LIA NIST-SRE'10 systems

Anthony Larcher, Christophe Lévy, Driss Matrouf
Jean-Francois Bonastre

University of Avignon
Laboratoire Informatique d'Avignon - CERI/LIA - France
{anthony.larcher, christophe.levy, driss.matrouf, jean-francois.bonastre}@univ-avignon.fr

1. Introduction

This paper is dedicated to the participation of the LIA laboratory in NIST-SRE'10 evaluation campaign. LIA submitted three system for the core-core task of the evaluation. These systems are combinations of three sub-systems which are also described in this paper. Two of these systems are based on the well known GMM-UBM paradigm as the third one is a GMM-SVM engine All LIA systems include Latent Factor Analysis (LFA) framework [1], [2] for session variability modelling in both training and testing phase. The same feature extraction algorithm is used for all systems. Development was performed on the 2008 evaluation. All LIA systems are mainly based on the open source ALIZE/MISTRAL toolkit [3]. SVM-based systems are using the LIB-SVM library [4].

Section 2 describes the feature extraction and frame selection process. In section 3, the world model training procedure is presented as well as the session variability modelling. The sub-systems combined to design the LIA submitted systems are described in Section 4. Finally, section 5 describes the combination of the sub-systems and gives some experimental results. Some perspectives are given in section 7.

2. Feature extraction

The feature extraction protocol described in this section is used for every LIA sub-system. Parameters extracted from speech signal are classical Linear Frequency Cepstral Coefficients (LFCC). 19 coefficient (c) + energy (e) are used augmented by their first (Δ) and second ($\Delta\Delta$) derivatives. Finally, each feature vector is composed by 50 coefficient ($19c + 19\Delta c + 11\Delta\Delta + \Delta e$). Coefficients are obtained as follows: 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate. Bandwidth is limited to the 300-3400Hz range. Moreover, it is important to notice that no echo-canceller is used on the speech material.

Here, the energy coefficients are first normalised using a mean removal and variance normalisation in order to fit a 0-mean and 1-variance distribution. The energy component is then used to train a three component GMM, which aims to select informative frames. The most energised frames are selected through the GMM: only about 30% of the frames are selected.

Once the speech segments of a signal are selected, a final process is applied in order to refine the speech segmentation:

- overlapped speech segments between both the sides of a conversation are removed,
- morphological rules are applied on speech segments to avoid too short ones.

Finally, feature vectors are normalised to fit a 0-mean and 1-variance distribution. The mean and variance estimators used for the normalisation are computed file by file on all the selected frames.

3. World model, session variability and score normalisation

3.1. World model

Universal background Models (UBM) used in the LIA systems are gender dependent. Three different UBM couples (male and female) are used for the whole submission. UBMs are trained using Fisher English Training Speech Part 1 [5] and NIST-SRE 2004 data collection for telephone data and NIST-SRE 2005 microphone data collection. Table 1 gives the number of components of each UBM and the constitution of the set of data used for UBM training.

UBM	**Micro '05**		**Mix2048**		**Mix512**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Number of distributions per model	512	512	2048	2048	512	512
Number of microphone segments	1262	1480	500	500	500	500
Number of telephone segments	0	0	4110	3225	4110	3225

Table 1: Description of the data used for UBM training. .

UBMs are trained using classical EM-ML algorithm. Covariance matrices are diagonal. For a better separation of initial classes, frames are randomly selected among the entire learning signal via a probability followed by an iteration of the EM algorithm, to estimate the GMM parameters. During the estimation of the world model parameters, instead of using all the learning signals in their temporal order, 10% of frames is selected randomly at each new iteration. For the ten last iterations, the entire signal is classically used in its temporal order. During all the process, a variance flooring is applied so that no variance value is less than 0.5.

3.2. Session variability

A speaker model can be decomposed into three different components: world, a speaker dependent and session dependent components [6], [7], [1]. A GMM mean super-vector is defined as the concatenation of the GMM component means. In the following, (h, s) will indicate the session h of the speaker s. The Latent Factor Analysis (LFA) model, can be written as:

$$\mathsf{m}_{(h,s)} = \mathsf{m} + Dy_s + Ux_{(h,s)} \tag{1}$$

where $\mathsf{m}_{(h,s)}$ is the session-speaker dependent supervector mean, D is $S \times S$ diagonal matrix (S is the dimension of the supervector), y_s the speaker vector (its size equal S), U is the session variability matrix of low rank R (a $S \times R$ matrix) and $x_{(h,s)}$ are the session factors, a R vector. Both y_s and $x_{(h,s)}$ are normally distributed among $\mathcal{N}(0, 1)$. D satisfies the following equation

$$I = \tau D^t \Sigma^{-1} D \tag{2}$$

where τ is the relevance factor required in the standard MAP adaptation.

The client model is obtained by performing the decomposition of equation 1 and by retaining only the speaker dependent components:

$$\mathsf{m}_s = \mathsf{m} + Dy_s \tag{3}$$

The success of the factor analysis model relies on a good estimation of the U matrix, thanks to a sufficiently high amount of data, where a high number of different recordings per speaker is available. In LFA modelling, the UBM drives the session variability estimation. Three matrices couples (male / female) are trained, driven by the previously introduced UBM-couples according to the process described in [1]. The data collections used to train the U matrices and coming from NIST-SRE'04 and NIST-SRE'05 databases are described in Table 2

U **Matrix**	**Micro '05**		**Mix2048**		**Mix512**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Number of distributions per model	512	512	2048	2048	512	512
Number of microphone speakers	45	186	45	186	45	186
Number of microphone segments	1277	1567	1277	1567	1277	1567
Number of telephone speakers	0	0	124	186	124	186
Number of telephone segments	0	0	2810	3975	2810	3975

Table 2: Description of the data used for FA matrices training.

3.3. Score Normalisation

For the three subsystems used for the LIA submission, scores are normalised using ZT-normalisation [8] (Z-norm first). For the GMM-UBM based systems, 380 male speakers (200 from the NIST-SRE'04 telephone data and 180 from the NIST-SRE'05 microphone data) and 327 female speakers (119 from the NIST-SRE'04 telephone data and 208 from the NIST-SRE'05 microphone data) are used as background data for TZ- norm. For the SVM systems, this data set is completed by a set of 180 male speakers and 180 female speakers from the Fisher database to perform score normalisation and be used as negative example to train the SVM classifiers.

4. LIA sub-systems

The three systems submitted by the LIA laboratory are based on three sub-systems described in this section.

4.1. SVM512

The first sub-system, denoted SVM512 is a classical GMM-SVM system (Gaussian Mixture Model - Support Vector Machine) using Latent Factor Analysis (LFA) [1], [2] for session variability modelling. This system is the one used for score calibration in the Human Assisted Speaker Recognition task (HASR) LIA submission.

According to Equation 1, the Factor Analysis model estimates speaker supervectors normalised with respect to the session variability. A distance between GMMs is computed by using a probabilistic kernel K [9]. This distance, well suited for a SVM classifier when only mean parameters of the GMM models are adapted, is given by Equation 4 for two sequences $\mathcal{X}_s$ and $\mathcal{X}_{s'}$ respectively spoken by speakers s and s'.

$$K(\mathcal{X}_s, \mathcal{X}_{s'}) = \sum_{g=1}^{M} (\sqrt{\alpha_g}\, \Sigma_g^{-\frac{1}{2}}\, \mathsf{m}_s^g)^t (\sqrt{\alpha_g}\, \Sigma_g^{-\frac{1}{2}}\, \mathsf{m}_{s'}^g) \tag{4}$$

where m_s is taken form the model in Equation 1 ($\mathsf{m}_s = \mathsf{m} + Dy_s$), and Σ_g is the covariance matrix of the component g shared by the two models.

The LIA SpkDet toolkit benefits from the LIB-SVM library [4] to estimate SVMs and classify instances. SVM are trained with an infinite (very large in practice) C parameter thus avoiding classification error on the training data (hard margin behaviour). The negative labelled examples are speakers from the normalisation cohort.

This system uses the UBM referred as **Mix512** in Section 3.1. Session variability is estimated according to the **Mix512** U matrix described in Section 3.2. The cohort used for score ZT-normalisation is composed of speakers from the NIST-SRE'04, NIST-SRE'05 and Fisher databases.

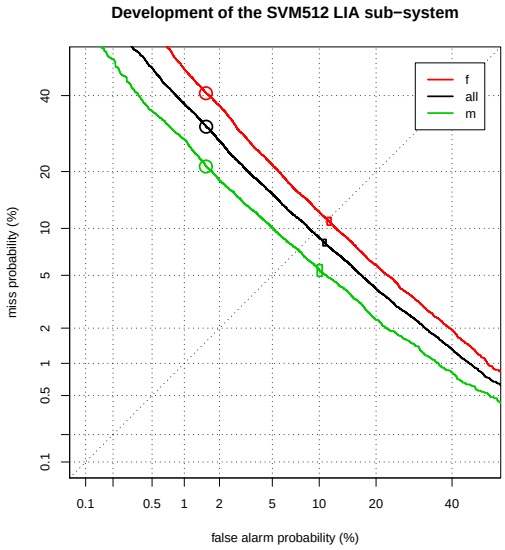


Figure 1: DET Curves for male, female and gender- independent in NIST-SRE'08 short2-short3 condition. SVM512 system, all trials.

Figure 1 shows that the system is much more powerful for males than for females, even though the system is the same for males and females. The only difference lies in the UBM and session variability estimation. As this shit can be observed for all LIA sub-systems, we suspect that the parametrisation used in LIA systems is not well adapted for female.

4.2. GMM2048

The second sub-system is a classical GMM-UBM system using Latent Factor Analysis (LFA) for session variability modelling. This system is based on the previously presented **Mix2048** UBMs and Matrices. Two versions of this sub-systems are used for final combination:

GMM2048 follows the classical UBM-GMM paradigm;

REV2048 perform the same trials by reversing the train and test speech segments roles (i.e. the LFA modelling is applied on test segments and scoring is processed using the train segments).

These two versions of the same system have shown interesting results in [10] as the UBM-GMM-LFA paradigm is not symmetric.

4.3. GMM512

The second UBM-GMM-LFA system is similar to the previous one, with different UBM and session variability estimation. This system is based on the previously presented **Micro'05** UBMs and Matrices. The collection of data used for acoustic modelling (male and female UBMs and FA matrices) differ from the one used for GMM2048 sub-system and is more adapted to microphone session variability.

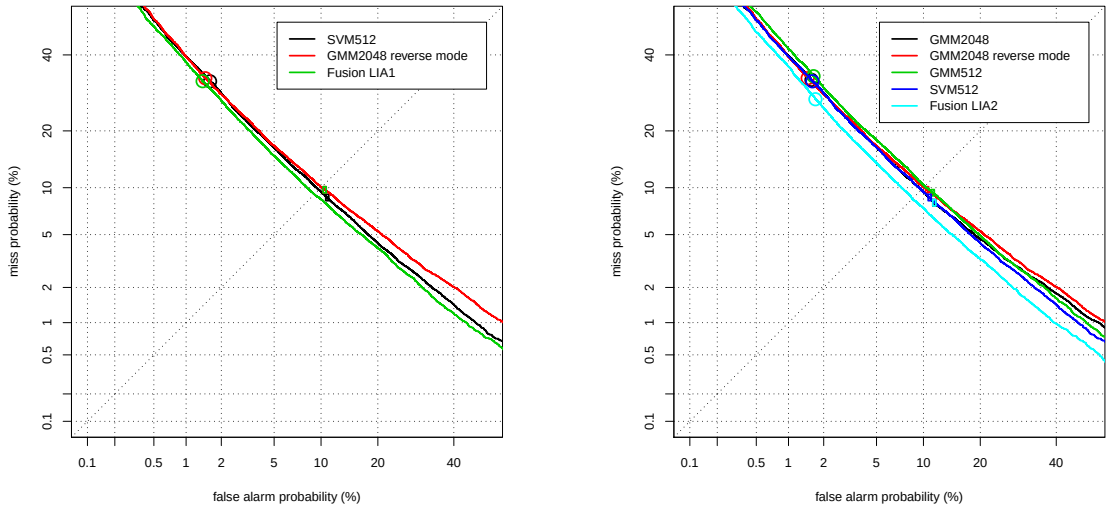
5. System combination

The three LIA submitted systems are obtained by combination of the three previously described sub-systems. Sub-systems are fused by Linear Logistic Regression (LLR) using the FoCal[1] toolkit by Niko Brümmer. No side channel information is used. The NIST SRE 2008 evaluation is used as development to determine the decision thresholds.

[1] `http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm`

6. Results

Performance of the three systems submitted by the LIA are given by Figure 2(a), 2(b) and 3 in terms of DET plots. Each submitted system DET curve comes with the DET curves of each sub-system combined.



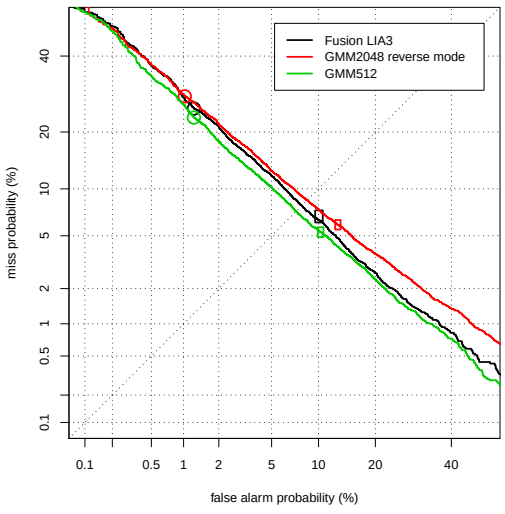
(a) Subsystems and fusion for the primary LIA1 system- all trials.

(b) Subsystems and fusion for the LIA2 system- all trials.

Figure 2: DET curves for gender independent trials in short2-short3 condition of NIST-SRE'08.

Considering the results obtained by the LIA system submitted in 2008, the 2010 primary system perform significantly better. For example, considering the det6 condition of the short2-short3 2008 task, the 2008 system reached 4.28% of DCF and 7.66% of EER when the LIA1 system obtains a DCFmin of 2.92 and 4.80% of EER for the same task. This represents a relative decrease of more than 30%.

As the LIA3 system was design to deal with microphone session variability, performance are given for trials involving interview speech segments for both training and testing phases (NIST-SRE'08 det1 condition).



ss

Figure 3: DET curves for gender independent trials in short2-short3 condition of NIST-SRE'08 (det1), subsystems and fusion for the LIA3 system.

Combination of GMM2048 in reverse mode and GMM512 allows to keep relatively good performances when microphone data are

used for both training and testing. Nevertheless, the combination does not benefits from the good performance of the GMM2048 system when considering telephone data. This phenomenon is probably due to the fact that LLR fusion coefficients are computed according to the whole test sets. However, different attempts do not succeed in improving fusion performance by splitting trials per condition.

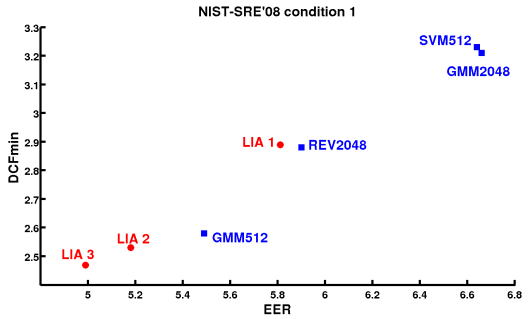
According to the development protocol (i.e. NIST-SRE 2008 short2-short3 experiment), two global conditions could be highlighted:

- a first condition in which train and test data involve the same recording condition (either telephone-telephone, or microphone-microphone);
- the second condition when train and test involve mixed recordings condition (for the 2008 evaluation, this condition involves microphone speech recordings for training and telephone data for testing).

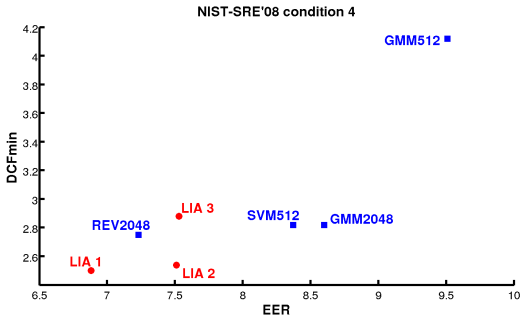
Even if Factor Analysis allows to deal with session variability and strongly improves performance by subtracting a part of the session effect, the session mismatch observed in the second global condition still a challenging issue. To deal with these global condition, the LIA laboratory submitted three systems, each dedicated to one of the following task:

- a condition in which train and test data involve the same recording condition:
  - microphone data for training and microphone data for testing;
  - telephone data for training and telephone data for testing;
- a condition in which train and test involve mixed recordings condition: microphone speech recordings for training and telephone data for testing.

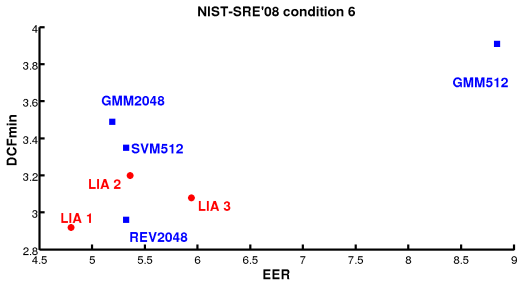
4.



(a) NIST-SRE'08 det1 condition: microphone-microphone



(b) NIST-SRE'08 det6 condition: telephone-telephone



(c) NIST-SRE'08 det4 condition: microphone-telephone

Figure 4: Performance of the LIA submitted systems and sub-systems in terms of EER % and DCFmin for the male part of the NIST-SRE'08 evaluation.

Figure 4 shows performance of the subsystems as well as the submitted combinations for three conditions of the NIST-SRE 2008 evaluation. In these three conditions (det1 = microphone-microphone, det6 = telephone-telephone and det4 = microphone-telephone), the three LIA combined systems show relatively good performance. However, it seems that each of the proposed combination remains strongly dedicated to one of these conditions. The LIA3 system is the most relevant for telephone condition. The LIA1 system seems to be the best one in case of microphone or mix (mic-phone) data and LIA2 seems to be quite a good compromise.
Figure 4(c) shows that fusing sub-systems is particularly efficient when different conditions are mixed between train and test, as illustrated by LIA1 and LIA2 systems. In this condition, LIA1 system performs surprisingly better than LIA2 or LIA3 though it is the only submitted system which does not involve the GMM512 sub-system while this system is the more dissimilar.

7. Conclusions

The submissions by LIA for SRE'10 were composed of 2 or 4 systems. Results show that the combination allows to improve performances along the whole conditions compared to single systems. Performance of the LIA submissions have been relatively improved

since the 2008 evaluation. Experiments have proved that tuning the UBM and FA training allows to strongly decrease the EER when considering relatively small session mismatch. Nevertheless, more important mismatch remains a challenging issue. Even if the results of the 2010 evaluation still not communicated at the time of writing this description, it is highly probable that mixing telephone and microphone data between train and test will remain a very challenging task for speaker verification systems. According to our point of view, this problem as to be focused by considering for example FA statistics adaptation. The gap between male and female results in the LIA submission (already existing in the previous evaluation) could also be considered as a prior issue which could find solutions in performing different parametrisation or by exploring the specificity of female voice variability.

8. References

[1] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, and Jean-Francois Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *International Conference on Speech Communication and Technology*, 2007.

[2] Patrick Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2005, vol. 1.

[3] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoît Fauve, and John S.D. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2008, http://mistral.univ-avignon.fr/.

[4] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," .

[5] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Fourth International Conference on Language Resources and Evaluation*, 2004.

[6] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[7] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA, 2005.

[8] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification System," *Digital Signal Processing*, vol. 1, no. 10, pp. 42–54, 2000.

[9] W.M. Campbell, DE Sturim, and DA Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308, 2006.

[10] Eluned S. Parris and Michael J. Carey, "Multilateral techniques for speaker recognition," in *Proceedings International Conference on Spoken Language Processing, ICSLP*, 1998.