving the performance of simple systems.

Comparison with the previous evaluation. In order to measure the progress of the systems in the last couple of years, we compare the performance between the best system of the first MOBIO evaluation [22], the best system of the current evaluation, and the fusion of all submitted systems[17]. DET curves in Fig. 2 clearly show a gain that might be due to the success of the TV modeling, together with the fusion techniques. Indeed, the average HTER of the best system in [22], Alpineon system, and the fusion of all submitted systems are 10.59%, 8.77% and 5.88%, respectively.

System requirements. One important point, especially in mobile environments, is the requirements of the speaker verification system in terms of speed and memory. Usually, the requirements can be split into an offline training and enrollment phase, and an online verification phase. Practically, it is difficult to compare the computation cost of the different systems because of the various tools (see Table 7). Obviously, UBM-GMM approaches need less computational power than i-Vector approaches since the latter needs to train additionally a TV matrix [23]. However, this is not very problematic since it is done offline. Furthermore, the processing time and the memory requirements of a system that relies on the fusion of several sub-systems make difficult its integration in mobile devices.

Table 5. PERFORMANCE SUMMARY I. *Equal error rate (EER %) on the development (DEV) set and half total error rate (HTER %) on the evaluation (EVAL) set. The best system and best single systems are in bold and bold italic, respectively.*

System	Female		Male	
	DEV	EVAL	DEV	EVAL
Alpineon*	**7.982**	**10.678**	**5.040**	**7.076**
ATVS$^+$	16.836	17.858	14.881	15.429
CPqD*	14.348	15.987	11.824	10.214
CDTA	19.471	22.640	12.738	19.404
GIAPSI	11.590	12.813	9.683	***8.865***
EHU	17.937	19.511	11.310	10.058
IDIAP	12.011	14.269	9.960	10.032
L2F*	13.484	22.140	10.599	11.129
L2F-EHU*	11.005	17.266	7.889	8.191
Mines-Telecom$^+$	11.429	***11.633***	10.198	9.109
Phonexia$^+$	***8.364***	14.181	***9.601***	10.779
RUN$^+$	25.405	23.112	24.643	22.524

Table 6. PERFORMANCE SUMMARY II. *Results of the fusion of all primary systems using linear logistic regression (EER % on DEV set and HTER % on EVAL set).*

System	Female		Male	
	DEV	EVAL	DEV	EVAL
Fusion	3.556	6.986	2.897	4.767

[13] A bug was found in one of the L2F sub-systems, this affects the L2F primary system and the L2F-EHU system. After fixing the bug, the rows in table 5 corresponding to L2F and L2F-EHU will be: $\{13.484, 14.733, 10.599, 11.051\}$ and $\{11.005, 13.591, 7.889, 8.137\}$

[14] Mines-Telecom has done an additional experiment using NIST SRE (03 and 04) data for UBM training. The EER on the DEV set are 14.80% for Female and 13.62% for Male, respectively.

[15] During the post-evaluation session, RUN has included the duration variability inside the PLDA training. Its new error rates are halved (EER on DEV: 13.39% and 13.73%, HTER on EVAL: 14.09% and 12.12% for Female and Male, respectively.

[16] The scores of all participants on both DEV and EVAL sets will be available on the MOBIO web page.

[17] Although the two protocols are not identical, there is a high overlap between the two datasets, and the difficulty of the data remained the same. This makes the comparison between them significant and fair.

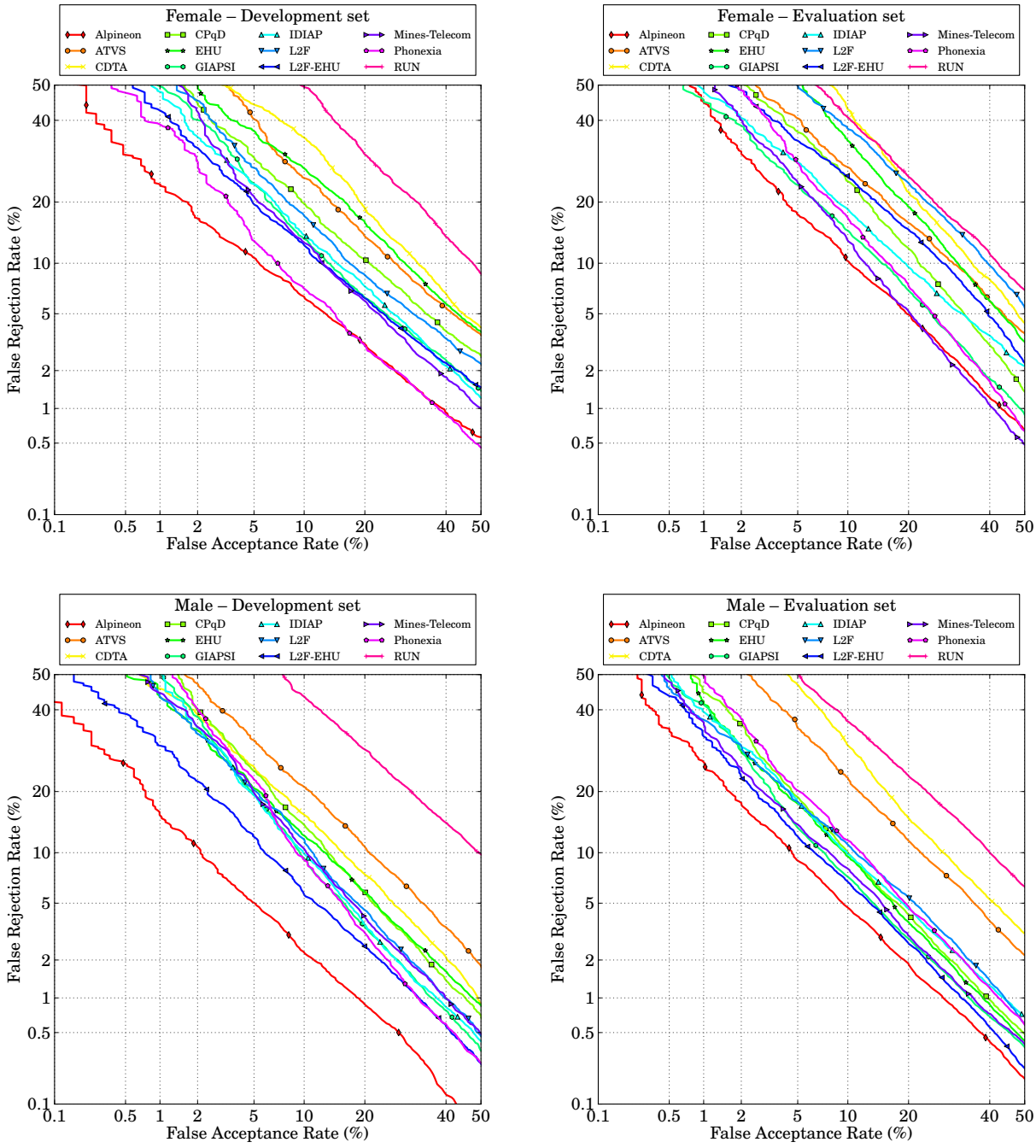

Figure 1. PERFORMANCE SUMMARY III. *DET curves of the different primary systems on the DEV and EVAL sets.*

6. Conclusions

This paper presents the results of the participants to the evaluation on speaker verification in mobile environment. This evaluation produced several interesting findings. First, the use of total variability modeling followed by a score fusion provides the best performances. This explains the boost in performance in comparison to the previous evaluation on MOBIO [22]. Second, the use of external but suitable data to train the background models as well as gender-dependent features can be helpful. Finally, future work will focus on the fusion of the speaker and face [24] modalities towards bi-modal verification system. It will also study the effect of using external training data.

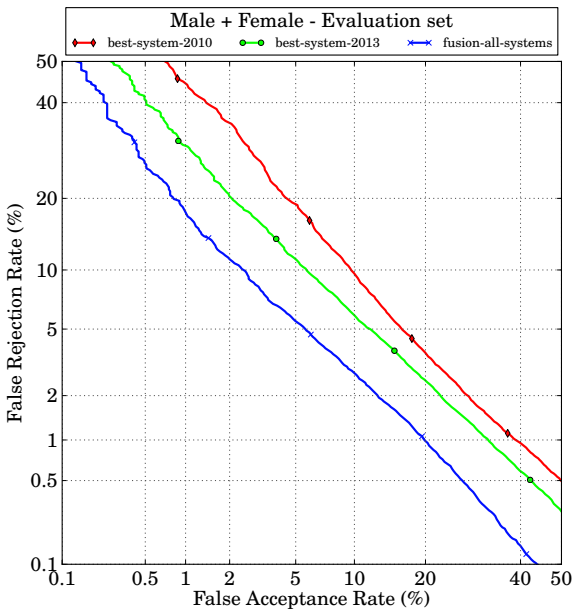

Figure 2. PERFORMANCE SUMMARY IV. *DET curves of the best system in [22], the best system of the current evaluation, and the fusion of all primary systems (for Female+Male).*

Table 7. THE TOOLS USED BY THE PARTICIPANTS. *The mark ($^+$) denotes an open-source software.*

System	Tools
Alpineon	HTK$^+$, Bob$^+$, Matlab, Bosaris$^+$, in-house C and Python tools
ATVS	Qualcomm-ICSI-OGI$^+$, SOX$^+$, in-house tool
CPqD	in-house C tool, in-house MATLAB tool, Bob$^+$, Python$^+$,
CDTA	in-house Matlab tool
GIAPSI	in-house C tool
EHU	SAUTRELA$^+$, Matlab, Bosaris$^+$
IDIAP	Bob$^+$, Qualcomm-ICSI-OGI$^+$
L2F	AUDIMUS (in-house), LibSVM$^+$, in-house matlab tool, Bosaris$^+$
Mines-Telecom	ALIZE-2.0$^+$, SPro4$^+$
Phonexia	BSCORE
RUN	Matlab, Bosaris$^+$

7. Acknowledgment

This evaluation was supported by the European Union under the project BEAT contract no. FP7-284989 , as well as the Swiss National Science Foundation under the LOBI project.

References

[1] J. Campbell. Speaker recognition: A tutorial. In *Proceedings of the IEEE*, 1997.

[2] C. McCool et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, 2012.

[3] M.-W. Mak and H.-B. Yu. Robust voice activity detection for interview speech in NIST speaker recognition evaluation. In *APSIPA ASC*, 2010.

[4] M. Nilsson and M. Ejnarsson. Speech recognition using hidden markov model. *Master Thesis, Department of Telecommunications and Speech Processing, Blekinge Institute of Technology*, 2002.

[5] A. El Hannani, D. Petrovska-Delacrétaz, B. Fauve, A. Mayoue, J. Mason, J.-F. Bonastre, and G. Chollet. Text-independent speaker verification. In *Guide to Biometric Reference Systems and Performance Evaluation*. Springer, 2009.

[6] M. McLaren and D. A. van Leeuwen. A simple and effective speech activity detection algorithm for telephone and microphone speech. In *NIST SRE 2011 workshop*, 2011.

[7] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *ITSAP*, 2003.

[8] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *ITSAP*, 2001.

[9] Y. Fu, M. Liu, and T. Huang. Conformal embedding analysis with local graph modeling on the unit hypersphere. In *CVPR*, 2007.

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

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. 2011.

[12] P. Matějka, O. Glembek, F. Castaldo, M.J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocký. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In *ICASSP*, 2011.

[13] P. Gómez, A. Álvarez, L. M. Mazaira, R. Fernández, V. Nieto, R.Martínez, C. Mu noz, and V. Rodellar. A hybrid parameterization technique for speaker identification. In *EUSIPCO*, 2008.

[14] M. Penagarikano and G. Bordel. Sautrela: a highly modular open source speech recognition framework. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, 2005.

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

[16] E. Khoury, L. El Shafey, and S. Marcel. The idiap speaker recognition evaluation system at NIST SRE 2012. In *NIST Speaker Recognition Conference*, 2012.

[17] 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. In *ACM Multimedia*, 2012.

[18] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for SVM speaker recognition. In *ICASSP*, 2005.

[19] R. Saeidi and D. A. Van Leeuwen. The RUN submission to SRE-2012. In *NIST Speaker Recognition Conference*, 2012.

[20] C. David, D. Miller, and K. Walker. The Fisher corpus: a resource for the next generations of speech-to-text. In *International Conference on Language Resources and Evaluation*, 2004.

[21] N. Brümmer et al. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Speech, Audio and Language Processing*, 2007.

[22] S. Marcel et al. On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation. In *Proceedings of the 20th International conference on Recognizing patterns in signals, speech, images, and videos*, ICPR'10, 2010.

[23] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny. Simplification and optimization of i-vector extraction. In *ICASSP*, 2011.

[24] M. Günther et al. The 2013 face verification evaluation in mobile environment. In *ICB*, 2013.