ATVS-UAM NIST SRE 2010 SYSTEM

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
{javier.gonzalez,ignacio.lopez,javier.franco,daniel.ramos,doroteo.torre,joaquin.gonzalez}
@uam.es

Abstract

This paper describes the system submitted by ATVS-UAM to the 2010 edition of NIST Speaker Recognition Evaluation (SRE). Instead of focusing on multiple, complex and heavy systems, our submission is based on a fast, light and efficient single system. Sample development results with English SRE08 data (data used in the previous evaluation in 2008) are 0.53% EER (Equal Error Rate) in tel-tel (telephone data used for training and testing) male data (optimistic evaluation), going up to 3.5% (tel-tel) and 5.1% EER (tel-mic, telephone data for training and microphone data for testing) in pessimistic cross-validation experiments. These results are achieved with an extremely light system in computational resources, running 77 times faster than real time.

Index Terms: speaker recognition, speaker recognition evaluation, factor analysis.

1. Introduction

Our group, ATVS-UAM, has been participating in NIST (National Institute of Standards and Technology) Speaker Recognition Evaluations (SRE) since 2001. In these years speaker recognition technology has evolved dramatically, passing through different phases. During the first years of this period technology has been dominated by the Gaussian Mixture Model – Universal Background Model (GMM-UBM) technique [1]. This technique was fast and accurate but suffered great degradation with inter-session variability. For this reason, it was constantly improved by new channel and in general inter-session variability compensation schemes such as Cepstral Mean Normalization (CMN), RASTA filtering [2], Feature Warping [3], Feature Mapping [4], and so on. Since 2003 and probably up to 2007 there was a generalized trend to fuse GMM-UBM systems with what were known as 'higher-level' systems [5] because they operated on higher levels of information of the speech signal (prosodic, phonotactic, lexical, dialogic, etc.) than the acoustic level used by the GMM-UBM systems. These systems exploited information that was not taken into account by GMM-UBM systems, and therefore provided additional information that tends to fuse well with acoustic-based GMM-UBM systems. However, higher-level systems tend to be computationally expensive and result in a multiplicity of systems that make computational complexity of the overall systems very high and even prohibitive. Since 2005 [6,7] a new inter-session compensation paradigm has appeared for the GMM-UBM framework that has improved so much the performance of this technology that has made it the mainstream again, letting higher-level systems as an interesting option to reduce a few decimals in the scores of the NIST competition, but a not so interesting option for real systems. This paradigm is generally known as Joint Factor Analysis and consists in working in a high-dimensional feature space, the super-vector space, in which the feature vector is composed by the concatenation of the means of the GMM. Provided that we work with diagonal covariance GMMs, with 1024 Gaussians and a speech parameterization that provides a vector of 39 features per frame, the super-vector will include 1024 x 39 = 39936 dimensions. Once an utterance is transformed in a vector in this high-dimensional space, the Joint Factor Analysis approach tries to determine low dimension sub-spaces of this high-dimensional space that cover most of the inter-session variance and most of the inter-speaker variance. Once these sub-spaces are indentified the speaker is identified using the information in the speaker variability sub-space. More recently a new approach called total-variability [8] has been proposed that does not try to disentangle speaker and inter-session variability and rather finds a sub-space (typically of 400 dimensions) that covers most of the variability (both speaker and inter-session) by means of Principal Component Analysis (PCA). The vectors in this sub-space are then compared, after compensation using Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN), with a simple cosine distance function, showing better performance than the more complex Joint Factor Analysis approach [6]. This is the approach that we have used in our system for NIST SRE 2010. The rest of the paper is organized as follows. Section 2 describes gives a brief overview of NIST SRE 2010, focusing in particular in the data used for the evaluation. Section 3 describes feature extraction with particular emphasis on the use of two voice activity detectors, a point that we consider crucial for the success in this evaluation. Then we describe the core of our system (section 4). Finally, we describe the development and evaluation results, including measures of computational complexity (section 5) and conclude the paper in section 6.

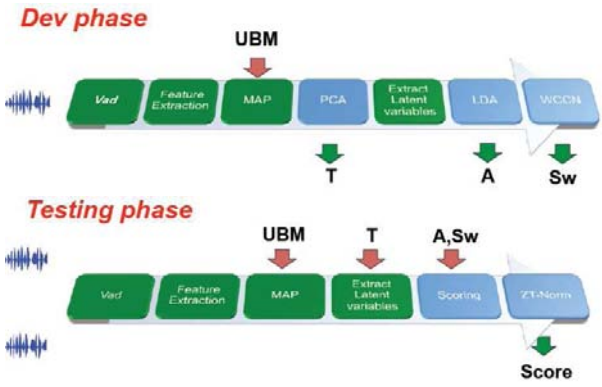

Figure 1. Developing (training) and testing phase of ATVS-UAM NIST SRE 2010 System.

Table1. Development data composition for total space training. (#Utterances/#speakers).

Gender		***Tel-Tel***	***Tel-Mic***
Male	***T/LDA***	5656/824	7868/452
	WCCN	5230/611	7838/437
Female	***T/LDA***	5155/889	10973/610
	WCCN	4521/572	10900/607

2. Overview of NIST SRE 2010

A complete description of the NIST speaker recognition evaluation is available in [9]. In general all these evaluations pose a speaker detection challenge in which the speaker models are trained on training data provided by NIST (and previously unreleased) and, after training the speaker models, these should be used to detect the speakers in test data also provided by NIST and also previously unreleased. The participants must submit their results without knowing the speaker assignments and without hearing the audios. In this paper we are only interested in one of the conditions, the so called core-core condition in which the training and testing material was one two-channel telephone conversational excerpt (we call this type of data *tel* data), of approximately five minutes total duration or a microphone recorded conversational segment (we call this type of data *mic* data) of three to fifteen minutes total duration involving the interviewee (target speaker) and an interviewer, in both cases with the target speaker channel designated. The type of data was known in advance for the systems. The evaluation established a maximum of 6000 speaker models and a maximum of 25000 test segments with a maximum of 750000 trials. The real evaluation was close to those figures.

3. Audio Processing and Feature Extraction

In our system, all audio except that used for tel-tel trials (tel data used for train and test) was first filtered with the QIO (Qualcomm-ICSI-OGI) Wiener filter in order to reduce noise [10]. Feature extraction is performed after noise reduction. It computes 38 coefficients per frame (19 Mel-Frequency Cepstrum Coefficients, MFCC, and deltas) using 20 ms. Hamming windows, overlapped 10 ms and 20 mel-spaced (300-3300 Hz) magnitude filters. Once these features are calculated three channel compensation methods are applied in sequence: CMN, RASTA [2] and Feature Warping [3] with 3 second windows.

Given that the data provided by NIST included speech from conversations, there were long periods in which the target speaker was in silence. In order to avoid processing those segments and achieve better performance we have used two different VAD (Voice Activity Detection) configurations depending on whether the data is *mic* or *tel*. *tel* audios are segmented into speech and non-speech segments combining an energy-based VAD developed by our group, and a VAD tool provided by Sound eXchange (SOX) [11] which uses speech enhancement and dynamic noise modelling. Only segments labelled as speech by both VADs are considered to be valid speech segments. For *mic* audios, we firstly remove the interviewer speech from the audio. In order to detect interviewer activity segments to remove, two different criteria have been used. The first criterion is based on an energy detector applied over the channel corresponding to the interviewer's microphone. Unfortunately for some recordings, the dynamic range was not enough for detecting any interviewer activity. In those cases, the energy based activity labels were replaced by the ASR (Automatic Speech Recognition) labels also provided by NIST (segments marked as silence was considered silence and segments with any word recognized as speech). After the interviewer speech was removed a VAD scheme equivalent to the one applied for *tel* data is used to detect valid speech segments.

4. Core Speaker Recognition

Figure 1 tries to represent the developing or training phase, and the testing phase of ATVS-UAM system. Our system is a single system based on Gaussian Mixture Models (GMM) where a 'Total Variability' modelling strategy [8] was employed in order to model both speaker and session variability. The 'total variability' scheme shares the same principles as Joint Factor Analysis (JFA) systems [6, 7], where variability (speaker and session) is supposed to be constrained, and therefore modelled, in a much lower dimensional space than the GMM-supervector space. However, unlike JFA, a *total space* which jointly includes speaker and session variability (represented by a low-rank T matrix) is computed instead of computing two separate subspaces as in JFA (matrices U and V). In our system we trained matrix T (Figure 1) with the development data shown in Table 1. After having the vectors computed in the total variability space defined by T, a session variability compensation stage is applied by means of Linear Discriminant Analysis (LDA), in which we train and use matrix A in Fig. 1, and Within-Class Covariance Normalization (WCCN), in which we train and use matrix Sw in Figure 1.

Instead of using a single total variability subspace, two gender dependent total subspaces of 200 dimensions were generated after applying LDA to a 400 (rank of T) dimensions space calculated via classical eigenanalysis from background data (Table 1). Two different *total spaces* were considered, namely tel-tel (telephone only) and tel_mic. The background, employed to construct the *total spaces* and the Universal Background Model from which GMM-supervectors models were derived (Table 1) contains a subset of data belonging to

Table 2: Breakdown timing for ATVS core system.

	GMM-FA
Testing (per 265s file)	
Total space hidden variables	0.05s
Scoring	1e-6 s
Z-norm	0.02s (~300 test)
T-norm	0.02s (~300 models)
Total (test)	3.66s
xRT test (CPU/speech)	0.013 RT

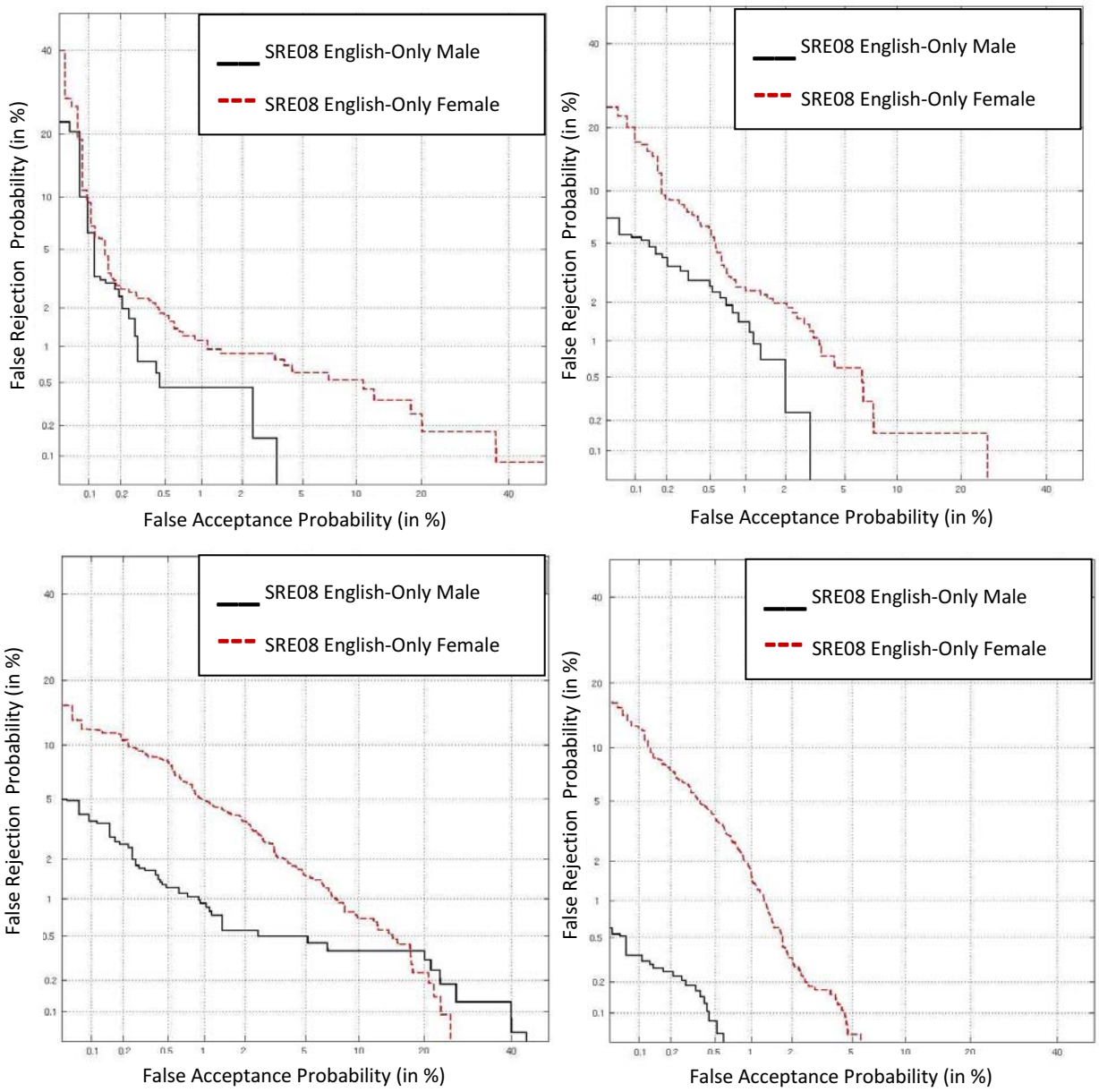

Figure 2: Development results for SRE08 english-only trials in different conditions: tel-tel (top-left), tel-mic (bottom-left), mic-tel(top-right), mic-mic(bottom-right).

Switchboard-I, Switchboard-II phase 2 and 3 and MIXER (from SREs 04, 05, 06 and 08).

The system uses a fast scoring procedure similar to [8]. Scores are then normalized using ZT-norm (Figure 1) and finally calibrated using linear logistic regression with the FoCal toolkit [12]. Calibration has been performed in a gender-independent way using different calibration rules for scores generated using microphone data in training, testing or both and scores generated using just telephone data.

5. Results

Figure 2 shows results obtained in the development phase for optimistic estimation of the T matrix (test data used for estimating it). Results range from a mere 0.53% EER (for tel-tel male and about 3% EER for tel-mic female. In order to have a less optimistic evaluation we used cross-validation excluding 25% of the test files for training 4 different T matrices and testing on the files excluded using the worst case in Figure 2. In this way we obtain Figure 3 in which EER increases up to a 5.13%, which is the result we expected in the real evaluation.

Figure 4 (a figure generated by NIST) shows the results attained by ATVS-UAM system in the real NIST SRE 2010 for the condition using interviews and the same microphone for train and test. This corresponds to our best result, a 3.5% EER. For comparison, best systems in this same condition obtain an EER slightly below 2%. Our results in other conditions can go up to 8.5% EER, which is only slightly worse than the 5.13% EER obtained in our worse development test.

Our emphasis in this evaluation was in developing an accurate and fast system. In this sense, Table 2 summarizes ATVS core system testing timing. All execution times have been obtained in a Red Hat Enterprise 5.0 server on a 2.2 GHz CPU, with cache memory of 1024 kB and RAM of 4GB. The speaker recognition process runs 77 times faster than real time, which makes the system widely applicable in real applications.

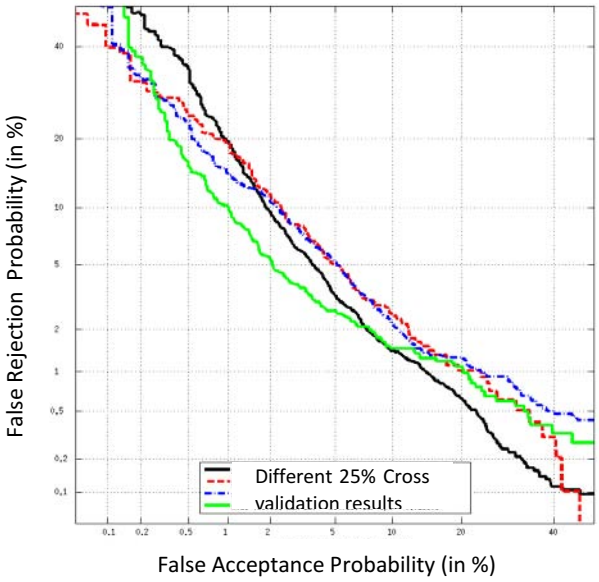

Figure 3. Cross validation development results for all SRE08 conditions, where each cross validation subset totally excludes the 25% of speakers in the subset test from the development.

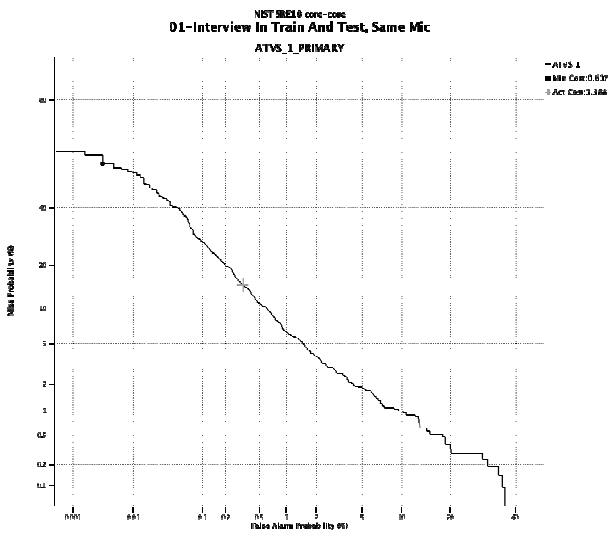

Figure 4. Actual evaluation results achieved at NIST SRE 2010. The figure corresponds to the sub-case using only interview data for train and test with the same microphone

6. Conclusions

This paper has presented the system submitted by ATVS-UAM to NIST SRE 2010. The system is based on a light and effective single system based on Total variability and achieved a 3.5% to 8.5% EER (depending on the condition) on the NIST SRE 2010 real evaluation, working over 75 times faster than real time.

7. Acknowledgements

This work has been partially supported by projects MEC TEC2009-14719-C02-01 and CAM S2009/TIC-1542 MA2VICMR.

8. References

[1] Reynolds D., Quatieri T., and Dunn R., Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10:19-41, 2000.

[2] Hermansky H. and Morgan N. Rasta processing of speech. IEEE Transactions on Speech and Audio Processing, 2(4):578-589, 1984.

[3] Pelecanos J. and Sridharan S., Feature Warping for Robust Speaker Verification, in 2001: A Speaker Odyssey: The Speaker Recognition Workshop, Crete, Greece, June 2001.

[4] Reynolds, D., Channel Robust speaker verification via Feature Mapping, in: IEEE International Conference on Acoustic Speech, and Signal Proccesing, 2003.

[5] D.A. Reynolds, et al. "The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition," Proc. ICASSP-03, Hong Kong, Apr 2003.

[6] Kenny, P. and Boulianne, G. and Dumouchel, P., "Eigenvoice Modeling With Sparse Training Data", IEEE Trans. on Speech and Audio Processing, vol. 13, no. , pp 345-354, 2005.

[7] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," Computer Speech & Language, vol. 22, no. 1, pp. 17–38, 2008.

[8] Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P and Dumouchel, P., Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification In Proc Interspeech 2009, Brighton, UK, September 2009.

[9] NIST SRE 2010 Evaluation Plan, available at http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf (accessed 23/09/2010).

[10] Qualcomm, ICSI, OGI (QIO) Front-End software, available at http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/qio/ (accessed 12/04/2010).

[11] "Sound Exchange" software, Available at http://sox.sourceforge.net/ (accessed 28/06/2010).

[12] Niko Brummer, "FoCal Toolkit", Available at http://sites.google.com/site/nikobrummer/focal (accessed 12/04/2010).