Loquendo - Politecnico di Torino System Description for NIST 2010 Speaker Recognition Evaluation

Fabio Castaldo^, Daniele Colibro, Claudio Vair*, Sandro Cumani^, Pietro Laface^*

Loquendo, Torino , Italy*
{Fabio.Castaldo, Daniele.Colibro, Claudio.Vair}@loquendo.com

Politecnico di Torino, Italy^
{Sandro.Cumani,Pietro.Laface}@polito.it

Overview of the submission

The score provided as results of the joint submission of Loquendo and Politecnico di Torino primary system for SRE10 is the combination of acoustic speaker models based on Gaussian Mixture Models (GMMs) estimated by using eigenvoices [1][2] and relevance MAP [3].

In particular, the system combines the results of 8 core acoustic systems, based on two modeling approaches and four sets of features of different dimensions.

The results of the primary system were supplied for all the proposed train and test conditions.

1. System overview

These are the main modules that have been used for this year evaluation:

- Voice Activity Detection (VAD)
- Echo cancellation
- Feature extraction
- Feature warping
- Joint Factor Analysis modeling
- Total Variability modeling
- Score normalization
- Score combination and calibration
- Speaker segmentation for the summed-channel condition

In the next sections we describe these modules and the databases that have been used for training the models and for the development of the systems.

2. Voice Activity Detection

Voice Activity Detection is performed by means of a phonetic decoder. The decoder is a hybrid HMM-ANN model trained to recognize 11 language independent phone classes. Each phone class is modeled by a three state left-to-right automaton with self-loops. The ANN is a Multilayer Perceptron that estimates the posterior probability of each phone class, given an acoustic feature vector. The ANN has been trained using 20 hours of speech of 10 different languages using corpora not specifically collected for speaker recognition evaluations. More details are given in [4].

3. Echo cancellation

Echo cancellation is obtained performing voice activity detection on the two channels of the telephone conversations, and comparing the energy of the regions where the VAD reports speech both on the alternate and on the target speaker channel. If the energy of the alternate channel region is greater than the corresponding energy of the target channel, the frames of the region is labeled as "echo". Files with more than 20% of echo labels are filtered removing the echo labeled frames.

In the interview tests, the interviewee speech regions are defined as the complement of the regions where the interviewer is speaking. These regions are detected by means of our phonetic decoder if the Signal to Noise Ratio of the interviewer channel is greater than 10 dB, otherwise our phonetic transcription would be not reliable. Thus, for low SNRs interviews we rely on the information provided by the NIST ASR for the same channel.

We don't use the NIST ASR information for all the interviews because it is obtained on the close talk microphone of the interviewer before the noise masking the speech has been added. Thus, possible interviewee echoes can erroneously be assigned to interviewer speech.

4. Feature extraction

Four sets of feature have been extracted for training the models used in this evaluation, two "small" and two "large". All the features are warped by means of short term gaussianization [5].

The first set (*MFCC-25*) is the "small" one we used in the SRE08 evaluation. It includes 12 Mel Frequency Cepstral Coefficients (MFCC) plus 13 delta cepstral parameters ($\Delta c0$-$\Delta c12$) computed every 10 ms. For this set of features, the analysis bandwidth is 300-3400 Hz, and feature warping to a Gaussian distribution is performed, for each static parameter stream, on a 3 sec sliding window excluding silence frames.

All the other feature sets are extracted analyzing the full 0-4000 Hz bandwidth, and feature warping is performed before the VAD has been applied, thus including silence frames.

The second set of "small" features (*PLP-26*) includes 13 PLP coefficients (c0-c12) and their first order derivatives.

The two set of "large" features consist of 60 parameters, 20 MFCC coefficients (c0-c20) and their first and second order

derivatives, and 20 PLP parameters and their first and second order derivatives.

5. Speaker models

For this evaluation we estimated models according to the Joint Factor Analysis (JFA) [6] and the Total variability [7] approaches, which allow obtaining accurate speaker models taking into account intersession variability.

In the JFA approach a speaker model is estimated as

$$\mathbf{s} = \mathbf{UBM} + \mathbf{U}\cdot\mathbf{x} + \mathbf{V}\cdot\mathbf{y} + \mathbf{D}\cdot\mathbf{z} \quad (1)$$

The next subsections will illustrate how we estimate the terms in the model equation.

5.1. Universal Background Model (UBM)

Gender dependent Universal Background Models were trained on telephone data only. In particular on Switchboard II Phases 3, Switchboard Cellular Parts 1 and 2, and *the English conversations* of the NIST SRE 2004, 2005 and 2006 databases. The final training set (*SWB+NIST*) includes 445 hours for speech selected from the 12498 conversations of 1183 female speakers and 328 hours from 9678 conversations of 963 male speakers. The models, consisting of 2048 Gaussian mixtures, were trained running 10 iterations of an approximation of the EM algorithm, which updates for each frame only the best Gaussian statistics for the sake of efficiency.

5.2. Joint Factor Analysis

The Joint Factor Analysis (JFA) models have been trained following the guidelines of [6] and [8] with some slight variations. Gender dependent models are trained using the corresponding UBMs to collect the zero-th and first order statistics necessary for estimating the eigenvoice matrix **V**.

5.2.1. Eigenvoice subspace estimation

The eigenvoice matrix **V** has been trained using the speaker models estimated by relevance MAP on a subset of the *SWB+NIST* dataset, including at least 4 conversations per speaker. The **V** matrix is trained, thus, on English telephone speech only. The number of eigenvoices is kept fixed at 300 for all the conditions in this evaluation. The estimation of matrix is initialized by EM Principal Component Analysis [9] followed by Maximum Likelihood estimation [8].

5.2.2. Eigenchannel subspace estimation

For each conversation of the same speaker collected from different sessions, a GMM is estimated by MAP estimation of the factor analysis vector **y** in

$$\mathbf{s} = \mathbf{UBM} + \mathbf{V}\cdot\mathbf{y} \quad (2)$$

by collecting the zero-th and first order statistics from a single conversation. In addition, the average model of every speaker is obtained from all the conversation of the same speaker. The difference supervector between each speaker model and its average supervector is collected for all the available speakers, and matrix **U** in

$$\mathbf{s} = \mathbf{UBM} + \mathbf{V}\cdot\mathbf{y} + \mathbf{U}\cdot\mathbf{x} \quad (3)$$

is obtained performing Principal Component Analysis (PCA) followed by Maximum Likelihood estimation [8] using as features the difference supervectors.

Three versions of gender dependent **U** matrices have been estimated for this evaluation:

- $\mathbf{U_t}$ trained on the telephone data selected from the NIST part of the *SWB+NIST database*. (6684 and 5487 recordings of 711 female and 622 male speakers, respectively).
- $\mathbf{U_m}$ trained on the microphone data of the NIST SRE 2005 e 2006, and also including telephone conversations of the speakers contributing to the microphone databases (3461 and 2893 recordings of 95 female and 82 male speaker, respectively).
- $\mathbf{U_i}$ trained on the small set of interview data provided as development for the NIST 2008 evaluation. Training has been performed by splitting the audio files into chunks of 3 minutes and estimating a supervector for each chunk, for a total of 1520 and 1560 recordings of 3 female and 3 male speakers, respectively. We then performed the difference with respect to the corresponding chunk supervector estimated on the "clean" condition of the same session (the interviewee near microphone, channel 2). Since the speaker and the phonetic content of parallel chunks are the same, the compensation is focused on channel and microphone differences.

The dimensions of the subspaces estimated for the "small" models are 60 for the $\mathbf{U_t}$, 60 for $\mathbf{U_m}$ and 20 for the $\mathbf{U_i}$ matrices, whereas for the "large" models the dimensions becomes 100, 100, and 20, respectively.

5.2.3. Residual variability estimation

The diagonal matrix **D** describing the residual variability in the JFA speaker model (1) is set to a constant value that allows obtaining the same behavior of relevance MAP.

5.3. Speaker model training

A speaker model is estimated by JFA, stacking the **V** and **U** matrices and jointly estimating the speaker and channel factors. Relevance MAP is performed in all conditions excluding 10sec-10sec. Finally the contribution $\mathbf{U}\cdot\mathbf{x}_{\mathbf{train}}$ is discarded.

5.4. Scoring

For these models scoring was performed computing and summing the frame by frame log-likelihoods on the channel dependent model obtained adding to the channel independent GMM speaker model (3) the estimated test channel contribution

$$\mathbf{s} = \mathbf{UBM} + \mathbf{V}\cdot\mathbf{y} + \mathbf{D}\cdot\mathbf{z} + \mathbf{U}\cdot\mathbf{x}_{\mathbf{test}} \quad (5)$$

5.5. Total Variability

A second set of models, using the same features described in Section 4, has been estimated according to the Total variability approach proposed in [7]. The approach is interesting because it get rid of the distinction between speaker and channel variability in its first dimensionality reduction step, where a total variability subspace, represented by a matrix **T**, is estimated.

5.5.1. Total subspace estimation

The **T** matrix has been trained using the same dataset and features of the **V** matrix. The same procedure that allows the eigenvoice **V** matrix to be obtained can be used for estimating the total variability matrix **T**, providing the procedure a supervector per conversation rather than a supervector per speaker. Since **T** is a low rank matrix, a large number of correlated variables in a supervector is projected into the total subspace producing a small number of speaker and channel dependent uncorrelated variables: the total factor vector **w** in the model

$$\mathbf{s} = \mathbf{UBM} + \mathbf{T} \cdot \mathbf{w} \tag{4}$$

5.5.2. Intersession compensation

Intersession compensation is then performed by means of Linear Discriminant Analysis (LDA), where all the total factor vectors of the same speaker are associated with the same class. The LDA transformation $\mathbf{w'} = \mathbf{A} \cdot \mathbf{w}$ seeks a rotation matrix **A** that project the total factor vectors **w** on new axes so that the differences between the classes are maximized. The matrix A is obtained by minimizing the intra-speaker variance (caused by intersession variability of the same speaker), while the variance between speakers is maximized.
The **A** matrix has been trained using not only telephone data (*SWB+NIST*), but also the microphone and interview data sets from NIST 2006.

In these experiments the dimension of total variability matrix **T** and of the LDA matrix have been set to 400 and 200, respectively, according to the setting proposed in [7], and confirmed by our experiments on the NIST 2008 evaluation data.

5.5.3. Within Class Covariance Normalization

After LDA transformation has further reduced the feature dimensions, removing the nuisance directions, a final step is performed to normalize the speaker features by means of Within Class Covariance Normalization (WCCN) [10][9][7].

$$\mathbf{w''} = \mathbf{B}^{t} \times \mathbf{w'} \qquad \mathbf{BB}^{t} = \mathbf{W}^{-1} \tag{6}$$

where **W** is the within class covariance matrix of a subset of the training data (NIST SRE 2005 and 2006 in our settings). All the conversations of a speaker are associated to a single class.

5.5.4. Fast scoring

Scoring for these models was performed computing the value of the cosine kernel between the target speaker factors $\mathbf{w}''_{\text{target}}$ and the test factors $\mathbf{w}''_{\text{test}}$

$$k(\mathbf{w}''_{\text{test}}, \mathbf{w}''_{\text{target}}) = \frac{(\mathbf{w}''_{\text{test}})^{t} \mathbf{w}''_{\text{target}}}{\sqrt{(\mathbf{w}''_{\text{test}})^{t} \mathbf{w}''_{\text{test}}} \cdot \sqrt{(\mathbf{w}''_{\text{target}})^{t} \mathbf{w}''_{\text{target}}}} \tag{7}$$

6. Score normalization

Similar to the 2008 evaluation, the scores of each system are subject to score normalization. First the raw score are speaker-normalized by means of Z-norm. Separate statistics have been collected for the female and male speakers both for the JFA and the Total variability models.

For the JFA telephone models, the Z-norm parameters for each speaker model have been evaluated using the audio samples of 323 female and 256 male impostor speakers, a subset of speaker samples included in the SRE04 and SRE05 database. The same data have been used for training the impostor models necessary for T-normalization [11]. The T-norm parameters for each test sample were estimated using the Z-normalized scores of the impostor voiceprints.

A much larger set can be used for the Total variability models due to the fast computation of the dot-product scores. In particular, 1183 female and 963 male impostor speakers have been used for this condition.

For the 10-sec and the 8conv training and test conditions, the list of the impostor speaker samples was selected in accordance with the condition, and the impostor models were trained with the appropriate amount of data.

The list of impostor speakers for the normalization of the scores of the microphone conditions is smaller due to the relatively poor amount of data: Z-norm and T-norm is performed in this case against 164 and 190 female and male microphone models, respectively.

The normalization of the interview conditions uses the impostor speakers of the microphone data.

Some core conditions have associated the new NIST Decision Cost Function, which weights False Alarms errors a thousand times more than Miss Classification errors. In our development experiments we have found that Adaptive T-normalization [12], which finds from a large set the T-norm impostor models more similar to the current model, improved the performance of the Total variability models. The same normalization does not perform as well in the JFA framework, possibly because the selection set is kept small for the sake of efficiency.

7. Score combination and calibration

The combination of the 8 GMM systems is obtained by linear fusion with prior-weighted Logistic Regression objective [13] estimating the combination parameters on the SRE 2008 data using the FOCAL tool [14]. The estimation is condition dependent.

Lacking development data for the microphone/microphone conditions, the weights combination is borrowed by the most similar interview conditions.

Table 1. Systems and models used for the evaluation, and their approximate average processing time

Systems & Models (MFCC or PLP)	**Training Average processing time per voiceprint (sec)**	**Testing Average processing time per audio file (sec)**
Small JFA	**26**	**15**
Large JFA	**47**	**28**
Small Total variability	**4**	**4**
Large Total variability	**6**	**6**
Fusion of 8 MFCC and PLP systems	**X**	**53*2 = 106**

8. Summed-channels trials

In addition to the four wires conditions, we performed speaker model training for the summed condition. In these conditions a set of 8 whole conversations between two speakers is supplied as training audio files, and a single speaker or a summed channel conversation is proposed as test.

For the multi-speaker conversations trials we use unsupervised speech segmentation to detect speaker clusters, followed by voiceprint creation and scoring.

For the two wire tests, speaker segmentation is performed, and each putative speaker cluster is scored against the speaker models in the index list. For each model, we select the speaker cluster that gives the best score.

Our procedure for speaker clustering is described in [15]. In our development experiments, performed on the NIST 2008 data, a relevant performance boost has been obtained by using the language dependent eigenvoices estimated as described in Section 5.2.1.

In the development experiments executed on the 2008 data, we found that mislabeled gender models affect the performance of our systems. In particular, the False Alarm rate increases due to the use of gender mismatched UBMs and speaker models. Thus, before speaker recognition is performed, we execute a gender detector, based on the gender dependent UBMs. If the gender detector does not agree with the NIST supplied gender labels, and if its confidence is greater than a given threshold, the trials against that model are considered impostor trials, and their scores are randomly set to very low values.

9. Threshold setting

The theoretical log likelihood-ratio threshold decision threshold is fixed for the scores calibrated by means of the logistic regression, according to the NIST evaluation plan Decision Cost Functions, to $\log 999 \cong 6.9$ for the core and 8conv core conditions, and to $\log 9.9 \cong 2.29$ for all the other conditions.

10. Processing speed

Table 1 summarizes the systems used for our primary submission, and their approximate average processing time per voiceprint training or per audio segment scoring. These times have been obtained for the core condition on a dual quad-core Xeon 2.53 GHz Linux processor, with 32 GB of memory, exploiting at our best the underlying hardware and the organization of the tests for this evaluation.

11. References

[1] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space", IEEE Trans. on Speech and Audio Processing, Vol.8, No.6, Nov. 2000, pp. 695-707.

[2] P. Kenny, P. Dumouchel, "Disentangling Speaker and Channel Effects in Speaker Verification", in Proc. ICASSP 2004, pp. I-37-40, 2004.

[3] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[4] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition", IEEE Trans. on Audio, Speech, and Language Processing. Vol. 15-7, pp. 1969-1978, 2007.

[5] J. Pelecanos, and S. Sridharan, "Feature Warping for Robust Speaker Verification," in Proc. 2001: A Speaker Odyssey, pp. 213-218, 2001.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis versus Eigenchannels in Speaker Recognition," IEEE Transaction on Audio Speech and Language Processing, vol. 15, no. 4, pp. 1435–1447, 2007.

[7] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, P. Dumouchel1, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification", in Proc. INTERSPEECH 2009, pp-1559-1562, 2009.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," IEEE Transaction on Audio, Speech and Language, vol. 16, no. 5, pp. 980–988, 2008.

[9] M. E. Tipping and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analysis," Neural Computation, vol.11, no.2, pp. 443-482, 1999.

[10] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-Class Covariance Normalization for SVM-Based Speaker Recognition," in Proc. ICSLP 2006, pp. 1471-1474, 2006.

[11] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", Digital Signal Processing, 10 (2000), pp. 42-54.

[12] D. E. Sturim, D. A. Reynolds, "Speaker Adaptive Cohort Selection for T-norm in Text-Independent Speaker Verification", in Proc. ICASSP 2005, pp. 741-744 , 2005

[13] N. Brummer and J. du Preez "Application-Independent Evaluation of Speaker Detection" Computer Speech & Language Vol. 20, 2-3, pp. 230-275, 2006.

[14] Available at http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm

[15] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, Stream-Based Speaker Segmentation Using Speaker Factors and Eigenvoices, Proc. ICASSP-2008, pp. 4133-4136.