University of the Basque Country System for NIST 2010 Speaker Recognition Evaluation

M. Penagarikano, A. Varona, M. Diez. L.J. Rodriguez-Fuentes, G. Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
e-mail:mikel.penagarikano@ehu.es

1 Introduction

This paper briefly describes the speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), and submitted to the NIST 2010 Speaker Recognition Evaluation. The system consists of a fusion of four subsystems: a GMM-SVM subsystem, a Linearized Eigenchannel GMM (LE-GMM) subsystem, a GLDS-SVM subsystem and a JFA subsystem.

2 Partitioning of the previous SRE databases

To implement the EHU Speaker Recognition system, the following sets were defined and used:

1. Universal Background Models (UBM)
2. Channel Compensation (CHC)
3. SVM Impostors (IMP)
4. Z-Norm score normalization (SN-ZNorm)
5. T-Norm score normalization (SN-TNorm)
6. Development set

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

In order to create these sets, SRE04 to SRE08 (including FollowUp SRE08) were used. A study of the databases was carried out to avoid including signals from the same speaker in two different sets. Table 1 shows the speaker distribution in all the databases. The main diagonal shows the number of speakers per database, elements outside the diagonal representing the number of common speakers in each pair of databases.

Table 1: Number of speakers per database (main diagonal) and and number of common speakers in each pair of databases (elements outside the diagonal).

	SRE04	***SRE05***	***SRE06***	***SRE08***	***FU08***
SRE04	310	0	0	0	0
SRE05	0	525	348	0	0
SRE06	0	348	949	112	0
SRE08	0	0	112	1336	150
FU08	0	0	0	150	150

2.1 SRE04 to SRE06

We found 1416 different speakers in the SRE04-06 sets: 180 of them (from SRE05 and SRE06) contained recordings with auxiliary microphones, whereas the remaining 1256 speakers were recorded only through different kind of telephones. Each set of speakers (either containing or not containing mic recordings) was divided into 4 different subsets (UBM, CHC, IMP and SN), and SN speakers were further divided into 2 additional sets (ZNorm and TNorm). Those speakers with the greatest number of signals acquired under different conditions where preferably assigned to the CHC set, whereas the remaining speakers were randomly distributed among the three other subsets. Table 2 shows the number of signals for the defined subsets.

Table 2: Number of signals from SRE04 to SRE06 in the Universal Background Models (UBM), Channel Compensation (CHC), SVM Impostors (IMP), and Score Normalization (ZNorm and TNorm) subsets.

	female	***male***	***Total***
UBM	2804	2119	4923
CHC	4586	3531	8117
IMP	2780	2094	4874
TNorm	1479	960	2439
ZNorm	1403	1146	2549

2.2 SRE08

Unlike previous competitions, SRE08 included in the training and test conditions, for the core test, not only conversational telephone speech data but also conversational telephone speech recorded through microphone channels in an interview scenario. 150 speakers were recorded in this new condition.

The full SRE08 database was used as development set. To avoid interactions with previous databases, the signals of the 112 speakers in common with SRE06 (see Table 1) were not used. The signals of the remaining 1224 speakers, both in train and test, were divided into two well-balanced sets for development.

Table 3: Distribution of signals in SRE08 into two balanced sets for development (devA and devB).

	SRE08	***SRE08_reduced***	***devA***	***devB***
train	3263	3149	1621	1528
test	6377	6211	3306	2905

2.3 FollowUp SRE08

The FollowUp SRE08 evaluation focused on speaker detection in the context of conversational interview speech. Test segments involved the same interview target speakers and interview sessions used in the SRE08 evaluation. Some involved the same microphone channels used in SRE08, whereas others were recorded through microphones not used previously.

The FollowUp SRE208 set, consisting of 6288 audio signals, was divided into two balanced subsets: CHC and SN, and the SN subset was further divided into two subsets: ZNorm and TNorm (see Table 4).

Table 4: Distribution of speakers and signals in FollowUp SRE08 database.

		Signals		
	Speakers	***female***	***male***	***Total***
CHC	38 *2	2432	1776	4208
TNorm	18 * 2	1145	848	1993
ZNorm	19 *2	1212	875	2087

3 The EHU Speaker Recognition System

The EHU system results from the fusion of four subsystems: a GMM-SVM subsystem, a Linearized Eigenchannel GMM (LE-GMM) subsystem, a GLDS-SVM subsystem and a JFA subsystem.

3.1 Preprocessing

The Qualcomm-ICSI-OGI (QIO)[1] noise reduction technique (based on Wiener filtering) was independently applied to the audio streams. The full audio stream was taken as input to estimate noise characteristics, thus avoiding the use of voice activity detectors on which most systems rely to constrain noise estimation to non-voice fragments.

3.2 Feature Extraction

Features were obtained with the Sautrela toolkit [2]. Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features, computed in frames of 25 ms at intervals of 10 ms. The MFCC set comprised 13 coefficients, including the zero (energy) coefficient. Cepstral Mean Subtraction (CMS) and Feature Warping were applied to cepstral coefficients. Finally, the feature vector was augmented with dynamic coefficients (13 first-order and 13 second-order deltas), resulting in a 39-dimensional feature vector.

3.3 UBM

Two gender dependent UBMs consisting of 1024 mixture components were trained with the Sautrela toolkit.

3.4 GMM-SVM & LE-GMM subsystem

The GMM-SVM and LE-GMM subsystems were built following the SUNSDV system description for SRE08 [3]. Channel compensation was trained for inter-telephone, inter-microphone and telephone-microphone variations, using 20, 20 and 40 eigenchannels respectively. For GMM-SVM, a linear kernel was trained using SMVTorch [4].

3.5 GLDS-SVM subsystem

Sufficient statistics space compensation was projected to feature space by applying the following expression:

$$\hat{f}_t = f_t - \sum_k \frac{\gamma_k(t)}{n_k} \Sigma_k^{\frac{1}{2}} c_k^x$$

where f_t is the feature vector at time t, $\gamma_k(t)$ is the posterior of gaussian k at time t, $n_k = \sum_t \gamma_k(t)$ is the zero-order statistic of gaussian k, Σ_k is the diagonal covariance matrix of gaussian k and c_k^x is the first-order statistics shift (sufficient statistics space compensation factor) of gaussian k given the input segment x. A polynomial expansion of degree 3 and a Generalized Linear Discriminant Sequence Kernel [5] were then applied.

3.6 JFA subsystem

The Joint Factor Analysis Matlab Demo from BUT [6, 7] was applied to the $MFCC + \Delta + \Delta\Delta$ features, using 200 eigenvoices and 100 eigenchannels.

3.7 ZT normalization

Trials were conditioned on three channel types: no microphone sessions (0MIC), one microphone session (1MIC) and two microphone sessions (2MIC). Gender dependent and channel type condition dependent ZT normalization was performed on trial scores.

3.8 Fusion and calibration

Side-info-conditional fusion and calibration was performed with FoCal [8], using channel type and gender conditioning. Fused scores were calibrated to be interpreted as detection log-likelihood-ratios, and the hard accept/reject decisions were made by applying a Bayes threshold of 6,907.

References

[1] Andre Adami, Lukas Burget, Stephane Dupont, Hari Garudadri, Frantisek Grezl, Hynek Hermansky, Pratibha Jain, Sachin Kajarekar, Nelson Morgan, and Sunil Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proceedings of ICSLP2002*, 2002.

[2] M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, San Juan, Puerto Rico, December 2005, pp. 386–391.

[3] Albert Strasheim and Niko Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

[4] R. Collobert and S. Bengio, "SVM Torch: Support Vector Machines for Large-Scale Regression Problems," *The Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.

[5] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proceedings of ICASSP*, 2002, pp. 161–164.

[6] *Joint Factor Analysis Matlab Demo*, http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, July 2008.

[8] *Tools for detector fusion and calibration, with use of side-information*, http://sites.google.com/site/nikobrummer/focalbilinear.