A speaker recognition system based on sufficient-statistics-space channel-compensation and dot-scoring

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

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

Abstract

This paper briefly describes the dot-scoring speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2010 Speaker Recognition Evaluation. The system does eigenchannel compensation in the sufficient statistics space and scoring is performed by a simple dot product. An optimized Matlab implementation of of the eigenchannels estimation, the channel compensation and the normalized mean vector computation is provided.

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

1. Introduction

This paper briefly describes the dot-scoring speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2010 Speaker Recognition Evaluation (SRE). This system was built following the SUNSDV system description for SRE08 [1]. The system combines two key technologies: sufficient statistics space eigenchannel compensation and dot scoring.

The rest of the paper is organized as follows. Sufficient statistics equations are described in Section 2. Eigenchannel compensation is discussed in Section 3. The linear scoring technique is introduced in section 4. The experimental setup is outlined in Section 5, including details about the partitioning of previous SRE databases, feature extraction (front-end) and configuration of the eigenchannel computation. Section 6 presents the results of the dot-scoring system in the SRE2010 evaluation. Finally, conclusions are summarized in Section 7.

2. Sufficient statistics

Let $\mathcal{N}\left(\omega, \mu^{ubm}, \Sigma\right)$ be a Gaussian Mixture Model (GMM) representing the Universal Background Model (UBM), consisting of K mixture components of dimension F and diagonal covariance matrix Σ. Let $f\left(t\right)$ be the feature vector at time t. Let $\gamma_k\left(t\right)$ be the posterior probability of mixture k at time t. Let Σ_k be the covariance matrix of mixture k. Let $repmat\left(M, i, j\right)$ be the function that replicates the matrix M $i \times j$ times, and let $vec\left(M\right)$ be the function that concatenates all the columns of matrix M in a single vector. We define:

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).

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

$$n = vec\left(repmat\left(\left[n_1\, n_2 \ldots n_K\right], F, 1\right)\right) \tag{2}$$

$$x_k = \sum_t \gamma_k\left(t\right) \Sigma_k^{-\frac{1}{2}} \left(f\left(t\right) - \mu_k^{ubm}\right) \tag{3}$$

$$x = vec\left(\left[x_1, x_2, \ldots, x_K\right]^t\right) \tag{4}$$

Vectors n and x (of size $F \times K$) are the so called zero-order and first-order sufficient statistics, respectively. Once the sufficient statistics are obtained, there is no need to use the UBM again, and therefore all the code is independent of the UBM. For example, the popular one-iteration relevance-MAP adapted and normalized mean vector $m = \frac{\mu_{map} - \mu_{UBM}}{\sigma}$ is obtained by:

$$m = \left(\tau \mathbf{I} + diag\left(n\right)\right)^{-1} \cdot x \tag{5}$$

where τ is the relevance factor and $diag\left(v\right)$ is a function that returns a diagonal matrix with values from vector v on the diagonal.

3. Eigenchannel compensation

Channel compensation in the space of sufficient statistics is performed using the eigenchannel recipe developed by the Brno University of Technology Speech Group [2]. The first order sufficient statistics are compensated as follows:

$$\hat{x} = x - diag\left(n\right) \cdot W L^{-1} W^t x \tag{6}$$

where W is the so called eigenchannel matrix, and matrix L is given by:

$$L = \mathbf{I} + W^t diag\left(n\right) W = \mathbf{I} + \sum_{k=1}^{K} n_k O_k \tag{7}$$

where $O_k = W_k^t W_k$. Channel compensation on the adapted and normalized mean vector m is performed using equation 5:

$$\hat{m} = \left(\tau \mathbf{I} + diag\left(n\right)\right)^{-1} \cdot \hat{x} \tag{8}$$

3.1. Estimation of the eigenchannel matrix

Given a data matrix $M = [m_1, \ldots, m_J]$ composed by adapted and normalized mean vectors, the eigenchannel matrix W consists of the D most significant eigenvectors of the data covariance, each eigenvector v weighted by the square roots of the corresponding eigenvalues λ:

$$W = \left[v_1 \cdot \sqrt{\lambda_1} \quad v_2 \cdot \sqrt{\lambda_2} \ \ldots \ v_D \cdot \sqrt{\lambda_D}\right] \tag{9}$$

Since the matrix M includes, among other sources of variability, speaker variability (which we would not like to compensate), the average speaker model must be subtracted from all sessions of a speaker prior to eigenchannel computation.

It is possible to obtain different eigenvectors from different types of channel variabilities (for example telephone-telephone, microphone-microphone and telephone-microphone variabilities), and then stack all of them in a single matrix. That is:

$$W = [W_{tel} \quad W_{mic} \quad W_{tel-mic}] \tag{10}$$

Note that the data covariance matrix $^1/_J MM^t$ has dimenssions $FK \times FK$ (F being around 40 and K around 1024), being unfeasible the direct computation of the eigenvectors. A possible solution is to compute the eigenvectors V of matrix $^1/_J M^t M$ (sized $J \times J$) and then project them by $W = MV$.

3.2. Matlab implementation of eigenchannel compensation

Some comments must be done about the Matlab implementation. Once the eigenchannel matrix W has been obtained, the compensated first order statistics, $\hat{x}$, can be computed from $(\boldsymbol{n}, \boldsymbol{x})$, the non-compensated zero and first order statistics, but note that the calculation of matrix L can be accelerated if products $O_k = \left\{W_k^t \cdot W_k\right\}$ are precomputed. Note also that being L positive definite, $L^{-1}\left(W^t x\right)$ can be solved by Cholesky decomposition. Finally, note that we are only interested in the D largest eigenvectors, which can be efficiently found using the Matlab `eigs` function (instead of the full `eig` version).

Listings 1, 2 and 3 show the Matlab implementations of the eigenchannels estimation, the channel compensation and the normalized mean vector computation, respectively.

4. Linear Scoring

Linear scoring (dot-scoring) makes use of a linearized procedure to score test segments against target models [1]. Given a feature stream f (the target signal) and a speaker spk, the first-order Taylor-series approximation to the GMM log-likelihood is:

$$\log P\left(f|spk\right) \approx \log P\left(f|UBM\right) + m_{spk}^t \cdot \nabla P(f|UBM) \tag{11}$$

where m_{spk} denotes the normalized mean vector of speaker spk and ∇ denotes the gradient vector w.r.t the standard-deviation-normalized means of the UBM, and $\nabla P(f|UBM) = x_f$ is the first order statistics vector of target signal f. The log-likelihood-ratio between the target model and the UBM is used for scoring, as follows:

$$score\left(f, spk\right) = \log \frac{P\left(f|spk\right)}{P\left(f|UBM\right)} \approx m_{spk}^t \cdot x_f \tag{12}$$

When channel compensation is applied, both the normalized mean vector of the speaker and the first order statistics vector of the target signal are compensated:

$$\hat{score}\left(f, spk\right) = \hat{m}_{spk}^t \cdot \hat{x_f} \tag{13}$$

The linear scoring is a very fast and effective method tha has proved to be comparable to (and sometimes even better than) Support Vector Machines (SVM) based scoring methods. Indeed, SVMs require much more computation and an extra set of impostor models.

Listing 1: Eigenchannel estimation function code in Matlab. The functions parametres are M, the data matrix containing normalized mean vectors as rows, K, the dimenssion os the feature vectors, c, a vector containing numeric speaker labels (identities) for each mean vector and D, the desired number of eigenchannels. Output values are the eigenchanel matrix W and the cell array $O = \{O_k\}$.

```
function [W,O] = EigChannEst(M,K,c,D)
  F=size(M,1)/K;
  J=size(M,2);
  % Step 1 - Speaker compensation
  for id=unique(c)
    ii=find(c == id);
    M(:,ii)=M(:,ii)-repmat(mean(M(:,ii),2)
        ,1,length(ii));
  end
  % Step 2 - Eigenchannel estimation
  opts.issym=1;
  opts.isreal=1;
  opts.disp=0;
  opts.tol=1E-3;
  [eigVec,eigVal]=eigs(1/J*M'*M,D,'lm',
      opts);
  V=eigVec*sqrt(eigVal);
  W=M*V;
  % Step 3 - Precompute O{k} matrices
  for k=1:K
    Wk=W(1+F*(k-1):F*k,:);
    O{k}=Wk'*Wk;
  end
```

5. Experimental setup

5.1. Partitioning of the previous SRE databases

To implement the dot-scoring speaker recognition system, the following sets were defined and used:

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

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.

5.1.1. 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 and SN), and SN speakers were further divided into 2 additional sets (ZNorm

Listing 2: Channel compensation function code in Matlab. The functions parametres are vectors n and x, the zero and first order sufficient statistics of the target signal, and matrices W and O, as returner by the eigenchannel estimation function. Output value is vector y, the channel compensated first-order sufficient statistics vector.

```
function y = ChannelComp(n,x,W,O)
  K=length(O);
  L=eye(size(W,2));
  for i=1:K
    L=L+n(i)*O{i};
  end
  y=x-n.*(W*(L\(W'*x)));
```

Listing 3: Normalized means function code in Matlab. The functions parametres are vectors n and x, the zero and first order sufficient statistics of the target signal, and tau, the relevance factor for MAP adaptation. Output value is vector m, adapted and normalized mean vector. Note that if x is channel compensated, them m is channel compensated too.

```
function m = NormalizedMeans(n,x,tau)
  m=x./(tau+n);
```

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.

5.1.2. 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 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

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

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

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

5.1.3. 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).

5.2. Preprocessing and Feature Extraction

The Qualcomm-ICSI-OGI (QIO)[3] 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.

Features were obtained with the Sautrela toolkit [4]. 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), RASTA 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.

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

5.3. System configuration

Two gender dependent UBMs consisting of 1024 mixture components were trained with the Sautrela toolkit, using binary spliting, ortphan mixture discarting and variance flooring.

Channel compensation was trained for telephone-telephone, microphone-microphone and telephone-microphone variabilities, using 20, 20 and 40 eigenchannels, respectively.

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.

Side-info-conditional calibration was performed with FoCal [5], using channel type and gender conditioning. 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 (derived from the SRE2010 competition costs, $P_{target} = 0.001$, $C_{miss} = 1$ and $C_{fa} = 1$).

6. Evaluation results

The year 2010 speaker recognition evaluation was part of an ongoing series of evaluations conducted by NIST. The core train and test conditions involved telephone conversational excerpts (of approximately five minutes total duration) and microphone recorded conversational segment (of three to fifteen minutes total duration), with 5460 train segments, 13344 test segments and a total of 610748 trials.

Five main conditions[1] were carried out in the core SRE2010 evaluation, according to train and test recording conditions mismatch:

1 - Interview in train and test, same mic.
2 - Interview in train and test, different mic.
3 - Interview in train and phonecall over tel channel in test.
4 - Interview in train and phonecall over mic channel in test.
5 - Phonecall in train and test, different telephone.

Figure 1 shows the DET curves for the dot-scoring system in the five core conditions. Minimum and actual cost operation points are marked with circles and asterisks, respectively. Whenever the test segment is related to microphone signals (conditions 1, 2 and 4), the DET curves show a calibration error (big distance between minimum and actual cost points). On the other hand, when the test is carried out over the telephon channel, the calibration is really good. A mismatch bewteen the designed development set and the evaluation set could explain this calibration issue.

7. Conclusions

The dot-scoring speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2010 Speaker Recognition Evaluation has been described. The system combines two key technologies: sufficient statistics space eigenchannel compensation and dot scoring. An optimized Matlab implementation of of the eigenchannels estimation, the channel compensation and the normalized mean vector computation has been provided.

The dot-scoring system attained competitive results at the NIST SRE 2010, despite being a much simpler approach compared to other methodologies. On the other hand, the calibration errors with microphone test segments suggest a mismatch bewteen the designed development set and the evaluation set.

[1] Another four conditions related to different vocal efforts were also evaluated, but they will be ignored in the current work.

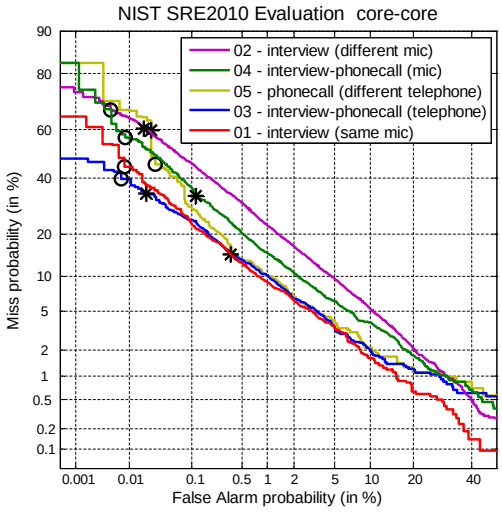

Figure 1: DET curves for the dot-scoring system in the five core conditions. DET curves are ordered in descending Equal Error Rate (ERR), the second condition (interviews with different mic) being the worst and the first one (interviews with same mic) being the best. Minimum cost operation point are marked with circles, and actual operation points with asterisks.

8. References

[1] A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

[2] 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.

[3] 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.

[4] 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.

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