Loquendo - Politecnico di Torino's 2006 NIST Speaker Recognition Evaluation System

Loquendo - Politecnico di Torino's 2006 NIST Speaker Recognition Evaluation System

Claudio Vair [1], *Daniele Colibro* [1], *Fabio Castaldo* [2], *Emanuele Dalmasso* [2], *Pietro Laface* [2]

[1] Loquendo, Torino, Italy , [2] Politecnico di Torino, Italy

{Claudio.Vair, Daniele.Colibro}@loquendo.com,
{Fabio.Castaldo, Emanuele.Dalmasso, Pietro.Laface}@polito.it

Abstract

This paper describes the Loquendo – Politecnico di Torino system evaluated on the 2006 NIST speaker recognition evaluation dataset. This system was among the best participants in this evaluation. It combines the results of two independent GMM systems: a Phonetic GMM and a classical GMM. Both systems rely on an intersession variation compensation approach, performed in the feature domain. It allowed a 30% error rate reduction with respect to our 2005 system. The linear combination of the two GMM engines gives a further 10% error rate reduction.

We also report the results of a set of post evaluation experiments, related to the training data for the intersession variation evaluation, both for the telephone and microphone datasets. The approach adopted for the two wire tests is also described, showing the effect of the speaker segmentation component of our system. Finally, we describe how we performed the incremental unsupervised adaptation tests.

Index Terms: Speaker Recognition, Speaker Segmentation, Intersession Feature Compensation

1. Introduction

One of the main causes of relevant performance degradations in automatic speaker recognition is the mismatch that occurs when models, trained in a set of conditions, are used to test speaker data collected from different microphones, channels, and environments. Moreover, system performance is heavily affected by the intrinsic variations of different utterances of the same speaker.

Several proposals have been made to contrast these effects by means of feature transformations. Some feature based transformations, such as feature warping [1], do not rely on a specific model and can be used as an additional front-end step for any recognition system. Feature mapping [2] exploits, instead, the a priori information of a set of models trained in known conditions to map the feature vectors toward a channel independent feature space. The drawback of this approach is that it does require labeled training data to identify the conditions that one wants to compensate. Model-based techniques have been recently proposed that are able to compensate variability without requiring the explicit identification of different conditions. These techniques share a common background: modeling the differences in speaker utterances constraining them to a low dimensional space. This approach has been shown to be effective for speaker adaptation both in speech [3] and speaker recognition [4], and for channel compensation [5-7]. All these methods use MAP adapted Gaussian Mixture Models (GMM) [8] for modeling the speakers.

The system submitted by Loquendo – Politecnico di Torino to NIST 2006 Speaker Recognition Evaluation (SRE06) consists of the linear combination of two independent GMM systems: a Phonetic GMM (PGMM) and a classical GMM. Both systems adopt a new Feature Domain Intersession Compensation (FDIC) technique that adapts the observation vectors exploiting the priori knowledge of a constrained intersession variation subspace [9]. A similar technique, described in [10], has been independently developed and assessed on speech recognition.

The paper is organized as follows: Section 2 summarizes the NIST SRE06 tasks. Section 3 presents the speaker recognition systems used for the experiments. Our intersession compensation approach in the feature domain is described in Section 4. The results obtained in a set of different test conditions are given in Section 5. Section 6 reports our concluding remarks.

2. NIST 2006 SRE

The National Institute of Standards and Technology (NIST) organizes an annual Speaker Recognition Evaluation (SRE) with the goal of encouraging the research and the development of advanced technologies in the field of text independent speaker recognition [11]. The 2006 evaluation, like the previous ones, focused on the speaker detection task, where the goal is to determinate 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) [11] and Detection Error Tradeoff (DET) curves [12].

SRE06 includes 5 training conditions and 4 testing conditions for a total of 15 different test configurations, including different amounts of speech (ranging from 10sec. to 8 conversations), 2/4 wire recordings and microphone data. A complete description of the data, tasks and rules of SRE06 can be found in the evaluation plan available in [11].

In this paper we discuss the results of the experiments performed in three test conditions:

- the core test, where both training and test are performed on one side of a two channel telephone conversation lasting ~5 minutes.
- the cross channel test condition, where the testing speech was collected through a set of different microphones.
- the summed channel test condition, where the two sides of the conversation are summed in a single track.

Moreover, the effect of unsupervised adaptation is evaluated on the core test condition.

3. Systems overview

Two GMM systems have been tested in this work: a Phonetic GMM, and a classical GMM [8]. A simple linear combination of the results of these two systems was our primary system for the SRE06 evaluation. Both systems used feature domain intersession compensation FDIC [9]. The PGMM system, without FDIC, was used for the 2005 NIST evaluation.

3.1. Phonetic GMM system

The PGMM system decodes the speaker utterance, both in enrollment and verification, producing phonetically labeled segments. The decoder is a hybrid Hidden Markov Model – Artificial Neural Network (ANN) model trained to recognize 11 language independent broad phone classes. Each phone class, excluding the silence, 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 state, given an acoustic feature vector. The ANN has been trained using 20 hours of speech in 10 different languages using corpora not specifically collected for speaker recognition. The UBM and the voiceprints consist of a set of phonetic GMMs, one for each state of a phone class. The maximum number of (diagonal covariance) Gaussians per mixture is 64, for a total of 1954. This gender-independent UBM has been trained on the same data that were used for training the ANN model.

In enrollment, the labels and the boundaries of the phonetic segments are used for MAP adaptation of the parameters of the class-dependent GMMs. In recognition, the phonetically labeled audio segments are scored against their corresponding GMMs. Thus, the likelihood of a given observation vector is computed by selecting the GMM corresponding to the phone class decoded at that time frame.

The system uses 19 Mel Frequency Cepstral Coefficients (MFCC). We perform feature warping to a Gaussian distribution on each static parameter, with a 3 second sliding window excluding silences [1]. Each observation includes 36 parameters obtained by discarding the c_0 cepstral parameter, and computing the delta parameters on a symmetric window of 5 frames. The FDIC technique, described in the following Sections, is used both in enrollment and verification.

3.2. GMM system

The GMM system is characterized by a reduced set of mixtures (512), and features (13 MFCC and their deltas, excluding c_0). The gender independent UBM has been trained using data from the NIST 2000, the OGI National Cellular, and HTIMIT databases. Moreover, feature mapping [2] is performed before applying the FDIC technique. Gender and channel dependent models have been used for feature mapping, with the channels labeled as Carbon, Electret, GSM, Analog, and Digital.

4. Feature Domain Intersession Compensation (FDIC)

Gaussian Mixture Models (GMMs) used in combination with Maximum A Posteriori (MAP) adaptation [8] represent the core technology of most of the state-of-the-art text-independent speaker recognition systems. In these systems the speaker models are derived from a common GMM root model, the so called Universal Background Model (UBM), by means of MAP adaptation. Usually, only mean vector adaptation is performed during model training. A supervector that by includes all the speaker specific parameters can be obtained simply appending the adapted mean value of all the Gaussians in a single stream. The speaker model can be seen as a point in a high dimensional space, whose coordinates are the supervector's parameters. When some kind of mismatch, like the use of different microphones or communication channels, the speaking style, the phonetic content, etc. affects the input speech, all the speaker supervector parameters may be modified.

The idea behind the intersession compensation is that the distortions in the large supervector space can be summarized by a small number of parameters in a lower dimensional subspace: the *channel factors* [12].

4.1. Feature-domain adaptation

Channel factor adaptation for an utterance i and a supervector k is performed, in the supervector model space, as follows:

$$\boldsymbol{\mu}^{(i,k)} = \boldsymbol{\mu}^{(k)} + \mathbf{U}\mathbf{x}^{(i,k)} \tag{1}$$

where $\boldsymbol{\mu}^{(i,k)}$ and $\boldsymbol{\mu}^{(k)}$ are the adapted and the original supervector of GMM k respectively. $\mathbf{U}$ is a low rank matrix projecting the channel factor subspace in the supervector domain. The N-dimensional vector $\mathbf{x}^{(i,k)}$ holds the channel factors for the current utterance i and GMM k.

Channel factor adaptation in the model domain has been shown to improve the performance of GMM speaker recognition systems. The feature domain method that we propose allows the benefits of the channel factor adaptation to be exploited, mapping the compensation supervector on the acoustic features. To obtain intersession compensation in the feature domain, however, some simplifications are required.

First, since $\mathbf{x}^{(i,k)}$ should account for the distortions produced in the supervector space by the intersession variations, we expect that $\mathbf{x}^{(i,k)}$ depends on utterance i, but only slightly on speaker model k. Thus, we drop the dependency on the channel factors from the speaker model by setting $\mathbf{x}^{(i,k)} = \mathbf{x}^{(i)}$ for each model k. This simplification allows significant saving of computation time, in particular when score normalization is performed by T-norm [14], which would require the estimation of a different $\mathbf{x}^{(i,k)}$ for every impostor model k. Moreover, it offers the possibility of applying the intersession compensation directly in feature domain. We rewrite, thus, (1) for each Gaussian component m of the supervector as:

$$\boldsymbol{\mu}_m^{(i,k)} = \boldsymbol{\mu}_m^{(k)} + \mathbf{U}_m\mathbf{x}^{(i)} \qquad \forall m \tag{2}$$

where $\boldsymbol{\mu}_m^{(i,k)}$, $\boldsymbol{\mu}_m^{k}$ and $\mathbf{U}_m$ refer to the m-th Gaussian of GMM k.

The adaptation of the feature vector at time frame t, $\mathbf{O}(t)$, is obtained by subtracting from the observation feature a weighted sum of the channel compensation offset values:

$$\hat{\mathbf{O}}^{(i)}(t) = \mathbf{O}^{(i)}(t) - \sum_m \gamma_m(t)\,\mathbf{U}_m\mathbf{x}^{(i)} \tag{3}$$

where $\gamma_m(t)$ is the Gaussian occupation probability, and $\mathbf{U}_m\,\mathbf{x}^{(i)}$ is the channel compensation offset related to the m-th Gaussian of the UBM model. In the actual implementation, the right side summation of (3) is limited, for the sake of efficiency, to the first best contributes only.

4.2. Training of the channel factor subspace

The channel factor subspace, modeled by the low rank matrix $\mathbf{U}$, is assumed to represent the distortion due to the intersession variations. This distortion can be estimated by analyzing how the models of the same speaker are affected, when trained with utterances collected from different channels or conditions. Thus, the intersession factor subspace is estimated off-line according to the following steps: for each utterance of the *same speaker* collected from *different sessions*, a supervector is estimated by MAP adaptation of the UBM. Then, the set of the differences among the supervectors of the same speaker is collected for all the available speakers [6]. Finally, matrix $\mathbf{U}$ is obtained with an Expectation-maximization Principal Component Analysis (EM-PCA) algorithm [15], using as features the difference supervectors.

4.3. Estimation of the channel factors parameters

To perform channel adaptation, the channel factors vector **x** must be estimated for each test utterance. A Maximum Likelihood Eigen-Decomposition solution to a similar problem has been proposed for speaker adaptation in [3]. For speaker verification, a technique called Probabilistic Subspace Adaptation (PSA), which uses MAP estimation of **x** has been presented in [4]. In our experiments, we perform a single iteration of the PSA estimation, obtaining the vector $\mathbf{x}^{(i)}$ for each test or training utterance *i* in (3).

5. Experimental results

In this section we show the results obtained by the GMM and PGMM systems submitted to SRE06, for different test conditions. The results are given in term of Equal Error Rate (EER) and minimum NIST Detection Cost Function (DCF).

All the results are obtained by normalizing the raw verification scores. First, the scores are normalized by means of Z-norm. The parameters for each speaker model have been estimated using a subset of speaker samples included in the NIST SRE04 database. Separate statistics have been collected for the female and male speakers, using two conversations of 80 speakers for each gender. Test dependent normalization is performed using T-norm [14]. A fixed set of impostor models was selected from the voiceprints enrolled with data belonging to SRE04. The T-norm parameters for each test sample were estimated using the Z-norm scores of the impostor voiceprints. We refer to the Z-norm followed by T-norm as ZT-norm.

5.1. Core test condition

The primary system submitted to SRE06 consists of the linear combination of the GMM and the PGMM classifiers. The performance of the combined system (PGMM+GMM) and those of the component systems are given in Table 1 in terms of Equal Error Rate (EER) and minimum Detection Cost Function (DCF), measured on the core test, all trials. The first row of the Table shows the performance of the "mothballed" system used in the SRE05 evaluation and tested on SRE06. The 2005 PGMM baseline system did not include the FDIC technique.

Table 1: *Equal Error Rate and minimum Detection Cost Function on SRE06 core test condition, all trials*

System	Subspace	EER(%)	DCF
PGMM 2005	NO	8.7	0.406
PGMM	40 tel.	6.0	0.280
GMM	40 tel.	5.9	0.271
PGMM+GMM	40 tel.	4.9	0.236

The number of intersession factors used by the GMM and PGMM systems is fixed at 40. The intersession compensation subspace was trained using data coming from the SRE04 and the SRE05 datasets.

5.2. Cross channel test

Since 2005, NIST encourages the submission of results on a cross channel test condition. These tests require that the enrollment of the voiceprints is done on telephone speech, and that the verification is performed on audio recordings made using microphones. The SRE06 includes conversation collected simultaneously through eight different devices, ranging from low cost PC microphones to high quality professional transducers.

For these tests we submitted again the results of a linear combination of a PGMM and a GMM classifier. The subspace of the GMM system was retrained, for these tests, using the differences between microphone and telephone recordings, according to the testing condition. The PGMM, on the other hand, used the same subspace matrix as the core test.

Table 2. *Equal Error Rate and minimum DCF on the SRE06 cross channel test condition, all trials*

System	Subspace	EER	DCF
PGMM	40 tel.	10.5	0.365
GMM	40 mic.	6.6	0.271
PGMM AGC	20 tel.	7.8	0.320
PGMM AGC	20 tel+mic	6.4	0.250
PGMM AGC+GMM	20 tel+mic, 40 tel	5.1	0.201

During post evaluation we noticed that the performance of the submitted PGMM was very poor when compared with the GMM system. This was due to the very low signal level of the many microphone recordings, which has a strong impact on the phonetic decoding in our PGMM, because many phonemes were recognized as silence and discarded. To overcome this problem we inserted an automatic gain control (AGC) in the PGMM system, so increasing its performance as shown in the third row of Table 2. Moreover, we extended the data for training the intersession subspace including the microphone recording available in SRE05. The results obtained (fourth row of Table 2), were comparable with the ones obtained with the GMM system, even using a reduced number of channel factors (20 vs. 40, for the sake of efficiency). The linear combination of the two systems further improves the performance as shown in the last row of the Table. It is worth noting that the intersession compensation using this new subspace gives good results even on telephone data: the performance on the core test condition is EER 6.1%, DCF 0.277.

5.3. Summed channel test

In addition to the four wires (4w) test condition, NIST proposes a set of tests involving two wire (2w) recordings. In the former condition, each audio file in the enrollment and verification lists includes a single side of a conversation, i.e. the voice of one speaker, whereas in the 2 wire condition a whole conversation between two speakers is supplied as training or test audio file.

We report here the results referring to the four wire training and two wire testing condition (1conv4w-1conv2w). This test has the goal of producing a score related to the probability that one of the two speakers involved in a conversation is the target speaker.

We performed the 2w tests, preprocessing the incoming audio using our automatic speaker segmentation approach [16] to produce two audio tracks, each containing the voice of a single unknown speaker. Since the automatic segmentation is not perfect, we are interested in evaluating the impact of the segmentation errors on speaker detection. Both audio tracks produced after segmentation are scored against the target model, and the best of the two scores is produced as the matching result. It is worth noting that this procedure, even neglecting the segmentation errors, will produce less accurate results compared with the corresponding 4w tests, due to the increase of the probability of false alarms (FA). The probability of not having a FA on the whole conversation is the product of the probabilities of not having FAs on both the segmented sides. This behavior is similar to that described for multi-target detection in [17].

To obtain a one-to-one comparison with the 4w test condition, we defined a new "unofficial" 2w test, by summing the two sides of all the recordings in the list of the core test.

Table 3. *Equal Error Rate and minimum DCF on 4 and 2 wire test conditions, all trials*

System	Test	EER(%)	DCF
PGMM	4w	6.0	0.280
PGMM	4w + 4w	7.3	0.345
PGMM	2w unofficial	8.4	0.385

The first row of Table 3 reports again the results of the baseline PGMM system on the 4w core test condition. The second row shows the result obtained, without segmentation, on a 4 wire extended test, where both sides of a conversation have been taken into account and the best matching decision rule, used for the 2 wire tests, has been applied. This result highlights the impact of the anticipated increase of the FAs on the system accuracy, even in the absence of automatic speaker segmentation. Finally, the last row of the Table shows the performance on the unofficial 2w test, including the effect of automatic segmentation. Comparing the results shown in Table 3, it is interesting to observe that the main source of accuracy degradation is the presence of both sides of the conversation in the trials. Further degradation of the results is due not only to automatic segmentation errors, but also to the occurrence of overlapped speech in the summed 2 wire signals.

5.4. Unsupervised adaptation test

Another task suggested in SRE is the "unsupervised adaptation mode". This test condition allows adaptation of the target models based on previous trial segments, whenever a correct match has been hypothesized. NIST rules require that the decoding trials are performed in the order given in the supplied test index files.

To select the trials that we used for adaptation, we performed the 1conv4w-1conv4w unsupervised adaptation test using the outcome of the unadapted PGMM system. The selection has been driven by a quite conservative threshold on the ZT-normalized score, set to 4. There are two reasons for using the unadapted scores for the selection: first, it reduces the risk of divergence of the target models due to the adaptation of the models with impostor data; second it simplifies the batch ZT-normalization process of the adapted voiceprints. On the other hand, using the adapted target models for trials selection could further increase the benefits of adaptation.

Table 4. *EER and minimum DCF on SRE06 core test condition, all trials, with and without adaptation*

System	Adaptation	EERs	DCF
GMM+PGMM	NO	4.9	0.236
GMM+PGMM	YES	4.5	0.202

Using a decision threshold set to 4 in our test we correctly selected 72% of the target data for adaptation. Among the selected data (trials with ZT-norm score > 4) only 5% were recordings of a speaker other than the target one. Table 4 compares the results of our unadapted and adapted systems. The unsupervised adaptation allows a 14% relative reduction on the DCF.

Future research will be devoted to studying the effect of using the adapted scores of the combined systems for selecting the adaptation trials.

6. Conclusions

The Loquendo – Politecnico di Torino system evaluated on the 2006 NIST speaker recognition evaluation has been described. The main features of the system are the FDIC technique, which allows a substantial reduction of the error rate, and the simple – and easily tunable – linear combination of two GMM engines. The FDIC shows its effectiveness even when the intersession variation subspace is trained with mixed telephone and microphone data. In the tests on the summed channel condition we have shown the effect of performing detection on both sides of a conversation, giving evidence that detecting a speaker on the 2 sides of a conversation has the same or greater relevance on the system accuracy as automatic segmentation. Finally, our unsupervised adaptation experiments highlight interesting perspectives that require further investigation.

7. References

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

[2] D. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in Proc. ICASSP 2003, pp. II–53–6, 2003.

[3] 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.

[4] S. Lucey, and T. Chen, "Improved Speaker Verification Through Probabilistic Subspace Adaptation," in Proc. EUROSPEECH 2003, pp. 2021-2024.

[5] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP Estimators for Speaker Recognition", Proc. EUROSPEECH-2003, pp. 2964-2967, 2003.

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

[7] R. Vogt, B. Baker and S. Sridharan, "Modelling Session Variability in Text-independent Speaker Verification," in Proc. INTERSPEECH 2005, pp. 3117-3120, 2005.

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

[9] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso and P. Laface, "Channel Factors Compensation in Model and Feature Domain for Speaker Recognition," in Proc. IEEE Odyssey 2006, San Juan, Puerto Rico, June 2006.

[10] P. Kenny, W. Gupta, G. Boulianne, P. Ouellet, and P. Dumouchel, "Feature Normalization Using Smoothed Mixture Transformations," in Proc. Interspeech-2006, pp. 25-28, 2006.

[11] National Institute of Standards and Technology, "NIST speech group website," http://www.nist.gov/speech/tests/spk/index.htm

[12] 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.

[13] P. Kenny, P. Dumouchel, "Disentangling Speaker and Channel Effects in Speaker Verification," in Proc. ICASSP2004, pp. I-37-40.

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

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

[16] E. Dalmasso, P. Laface, D. Colibro, C. Vair, "Unsupervised segmentation and verification of multi-speaker conversational speech", in Proc. INTERSPEECH 2005, pp. 3053-3056.

[17] E. Singer, D. A. Reynolds, "Analysis of multitarget detection for speaker and language recognition", in Proc. Odyssey 2004, pp. 301-308.