University of the Basque Country System for the NIST 2011 SRE Analysis Workshop*

Mireia Diez**, Mikel Penagarikano, Amparo Varona, Luis Javier Rodriguez-Fuentes, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
<mireia_diez@ehu.es>

Abstract. The GTTS submission for the NIST 2011 Speaker Recognition Evaluation (SRE) Analysis Workshop comprises two systems, which perform eigenchannel compensation in the sufficient statistics space and dot product scoring. Both systems differ in the method used for Channel Matrix computation: the former applies Speaker Mean Subtraction to eliminate speaker variability (which should not be modeled as eigenchannels) and then estimates the matrix by Principal Component Analysis, while the latter makes use of Speaker Locations and the Maximum-Likelihood Minimum-Divergence algorithm.

Index Terms: Speaker Recognition, NIST SRE, Dot Scoring, Sufficient Statistics, Eigenchannel Compensation

1 Introduction

In this paper we describe the speaker recognition systems developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2011 Speaker Recognition Evaluation (SRE) Analysis Workshop. These two systems combine two technologies: sufficient statistics space eigenchannel compensation and dot scoring. The first system is based on a previous work [1], where channel matrix is estimated by Principal Component Analysis (PCA). In the second system, the channel matrix is computed by means of the Maximum-Likelihood and Minimum-Divergence (ML-MD) algorithm. Furthermore, to avoid modeling speaker variability as eigenchannels, the system based on PCA applies Speaker Mean Subtraction, whereas the system based on the ML-MD algorithm makes use of Speaker Locations, as in [2].

* This work has been supported by the University of the Basque Country under grant GIU10/18, by the Government of the Basque Country, under program SAIOTEK (project S-PE10UN87), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

** Supported by a research fellowship from the Department of Education, Universities and Research of the Government of the Basque Country.

2 Sufficient statistics

Let $\lambda \equiv \{\omega_k, \mu_k, \Sigma_k | k = 1..K\}$ be a GMM consisting of K Gaussians in a F-dimensional space, with diagonal covariance matrices Σ_k. Let f_t be the feature vector at time t. Let $\gamma_k(t)$ be the posterior probability of Gaussian k at time t. We define:

$$n_k = \sum_t \gamma_k(t) \tag{1}$$

$$x_k = \sum_t \gamma_k(t) \Sigma_k^{-\frac{1}{2}} (f_t - \mu_k) \tag{2}$$

The parameter vectors $n = [\overbrace{n_1, \ldots, n_1}^{F}, \ldots, \overbrace{n_K, \ldots, n_K}^{F}]'$ and $x = [x_1, \ldots, x_K]'$ (both having size $F \cdot K$) are known as the zero and first order sufficient statistics, respectively ($'$ denotes transpose). The one-iteration relevance-MAP adapted and normalized mean vectors $m = \Sigma^{\frac{1}{2}} (\mu_{\text{MAP}} - \mu_{\text{UBM}})$ can be then computed according to the following expression [3,1]:

$$m = (\tau \mathbf{I} + diag(n))^{-1} \cdot x \tag{3}$$

3 Eigenchannel compensation

3.1 PCA System

Channel compensation in the space of sufficient statistics is performed using the eigenchannel recipe developed by the Brno University of Technology Speech Group [4].

In this method, we first model and then compensate for channel variability. Therefore, before estimating eigenchannels, we try to remove speaker variability by taking the set of feature vectors corresponding to each speaker and subtracting them the mean vector computed on all the sessions for that speaker. Eigenchannel estimation is then performed by means of PCA. All the algorithms for channel compensation were implemented using Matlab (further details can be found in [1]).

3.2 ML-MD System

In this system, eigenchannel estimation is based on the method proposed for Language Recognition in [2]. With this technique, speaker variabilities (the ones to be subtracted before eigenchannel estimation) are modeled by Speaker Locations, which are defined as follows:

$$t_s = \left(\tau I + diag \left(\sum_{f \in D(s)} n_f \right) \right)^{-1} \cdot \sum_{f \in D(s)} x_f \tag{4}$$

In Eq. (4), x_{fk} and n_{fk} stand for the zero and first order statistics computed on the feature stream f; t_s denotes the speaker location for speaker s, τ is the relevance factor and $D(s)$ denotes the set of feature streams corresponding to speaker s. Once speaker locations are estimated, a factor analysis model is defined as follows:

$$m_f = t_f + U \cdot c_f \tag{5}$$

where m_f denotes the normalized mean vector computed on the feature stream f, t_f is the speaker location corresponding to f, U is the channel matrix and c_f is a f-dependent channel factor vector. Finally, starting from the model given by Eq. (5), the channel matrix U is computed by ML-MD [5].

4 Linear Scoring

Linear scoring (dot-scoring) computes the similarity of test segments to target models by means of a linearized procedure [3]. Given a speaker s and a feature stream f (the target signal), the metric used for scoring is computed as follows:

$$\text{score}(s, f) = \hat{m}_s^T \cdot \hat{x}_f \tag{6}$$

where $\hat{m}_s^T$ is the (transposed) compensated normalized mean vector of speaker s, and $\hat{x}_f$ is the compensated first order statistics vector of target signal f.

5 Experimental setup

5.1 Partitioning of the previous SRE databases

To implement the dot-scoring speaker recognition systems, five datasets were defined and used (their names identifying the use they were given): (1) Universal Background Models, (2) Channel Compensation, (3) Z-Norm score normalization, (4) T-Norm score normalization and (5) Development. In order to create these sets, SRE04 to SRE08 (including FollowUp SRE08) were used. A partitioning of the databases was carried out to avoid including signals from the same speaker in two different sets.

5.2 Preprocessing and Feature Extraction

The Qualcomm-ICSI-OGI (QIO)[6] noise reduction technique 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.

Features were obtained with the Sautrela toolkit [7]. 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), RelAtive SpecTrAl

(RASTA) processing and short time gaussanization (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.

5.3 System configuration

The Sautrela toolkit was used to train two gender dependent UBMs consisting of 1024 mixture components, applying binary splitting, orphan mixture discarding and variance flooring.

In the both PCA and ML-MD approaches, channel compensation was trained for telephone-telephone, microphone-microphone and telephone-microphone variabilities, using 20, 20 and 40 eigenchannels, respectively. The relevance factor was set to 16.

Trials were conditioned on four different channel type conditions: 0INT-0MIC, 0INT-1MIC, 1INT-1MIC and 2MIC, where INT denotes the number of interview speech sides in the trial (telephone conversational speech otherwise) and MIC denotes the number sides in a trial recorded over a microphone channel (telephone channel otherwise). Gender dependent and channel type condition dependent ZT normalization was performed on trial scores.

The fusion of the two systems developed for this evaluation, PCA and ML-MD, was calibrated on the operating point defined for the NIST 2008 SRE (SRE08) (P_{fa}=0.01, C_{miss}=10, C_{fa}=1). Side-info-conditional calibration was performed using FoCal [8], with channel type and gender conditioning.

Table 1. DCF, Minimum DCF and Equal Error Rate (EER) for the fusion of PCA and ML-MD systems for the development set, evaluated on SRE08 (Old) and SRE10 (New) operating points, (P_{fa}=0.01, C_{miss}=10, C_{fa}=1) and (P_{fa}=0.001, C_{miss}=1, C_{fa}=1), respectively.

	Gender	EER (%)	OldDCF	OldMinDCF	NewDCF	NewMinDCF
0INT-0MIC	Male	3.04	0.150	0.132	1.180	0.620
	Female	3.16	0.144	0.138	0.644	0.610
0INT-1MIC	Male	4.84	0.240	0.228	0.383	0.323
	Female	8.78	0.332	0.313	0.639	0.421
1INT-1MIC	Male	3.03	0.159	0.154	0.386	0.272
	Female	6.40	0.288	0.273	0.613	0.503
2MIC	Male	2.84	0.125	0.119	0.818	0.456
	Female	3.12	0.183	0.179	1.248	0.792

6 Results

Table 1 shows the Detection Cost Function (DCF) as defined by the NIST for SRE08 and SRE10 [9], the Minimum DCF and the Equal Error Rate on the

development set. In order to obtain the scores for the development set, the corpus was divided into two halves. One of the halves was used for training the calibration parameters and the other for testing. The same experiment was repeated after swapping the roles of the halves. Performance was computed by accumulating results for both experiments.

Results in Table 1 show performance on the four conditions used for calibration, that is, 0INT-0MIC, 0INT-1MIC, 1INT-1MIC and 2MIC.

Table 2. DCF, Minimum DCF and Equal Error Rate (EER) for the fusion of PCA and ML-MD systems for the SRE10 (not extended) dataset for five core conditions evaluated on SRE08 (Old) and SRE10 (New) operating points, (P_{fa}=0.01, C_{miss}=10, C_{fa}=1) and (P_{fa}=0.001, C_{miss}=1, C_{fa}=1), respectively.

	Gender	EER (%)	OldDCF	OldMinDCF	NewDCF	NewMinDCF
Condition 1	Male	2.53	0.977	0.106	14.357	0.370
	Female	3.44	0.791	0.161	7.968	0.573
Condition 2	Male	4.13	0.318	0.205	2.313	0.625
	Female	6.56	0.369	0.307	1.288	0.717
Condition 3	Male	3.46	0.176	0.135	0.747	0.358
	Female	4.52	0.235	0.198	0.613	0.597
Condition 4	Male	3.67	0.558	0.157	6.502	0.470
	Female	5.00	0.514	0.224	4.460	0.803
Condition 5	Male	4.25	0.213	0.195	0.959	0.799
	Female	4.51	0.182	0.182	0.925	0.854

Table 2 shows results on the evaluation set for the system calibrated and evaluated on the SRE08 and SRE10 operating points, for the five core conditions: interview speech from the same microphone in training and test (condition 1), interview speech from different microphones in training and test (condition 2), interview training speech and normal vocal effort conversational telephone test speech (condition 3), interview training speech and conversational telephone test speech recorded over a room microphone channel (condition 4) and conversational telephone speech in training and test (condition 5). The system achieves competitive EER, but it can be seen that conditions 1 and 4 are poorly calibrated, as they are not suitably covered with the conditions used for calibration. The side-info used to calibrate our systems did not cover the type of microphone used in each trial (useful for core condition 1), and our side-info was not-directional, that is, when a trial involved different types of speech (Interview, Telephone conversational) or different recording channels (microphone, telephone) in training and test, we could not distinguish which conditions corresponded to train and which to test (useful for core conditions 3 and 4).

As expected for SRE10, the DCF and Minimum DCF in all conditions degraded significantly with regard to SRE08. Calibration errors increased in all cases in this operating point, specially for conditions 1 and 4.

7 Conclusions

The fusion of PCA and ML-MD systems achieves competitive results, but calibration errors suggest a mismatch between the development and evaluation sets. Future work should focus on developing new (more robust) techniques and on improving calibration on "extreme" operating points, such as that defined in the SRE10. Also, the extended NIST SRE 2010 dataset will be included in future experiments in order to improve performance.

References

1. M. Penagarikano, A. Varona, M. Diez., L. J. Rodriguez-Fuentes, and G. Bordel, "A speaker recognition system based on sufficient-statistics-space channel-compensation and dot-scoring," in *Proceedings of the II Iberian SLTech Workshop*, (Vigo, Spain), 2010.
2. N. Brümmer, A. Strasheim, V. Hubeika, P. Matejka, l. Burget, and O. Glembek, "Discrimitative Acoustic Language Recognition via Channel-Compensated GMM Statistics," in *Proceedings of Interspeech*, (Brighton, United Kingdom), 2009.
3. A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.
4. L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocký, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.
5. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1435–1447, May 2007.
6. A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proceedings of ICSLP2002*, 2002.
7. M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.
8. *Tools for detector fusion and calibration, with use of side-information.* http://sites.google.com/site/nikobrummer/focalbilinear.
9. *The NIST Year 2010 Speaker Recognition Evaluation Plan.* http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.