LOQUENDO - POLITECNICO DI TORINO'S 2010 NISTSPEAKER RECOGNITION EVALUATION SYSTEM

LOQUENDO - POLITECNICO DI TORINO'S 2010 NIST SPEAKER RECOGNITION EVALUATION SYSTEM

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

Loquendo, Torino, Italy*
{first.lastname}@ loquendo.com

Politecnico di Torino, Italy^
{first.lastname}@polito.it

ABSTRACT

This paper describes the improvement introduced in the Loquendo–Politecnico di Torino (LPT) speaker recognition system submitted to the NIST SRE10 evaluation campaign. This system combines the results of eight core acoustic systems all based on Gaussian Mixture Models (GMMs).
We illustrate the key factors, in the selection of the development data and in engineering state-of-the art technology, which contributed to the very good performance and calibration of our system in all the test conditions proposed in this evaluation.

***Index Terms*—**Speaker Recognition, Speaker Segmentation, Joint Factor Analysis, Total variability models

1. INTRODUCTION

The 2010 Speaker Recognition Evaluation (SRE10) organized by the National Institute of Standards and Technology (NIST), focused, as usual, on the speaker detection task, where the goal is to decide whether a target speaker is speaking in a segment of conversational speech. System performance is assessed using the Detection Cost Function (DCF) defined in the evaluation plan [1] and by means of Detection Error Tradeoff (DET) curves [1].

The main difference of the 2010 evaluation with respect to the previous ones is that the core test includes speech from telephone conversations, conversations recorded over a room microphone channel, and conversational speech from an interview scenario recorded over a room microphone channel. Some of the telephone conversations have been collected in a manner to produce particularly high, or particularly low, speaker vocal efforts. Moreover, the evaluation of the systems was performed according to a new Detection Cost Function that severely penalizes false acceptance costs. SRE10 included 4 training and 3 testing conditions, but only 9 different test configurations, with different amounts of speech, such as 10sec, ~5 minutes (core condition) or 8 conversations, and 2/4 wire recordings. A detailed description of the data, tasks and rules of SRE10 can be found in the evaluation plan available in [1].

One of the most important factors for the success of our system in this evaluation was the use of models obtained by Joint Factor Analysis (JFA) [3] and by the Total Variability [4] approach, which perform better than our Feature Domain Compensation technique [5] at the expense of a higher computational cost. These two technologies have been exploited to train eight systems, differing only for the number and type of acoustic features chosen to generate "complementary" systems: The scores of these systems are combined and normalized in order to obtain the final scores.

A wise usage of the development data was the second key factor that allowed our fused systems to obtain a good calibration. English speaker segments only were selected, the development set has be extended so that it was possible to reliably estimate the parameters that optimize the new DCF, and finally, we used only the interview segments in the SRE08 *development* subset for channel compensation, leaving the SRE08 training and test subsets for back-end estimation and for evaluation. In other words, we avoid partitioning the SRE08 train and test subsets to set aside interview speakers segments for channel compensation.

Complying with the new DCF raised new issues on the normalization and calibration process that has been faced using Adaptive T-norm [6] and custom development sets with many impostors.

We submitted results for all the test conditions, including the summed channel test conditions, where the speaker segments were obtained by means of the diarization technique presented in [7], using English trained eigenvectors, rather than multilingual eigenvectors. This simple replacement has shown to be effective compared with the best results reported in [8].

2. 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 subject to short term gaussianization.
The first set (MFCC-25) is the "small" one that was 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 voice activity detection 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-c19) and their first and second order derivatives, and 20 PLP parameters and their first and second order derivatives.

3. SPEAKER MODELS

For this evaluation we estimated models according to the Joint Factor Analysis (JFA) and the Total variability approaches, which allow obtaining accurate models taking into account intersession variability. Both approaches rely on GMMs estimated from a Universal Background Model (UBM).

Gender dependent UBMs were trained on telephone data only 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.

3.1. Joint Factor Analysis

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 Joint Factor Analysis (JFA) models have been trained following the guidelines of [3] and [9]. 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**.

3.1.1. Eigenvoice subspace estimation

The eigenvoice matrix **V** was trained on a subset of the *SWB+NIST* dataset, including at least 4 conversations per speaker. The **V** matrix is trained on English telephone speech only. The number of eigenvoices was kept fixed at 300 for all the conditions in this evaluation. The estimation of matrix is initialized by EM Principal Component Analysis [10] on speaker models estimated by relevance MAP, followed by Maximum Likelihood estimation [3].

3.1.2. Eigenchannel subspace estimation

For each conversation of the same speaker collected from different sessions, a GMM is estimated through 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 followed by Maximum Likelihood on centralized statistics.

Three versions of gender dependent U matrices were estimated:

- $\mathbf{U_t}$ 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 including also telephone conversations of the speakers contributing to the microphone databases (3461 and 2893 recordings of 95 female and 82 male speakers, 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 become 100, 100, and 20, respectively.

3.1.3. 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.

3.1.4. 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 $\mathbf{U}\cdot\mathbf{x_{train}}$ is discarded.

3.1.5. 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

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

3.2. Total Variability

A second set of models, using the same previously described features has been estimated according to the Total variability approach proposed in [4]. 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.

3.2.1. 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**, supplying the procedure with 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

$$\mathbf{s} = \mathbf{UBM} + \mathbf{T}\cdot\mathbf{w} \quad (6)$$

3.2.2. 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 projects the total factor vectors **w** on new axes so that the differences between the classes are maximized. 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 from NIST 2006 and interview data sets from NIST 2008.

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 [4], and confirmed by our experiments on the NIST 2008 evaluation data.

3.2.3. 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) [11][4].

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

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.

3.2.4. 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 $\mathbf{w}''_{\text{test}}$

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

4. SCORE NORMALIZATION

The scores of each system are subject to score normalization. First the raw scores are speaker-normalized by means of Z-norm. Separate statistics are 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 databases. The same data have been used for training the impostor models necessary for T-normalization. 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.

The core and 8-conv conditions were evaluated according to the new NIST DCF, which weights False Alarms errors a thousand times more than Miss Classification errors. In our development experiments we have found that Adaptive T-normalization [6], which finds, in 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.

4.1.1. Score combination and calibration

The combination of the 8 GMM systems is obtained by linear fusion with prior-weighted Logistic Regression objective, estimating the combination parameters on the SRE 2008 data using the FOCAL toolkit [12]. Parameter estimation is condition dependent. Lacking development data for the microphone/microphone conditions, the weights combination is borrowed by the most similar interview conditions.

5. 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 model 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.

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

6. THRESHOLD SETTING

To compute the Actual DCF, the theoretical log likelihood-ratio decision threshold for the scores calibrated by means of the logistic regression was fixed, according to the NIST evaluation plan DCF, 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.

7. RESULTS

The same combination of systems has been used for all the test conditions, the only difference being a condition dependent estimation of the back-end parameters.

Figure 1 and Table 1summarize the results of our system for all the *telephone call* train-test conditions (DET5). In the figure, the white and black marks refer to actual and min DCF, respectively.

Looking at the DET curves, it can be noticed that, as expected, on the 10sec test conditions, the performance improves using more training data. The DET curve of the *8summed-core* condition is near to the curve of the *8conv-core* condition because it is rather easy to detect the training speaker in eight conversations. The blue and brown curves referring to the summed-test conditions show higher errors, but again the *summed-train* condition does not affect the results as much as the *summed-test*. This is confirmed by the yellow and violet curves referring to the core conditions.

It is worth noting that in Table 1 the DCF values reported in the conditions marked by an asterisk are much higher than the others because they refer to the new more challenging DCF.

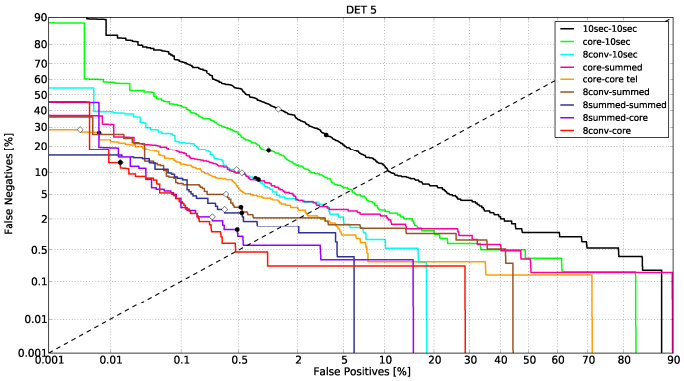

Figure 1. DET plots of the results on the telephone test conditions.

Table 1. Summary of the results on the phonecall test conditions

Condition	minDCF	actDCF	EER %
10sec-10sec	0.534	0.614	10.44
core-10sec	0.273	0.278	5.83
8conv-10sec	0.146	0.156	3.46
core-summed	0.143	0.151	3.22
core-core-tel *	0.285	0.334	2.40
8conv-summed	0.070	0.075	2.10
8summed-summed	0.057	0.065	1.45
8summed-core	0.037	0.050	0.62
8conv-core *	0.225	0.225	0.45

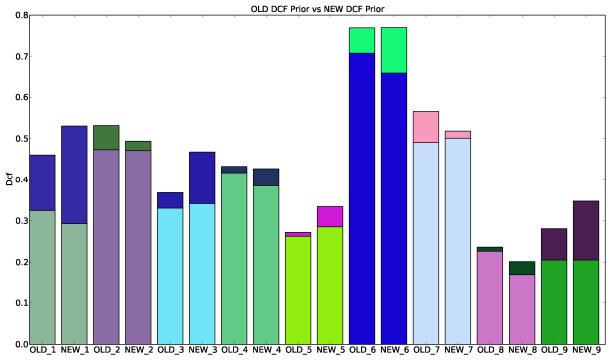

Figure 2. Actual and minimum DCF on other NIST test conditions

Table 2. Legend for Figure 2

Condition	No	vocal effort
interview_interview_same_mic	1	
interview_interview_different_mic	2	
interview_nvephonecall_tel	3	nve: normal
interview_nvephonecall_mic	4	
nvephonecall_nvephonecall_different_tel	5	
nvephoencall_hvephonecall_different_tel	6	hve: high
nvephoencall_hvephonecall_mic	7	
nvephoencall_lvephonecall_different_tel	8	lve: low
nvephoencall_lvephonecall_mic	9	

Figure 2 shows the actual and minimum DCFs in 9 conditions including recordings of interviews and of telephone calls produced by high or low vocal efforts. Each pair of bars shows the actual and minimum DCF obtained in the condition shown in Table 2, by training the back-end parameters to optimize the old or the new Decision Cost Function, respectively. Most of the times calibrating the system on the new DCF is more difficult, as shown by the differences between the actual and minimum DCFs in the Figure. Surprisingly, the low vocal effort tests do not seem to affect the performance, while the high effort condition harms our system. It is also interesting to note that in four conditions the systems tuned for the old DCF achieve better performance than the systems properly tuned for the new DCF. This raises some issues about the amount of data needed for reliably estimate the new DCF.

8. CONCLUSIONS

The experience gained in this evaluation suggests that using complementary features and models is effective. A key factor for the success of our system on the interview conditions was the use of the SRE08 development data only for channel compensation. For all conditions it was important to use a large training set, and gender dependent models. Moreover it was beneficial to use a large number of speakers for score normalization, and for intersession compensation condition-dependent matrices. Although our experiments have shown that it is possible to obtain even better results combining just 4 systems, the set of the best systems would be condition dependent.

More experimentation, possibly with larger and different amount of data is required to face still open issues such as the large variations of calibration errors of the subsystems, the effect of priors in back-end training, and the effects of vocal efforts.

9. REFERENCES

[1] National Institute of Standards and Technology, "NIST speech group web," www.itl.nist.gov/iad/mig//tests/sre/2010/ index.html

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in Proc. Eurospeech-1997, vol. 4, pp. 1895–1898.

[3] Kenny, P., Ouellet, P., Dehak, N., Gupta, V. and Dumouchel, P. "A Study of Inter-Speaker Variability in Speaker Verification" IEEE Transactions on Audio, Speech and Language Processing, Vol. 16-5 pp. 980-988, 2008.

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

[5] 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.

[6] 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.

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

[8] Kenny, P., Reynolds, D., and Castaldo, F. Diarization of Telephone Conversations using Factor Analysis IEEE Journal of Selected Topics in Signal Processing, December 2010.

[9] Burget, L., Matějka, P., Hubeika, V., Černocký, J.: Investigation into variants of Joint Factor Analysis for speaker recognition, In: Proc. Interspeech 2009, p. 1263-1266, 2009.

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

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

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