Loquendo - Politecnico di Torino’s 2008 NIST Speaker Recognition Evaluation System

LOQUENDO - POLITECNICO DI TORINO'S 2008 NIST SPEAKER RECOGNITION EVALUATION SYSTEM

Emanuele Dalmasso^, Fabio Castaldo^, Pietro Laface^, Daniele Colibro, Claudio Vair*,*

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

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

ABSTRACT

This paper describes the improvements introduced in the Loquendo–Politecnico di Torino (LPT) speaker recognition system submitted to the NIST SRE08 evaluation campaign. This system, which was among the best participants in this evaluation, combines the results of three core acoustic systems, two based on Gaussian Mixture Models (GMMs), and one on Phonetic GMMs.
We discuss the results of the experiments performed for the 10sec-10sec condition and for the core condition, including the challenging tasks involving a target speaker and an interviewer.
The error rate reduction of our SRE08 system compared to the SRE06 system ranges from 25% of the telephone-interview condition to 57% of the interview-interview condition. On the test with telephone and microphone conversations, the improvements range from 9% to 32%.

Index Terms—Speaker Recognition, Speaker Segmentation, Intersession Feature Compensation, Eigenvoices

1. INTRODUCTION

The National Institute of Standards and Technology (NIST) organizes periodically Speaker Recognition Evaluations (SRE) with the goal of encouraging the research and the development of advanced technologies in the field of text independent speaker recognition [1]. The 2008 evaluation, like the previous ones, focused on the speaker detection task, where the goal is to decide whether a target speaker is speaking in a segment of conversational speech. The performance of a system is assessed using the Detection Cost Function (DCF) [1] and Detection Error Tradeoff (DET) curves [2].

SRE08 includes 6 training conditions and 4 testing conditions for a total of 13 different test configurations, with different amounts of speech (ranging from 10sec. to 8 conversations), 2/4 wire recordings and microphone data. A new, challenging, condition was introduced for this evaluation, based on conversational segments involving a target speaker and an interviewer, collected through different microphone channels. A complete description of the data, tasks and rules of SRE08 can be found in the evaluation plan available in [1].

We had three main goals for this evaluation: considering the very good results obtained in the previous evaluation, we tried to improve our system for the most difficult conditions, i.e. the short durations and the mismatched conditions, in particular the new trials based on interview data.

In this paper we present only the results of the experiments performed for the core and for the 10sec-10sec conditions, but we submitted results for all the conditions proposed by NIST [1]. In particular, we submitted the results for the summed channel test condition obtained by using the diarization technique presented in [3], and we performed the test of unsupervised adaptation for the most difficult 10sec-10sec condition. To evaluate the technology improvement in the last two years, we also ran our SRE06 mothballed system on the SRE08 core condition.

One of the most important factor for the success of our system in this evaluation was the use of eigenvoice models [4][5] .

2. SYSTEM OVERVIEW

The system is based on the combination of recognizers based on three acoustic models: two Gaussian Mixture Models (GMMs), and a model based on Phonetic GMMs, which share the procedures for speaker modeling and for score normalization.

2.1. Acoustic features

The acoustic features that have been used for all the core systems are the standard MFCC cepstral parameters. In particular we will refer to these settings:
GMM-25: 12 cepstral (c1-c12) + 13 delta (Δc0-Δc12)
GMM-43: 18 cepstral (c1-c18) + 19 delta (Δc0-Δc18) + 6 double delta (ΔΔc0-ΔΔc5)
PGMM-36: 18 cepstral (c1-c18) + 18 delta (Δc1- Δc18)
Feature warping to a Gaussian distribution is performed on a 3 sec sliding window excluding silence frames [6].

2.2. Phonetic System

The phonetic GMM (PGMM) system used for SRE08 has the same architecture as used in SRE06 and described in [7], but uses different features and models. The system, decodes the speaker utterance, both in enrollment and in verification, producing phonetic labeled segments. The decoder is a hybrid HMM-ANN model trained to recognize 11 language independent phone classes [8]. The UBM [9] and the voiceprints for the PGMM, thus, consist of a set of phonetic GMMs, one per phone class. The number of (diagonal covariance) Gaussians per mixture per phone class is 128, for a total of 1408 Gaussians. Gender dependent UBMs have been trained using the SRE04 and SRE05 data.

The PGMM system uses the *PGMM-36* features.

We use the Phonetic decoder also as a voice activity detector (VAD), by discarding the speech intervals recognized as silence.

2.3. GMM Systems

Five GMM systems have been trained for this evaluation:

- *GMM-25-512* characterized by a small set of mixtures (512) and features (*GMM-25*).
- *GMM-43-1024* characterized by 1024 Gaussian mixtures and *GMM43* features.

- *GMM-43-2048* with the same features of the *GMM-43-1024* models, but it uses 2048 Gaussian mixtures.
- *GMM-25-512-V* has the same features and number of Gaussians of the *GMM-25-512* models, but uses a different VAD and estimates the eigenvoice matrix **V** according to [5].
- *GMM-25-1024-V* Same as *GMM-25-512-V*, but using 1024 Gaussians.

3. TRAINING PROCEDURE

Our training procedure differs from the Joint Factor Analysis approach [10] because the models are created by means of three sequential steps without iteration: intersession feature compensation in the domain of the features [8], speaker modeling by means of eigenvoices [5], and possibly final relevance MAP adaptation [9].

3.1. Intersession feature compensation

The adaptation of the feature vector at frame t of a speaker utterance is obtained by subtracting from the observation feature a weighted sum of the intersession compensation offset values according to:

$$\hat{\mathbf{o}}(t) = \mathbf{o}(t) - \sum_m \gamma_m(t) \cdot \mathbf{U}_m \cdot \mathbf{x} \quad (1)$$

where γ_m is the Gaussian posterior probability, and $\mathbf{U}_m \cdot \mathbf{x}$ is the intersession compensation offset related to the *m-th* Gaussian of the UBM model.

The low rank matrix **U**, defining the intersession variability subspace, has been trained on the SRE04 and SRE05 data. It is estimated off-line according to the following steps. For each utterance of the same speaker collected from different sessions, a GMM supervector is estimated by relevance MAP adaptation of the UBM model. Then, the set of the differences among the supervectors of the same speaker is collected for all the available speakers [11]. Finally, the matrix **U** is obtained performing Principal Component Analysis (PCA) using as features the difference supervectors by means of an EM training algorithm [12].

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

- $\mathbf{U_t}$ trained with supervectors estimated on telephone data,
- $\mathbf{U_{t+m}}$ including also microphone data
- $\mathbf{U_{t+m+i}}$ including also interview data

The number of the eigenvectors in $\mathbf{U_t}$ and in $\mathbf{U_{t+m}}$ was set to 60 and 50 respectively based on the results obtained on development data. Matrix $\mathbf{U_{t+m+i}}$ was obtained by appending the 20 most significant eigenvectors estimated on the interview development data provided by NIST for this evaluation to the first 30 eigenvectors of $\mathbf{U_{t+m}}$.

A key point for the success of our system in this evaluation has been the use of feature domain intersession compensation even for the interview data, exploiting the small development set provided by NIST as explained in Section 5.1.3.

Feature compensation is applied both in enrollment and in recognition using the $\gamma_m(t)$ statistics computed on the UBM model.

3.2. Speaker model

The speaker models are trained using intersession compensated features by means of eigen-speaker MAP modeling:

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

where the columns of matrix **V** are the so called eigenvoices [5], and **y** is a vector including the speaker factors. The $\gamma(t)$ statistics for the estimation of y are computed again on the UBM model, but using the intersession compensated features.

Matrix **V** has been trained using the speaker models estimated by relevance MAP on the suite of data of the NIST SRE evaluations of the years 1999, 2000, 2003, and 2005. We used also 1029 female and 828 male speaker models randomly selected among the speakers having at least 3 utterances in the Fisher English Training Speech Part 1 and Part 2. Overall, the number of speakers models contributing to train matrix **V** were 2079 female and 1634 male respectively. In these experiments we used 300 eigenvectors.

No use has been done of the SRE06 data for training. This database was reserved as development data for testing our models.

Finally, relevance MAP adaptation is performed using the $\gamma(t)$ statistics computed on the speaker model **s** (relevance = 16):

$$\mathbf{s}' = \mathbf{s} + \mathbf{D}\mathbf{z} \quad (3)$$

4. SCORE NORMALIZATION

The raw score are speaker-normalized by means of ZT-norm [13]. The ZT-norm parameters for each speaker model have been evaluated using a subset of speaker samples included in the SRE04 and SRE05 databases.

Separate statistics have been collected for the female and male speakers, using audio samples of 1252 female and 1103 male speakers for the one conversation telephone and microphone conditions. We refer to these data as extended normalization set with respect to the previously used normalization set including 80 female and 80 male speakers only. For the normalization of the interview conditions, the impostor speakers were selected among the microphone channel audio files. ZT-norm is performed in this case against 204 and 180 female and male models respectively.

5. EXPERIMENTAL RESULTS

Since there was a possible non null intersection between the speakers in the SRE06 and SRE08 datasets, the data of SRE06 have been used only as development set for testing the quality of our models.

5.1. Development experiments

We report the results referring to GMM systems to illustrate the improvements obtained using different models.

5.1.1. 1conv4w conditions

Table 1 is divided in two parts referring to the telephone-telephone and telephone-microphone all trial conditions respectively. The color of the rows in the Table indicates GMMs with the same number of Gaussians and features.

The first row of the table shows the baseline performance of our GMM system evaluated in 2006. The models were 512 Gaussians, gender-independent GMMs, using 25 channel compensated MFCC features [7]. The next three rows show the improvements obtained jointly using gender dependent modeling, speaker factors modeling plus relevance MAP adaptation, and the extended normalization set. The labels *UnTM* in the Table refer to the use of an intersession compensation matrix $\mathbf{U_{t+m}}$ including *n* eigenvectors.

Table 1. SRE06 development tests using different models

	SRE2006 1conv4w-4w target/impostor trials: 3552/47689		SR2006 1conv4w-mic target/impostor trials: 2566/21540	
Model	**EER %**	**MinDCF**	**EER %**	**MinDCF**
GMM SRE06	**5.88**	**0.278**	**6.42**	**0.270**
GMM-25-512 MAP U40TM	**5.57**	**0.278**	**4.72**	**0.198**
V300+D16	**5.23**	**0.264**	**3.73**	**0.179**
ExtNorm	**5.01**	**0.257**	**3.43**	**0.149**
GMM-43-1024 MAP U60TM	**5.66**	**0.272**	**4.68**	**0.217**
V300+D16	**4.62**	**0.243**	**4.48**	**0.216**
ExtNorm	**4.59**	**0.239**	**3.67**	**0.191**
U60Tel	**4.28**	**0.221**		

Increasing the number of Gaussian and parameters, a similar behavior is obtained, as shown in the successive three rows.

For the *GMM-43-1024* models, speaker factor modeling is more effective compared to *GMM-25-512*, whereas the extended normalization set produces reduced benefits. The use of a condition dependent intersession compensation matrix is beneficial as shown by the results in the last row of Table 1. The relative error reduction of the latter model compared to the baseline is greater than 20% for the SRE2006 1conv4w-1conv4w condition.

Similar results have been obtained on the microphone test condition, shown on the right side of Table 1. In this condition, however, the benefit of speaker factor modeling is more evident for 512 Gaussian GMMs. The best configuration allows a 45% relative error reduction with respect to our baseline SRE06 system.

5.1.2. 10sec-10sec condition

For the 10sec-10sec test condition, pure eigenvoice models have been trained excluding the final relevance MAP adaptation step. Table 2 shows the performance improvements obtained increasing the model complexity in term of number of Gaussians, parameters and of eigenchannels. Surprisingly, even for the 10sec-10sec test condition, bigger models do better than smaller ones. The relative error reduction is 28% for the EER and 15% for the MinDCF.

5.1.3. Interview condition

For the challenging task of the interview condition, we tried to exploit at our best the small development set made available by NIST including 3 female and 3 male speakers, 6 sessions per speaker and 9 channels per session, for a total of 324 audio files.

A critical aspect dealing with interview data is Voice Activity Detection. NIST supplied the data for testing the interview conditions together with the corresponding automatic VAD markers and ASR transcriptions, performed on the best channel for the interviewee. Although this information cannot be estimated in real conditions, we decided to use it for the evaluation, assuming that the selection of the interviewee voice was more reliable than the one obtained using our own VAD working on the actual rather than on the interviewee channel.

We used two VAD procedures. For VAD on the development set, we detected the interviewee speech comparing the energy levels of the interviewer and interviewee near field microphones. In evaluation we adopted a more complex procedure taking into account the concordance of the VAD markers and ASR transcriptions supplied by NIST.

Table 2. Model comparison using the SRE06 10sec-10sec development tests

Model	**GMM SRE06**	**GMM-25-512 V300 U40TM**	**GMM-25-1024 V300 U40TM**	**GMM-43-1024 V300 U60TM**	**GMM-43-2048 V300 U60TM**
EER %	**24.24**	**21.32**	**19.63**	**18.02**	**17.39**
MinDCF	**0.884**	**0.801**	**0.772**	**0.757**	**0.748**

In particular, we rely on the joint information of the VAD and ASR transcriptions whenever they agree for more than a given percentage of the frames in the audio file. This percentage is 40% for the short interview condition. Otherwise, we use the information given by one of the two decoders, privileging VAD over ASR, whenever it is able to cover alone more than 40% of the file frames. This percentage reduces to 30% for the long interview condition. If no useful information can be obtained from the two decoders, the speaker model is estimated using all the frames in the file, assuming that the duration of the intervals of the interviewer voice is negligible compared with the target speaker voice.

In all cases the VAD information was further filtered by the Loquendo ASR decoder.

We defined also a development test set for the interview test condition by splitting the audio files into chunks of 3 minutes according to the short evaluation condition. These chunks were used as segments for training and testing. Moreover, they play an important role for training the interview compensation matrix. In particular, we estimated a supervector for each chunk, and then we computed its 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. As usual, the compensation matrix $\mathbf{U_i}$ was trained using EM-PCA algorithm for computing 20 eigenvectors, which were appended to the first 30 eigenvectors of matrix $\mathbf{U_{t+m}}$ trained using the telephone and microphone data.

To avoid overlapping of the data used for training the $\mathbf{U_i}$ matrix with the ones used for testing the chunk models, the tests were performed using the female $\mathbf{U_i}$ matrix for recognizing male segments, and viceversa.

The interview tests were gender dependent, with uniform cross channel test distribution, and we avoided same session tests. Table 3 shows the results obtained using different models and VAD approaches on the 3 male speaker interview development tests (7200 target and 17280 impostor chunk tests). The GMM system evaluated is the one best performing on the SRE2006 1conv4w-1convmic test condition, *GMM-25-512*. In the first three columns different compensated feature are compared, starting from the ones obtained using the $\mathbf{U_{t+m}}$ estimated without interview data. The second result has been obtained using the interview data for estimating matrix $\mathbf{U_{t+m+i}}$. Label MF in the third column refers to models obtained using a matrix $\mathbf{U_{t+m+i}}$ trained on all the interview development data, pooling together male and female speakers. This matrix is the one used in the evaluation.

All these tests were performed using the energy based VAD approach for the development set. Since this approach would be unfeasible in the actual evaluation, the "concordance" approach previously described has been used. The last two columns of Table 3 compare the results obtained without and with the contribution of the VAD and ASR information provided by NIST.

Table 3. Interview development tests: results using different models and VAD procedures

Model: GMM-25-512 V300	U40TM	U30TM + U20I	U30TM +U20I MF	U30TM + U20I MF	U30TM + U20I MF
VAD	Energy based			LoqASR	Nist vad/asr +LoqASR
EER %	7.25	6.48	6.31	8.57	7.15
MinDCF	0.363	0.332	0.318	0.400	0.338

Table 4. Results on the SRE08 core condition tests (all trials)

Condition	Short2Int Short3Int	Short2Int Short3Tel	Short2Tel Short3Tel	Short2Tel Short3Mic
ERR %	3.02	4.98	6.46	5.16
ActDCF	0.169	0.233	0.358	0.226
MinDCF	0.169	0.219	0.357	0.216
MinDCF 06	0.395	0.399	0.391	0.319

Comparing columns 2 and 4 it is also worth noting that, even exploiting the NIST VAD and ASR, the obtained performance is 10% worse than using a better (but unfeasible**)** VAD procedure to detect the interviewee speech. This highlights the importance of the diarization task for the interview condition, which remains a challenging problem.

5.2. SRE08 Tests

Based on the results obtained by our systems on the SRE 2006 development data, several combinations of the core systems have been used for this evaluation:

- Telephone-telephone conditions

  Phonetic GMMs, GMM1024-43 and *GMM512-25V*.

  All models were trained using their 60 eighenchannel $\mathbf{U_t}$.

  The exception is the 10sec test conditions, where the *GMM1024-25V* replaces the *GMM512-25V* and the *GMM2048-43* replaces the *GMM1024-43*.

- Telephone-microphone conditions

  Phonetic GMMs, GMM512-25 and GMM512-25V

  All models where trained using their 50 eigenchannel $\mathbf{U_{t+m}}$

- Interview conditions

  Phonetic GMMs, GMM512-25 and *GMM512-25V*.

  All models where trained using their 50 eigenchannel $\mathbf{U_{t+m+i}}$

The combination of the systems is obtained by linear fusion with prior-weighted Logistic Regression objective. The estimation of the combination parameters was done on the most similar conditions of the SRE 2006 using the FOCAL tool [14] .

Since we lack development data for the interview conditions, the weights combination is borrowed by the most similar conditions, substituting the microphone to the interview condition.

Table 4 shows the results on the SRE08 core tests all trials. The minDCF reduction of the SRE08 system compared to the SRE06 system ranges from 25% of the telephone-interview condition to 57% of the interview-interview condition. On the telephone and microphone conditions, the improvements range from 9% to 32%. The deviation of the actual DCFs from the minDCF is small.

Table 5 shows the results obtained on the SRE08 10sec-10sec condition. EER, minimum and actual DCF are shown for the unadapted and unsupervised-adapted models (see evaluation plan in [1]), where a small improvement can be observed for the latter.

Table 5. Results on the SRE08 10sec-10sec condition

System	Unadapted	Unsupervised adaptation
ERR %	15.81	15.56
ActDCF	0.741	0.726
MinDCF	0.731	0.724

6. CONCLUSIONS

The experience gained in this evaluation suggests that, for all conditions it was important to use a large training set, gender dependent models, and speaker factor modeling. Moreover it was beneficial to use a large number of speakers for score normalization, and different, condition-dependent, matrices for intersession compensation.

In particular, the impact of speaker factors was relevant for the short duration and mismatched conditions. For the latter, also important was to estimate the **U** matrices with more data, and tuning the complexity of the models in terms of number of Gaussians and features. A key factor for the success of our system on the interview conditions was the use of the development data for channel compensation.

7. REFERENCES

[1] National Institute of Standards and Technology, "NIST speech group web,"www.nist.gov/speech/tests/sre/2008/official_results/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] 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.

[4] 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.

[5] Kenny, P., Boulianne, G. and P. Dumouchel. "Eigenvoice Modeling with Sparse Training Data" IEEE Transactions on Speech and Audio Processing, Vol.13-3, pp. 345-359, 2005.

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

[7] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, P. Laface, "Loquendo-Politecnico di Torino's 2006 NIST Speaker Recognition Evaluation System", Proc. Interspeech 2007, pp.1238-1241, 2007.

[8] 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.

[9] 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.

[10] 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.

[11] N. Brümmer, "The Spescom Data Voice NIST SRE 2004 System," presented at NIST SRE 2004 Workshop, Toledo, Spain

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

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

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