ATVS-UAM NIST SRE 2010 SYSTEM DESCRIPTION

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

1. Abstract

ATVS-UAM submits a fast, light and efficient single system. The use of a task-adapted non-speech-recognition-based VAD (apart from NIST conversation labels) and gender-dependent total variability compensation technology allows our submitted system to obtain excellent development results with SRE08 data with exceptional computational efficiency. In order to test the VAD influence in the evaluation results, a contrastive equivalent system has been submitted exclusively changing ATVS VAD labels with BUT publicly contributed ones. In all contributed systems, two gender-independent calibrations have been trained with respectively telephone-only and mic (either mic-tel, tel-mic or mic-mic) data. The submitted systems have been designed for English speech in an application-independent way, all results being interpretable in the form of calibrated likelihood ratios to be properly evaluated with Cllr. Sample development results with English SRE08 data are 0.53% (male) and 1.11% (female) EER in tel-tel data (optimistic as all English speakers in SRE08 are included in total variability matrices), going up to 3.5% (tel-tel) to 5.1% EER (tel-mic) in pessimistic cross-validation experiments (25% of test speakers totally excluded from development data in each xval set). The submitted system is extremely light in computational resources, running 77 times faster than real time. Moreover, once VAD and feature extraction are performed (the heaviest components of our system), training and testing are performed respectively at 5300 and 2950 times faster than real time.

2. Feature Extraction

Except all the tel-tel data (dev and eval), development and eval files have been filtered with the ICSI Wiener filter [1].

After this audio-in audio-out filtering, feature extraction is performed as follows:

- 20 ms. Hamming window length, overlapped 10 ms.
- 20 mel-spaced (300-3300 Hz) magnitude filters.
- 38 coefficients per frame (19 MFCC + delta).
- CMN, Rasta and 3-second window Feature Warping.

3. Voice Activity Detector (VAD)

ATVS-UAM VAD scheme used for NIST SRE 2010 has been designed as a light detector that limits the number of valid input speech segments to those proceeding only from the speaker of interest, and avoids the usage of computationally expensive VAD's such as those based on phoneme or speech recognizers. Two VAD schemes have been used according to the input data. Phonecall utterances are segmented into speech and non-speech segments combining an energy-based VAD, and a VAD tool provided by Sound eXchange organization [2] which uses speech enhancement and dynamic noise modelling. Only segments labeled as speech by both VAD's approaches are considered to be valid speech segments. For interview conversations, we firstly remove the interviewer speech from the audio. Then a VAD scheme equivalent to the one applied for phonecall data is used to detect valid speech segments. 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 B, corresponding to the interviewer's head mounted close-talking microphone. Unfortunately for some of the B channels utterances, the recorded dynamic range was not enough for detecting any interviewer activity. In those cases, the energy based activity labels were replaced by the ASR labels provided by NIST.

4. Core Speaker Recognition

The ATVS primary submitted system consisted of a single system based on Gaussian Mixture Models (GMM) where a 'Total Variability' modelling strategy [3] was employed in order to model both speaker and session variability. The 'total variability' scheme shares the same principles as Joint Factor Analysis systems [4][5], 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* (represented by a low-rank T matrix) which jointly includes speaker and session variability is computed instead of computing two separate subspaces as in JFA (matrices U and V). Then a session variability compensation stage is applied directly to the low dimensional space driven by T by means of Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN).

In order to build the ATVS core system, 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. Two different *total spaces* were considered, namely Tel (telephone only) and Tel_Mic, where phonecall-mic and interview-mic were included besides telephone data.

The background, employed to construct the *total spaces* and the Universal Background Model from which GMM-supervectors models were derived contains a subset of data belonging to SWB-I, SWB-II phase 2 and 3 and MIXER (from SREs 04, 05, 06 and 08). Tables 1 and 2 sum up the development background data composition for each total space and gender respectively.

	Tel	*Tel_Mic*
T/LDA	5656/824	7868/452
WCCN	5230/611	7838/437

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

	Tel	*Tel_Mic*
T/LDA	5155/889	10973/610
WCCN	4521/572	10900/607

Table2. Development data composition for female total space training.
(#Utterances/#speakers).

5. Calibration

A linear logistic regression scheme has been used for calibration, using the FoCal toolkit [6]. Calibration has been performed in a gender-independent way. Two different calibration rules have been used: i) scores generated using microphone data in training, testing or both utterances; and ii) scores generated using just telephone data. In any case, the pessimistic xval sets have been used to train the calibration. Comparisons containing empty files detected with VAD procedures have been scored with a log-likelihood-ratio of 0.

6. Timing/Computational Resources

Table 3 summarizes ATVS core system 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.

	GMM-FA
Development	
UBM train	• Tel: UBM, 4M feature vectors: 10h (male), 10h (female) • Tel_Mic: UBM, 5M feature vectors: 12h (male), 12h (female)
Total Variability train	Per condition: • Tel: T: 30 m (male), 30 m (female) LDA: 8 m (male), 8 m (female) WCCN: 6 m (male), 6 m (female) • Tel_Mic:

	T: 45 m (male), 1h 10 m (female) LDA: 10 m (male), 12 m (female) WCCN: 8 m (male), 10 m (female)
Feature extraction (per 265s file)	
MFCC	2s
VAD	1.57s
Training (per 265s file)	
Total space hidden variables	0.05s
Total (train)	3.62s
xRTt train (CPU/speech)	0.013 RT
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

Table3: Breakdown timing for ATVS core system.

7. References

[1] http://www.icsi.berkeley.edu/ftp/global/pub/speech/ papers/qio/

[2] http://sox.sourceforge.net/

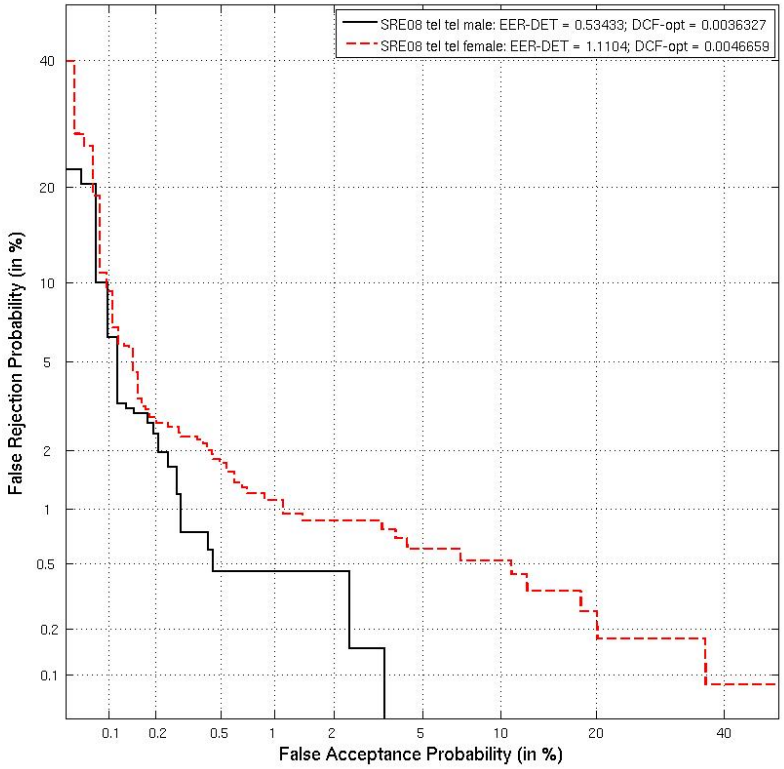
[3] Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P and Dumouchel, P., Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification In Proc Interspeech 2009, Brighton, UK, September 2009.

[4] 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.

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

[6] http://sites.google.com/site/nikobrummer/focal

ANNEX A: Development Results



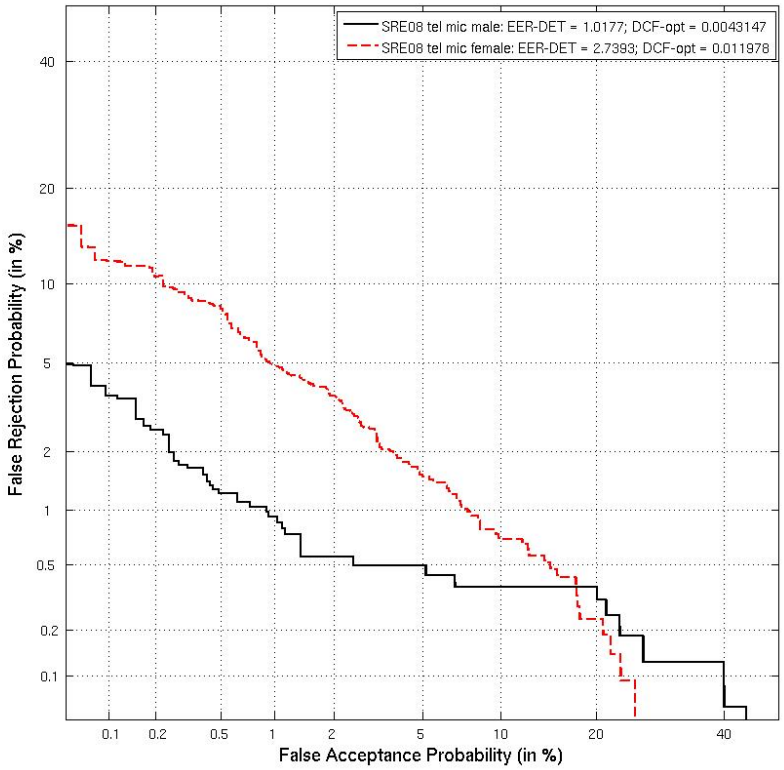
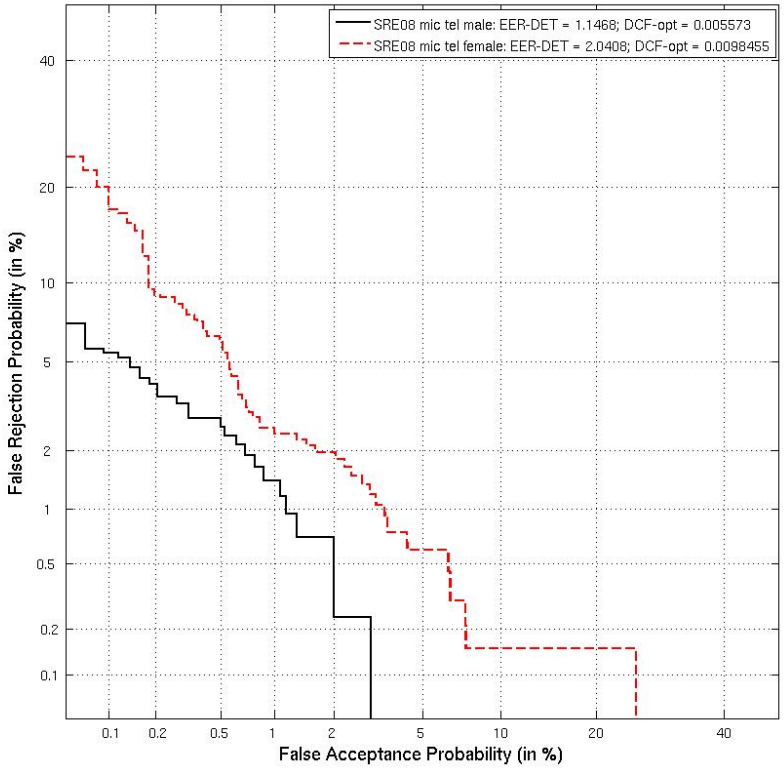


Figure1: Development results for SRE08 english-only trials in tel-tel and tel-mic conditions



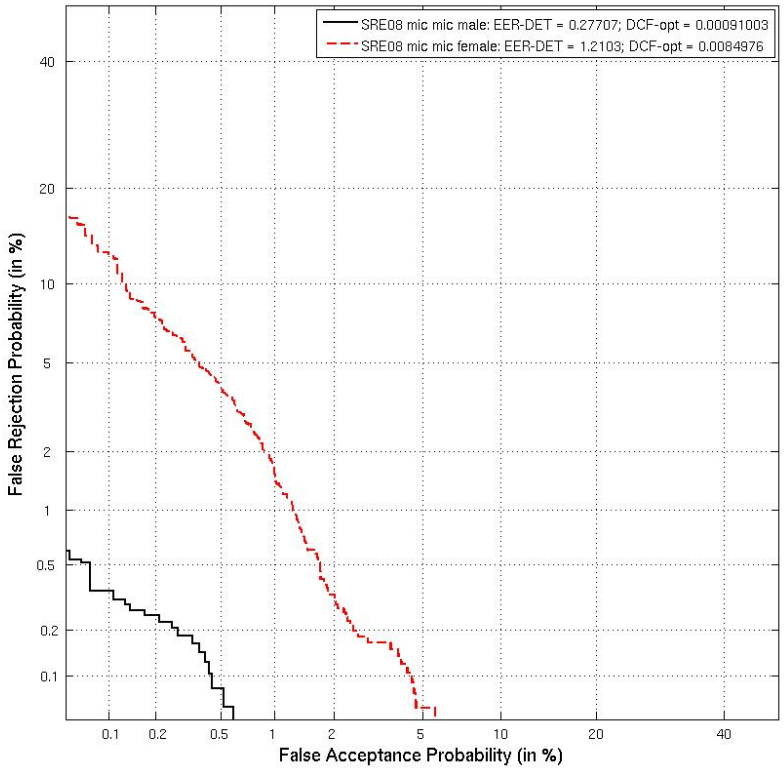


Figure2: Development results for SRE08 english-only trials in mic-tel and mic-mic conditions

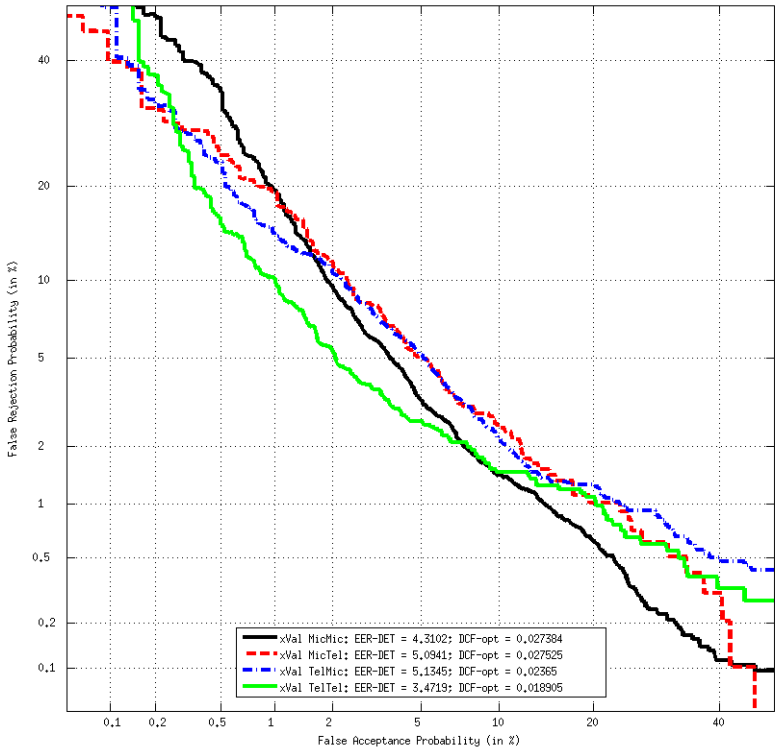


Figure3: Cross validation development results for all SRE08 conditions, where each xval subset totally excludes the 25% of speakers in the subset test from the development