14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPEAKER RECOGNITION USING CHANNEL FACTORS FEATURE COMPENSATION

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

Loquendo, Torino , Italy*
{Daniele.Colibro,Claudio.Vair}@loquendo.com
Politecnico di Torino, Italy^
{Fabio.Castaldo,Emanuele.Dalmasso,Pietro.Laface}@polito.it

ABSTRACT

The variability of the channel and environment is one of the most important factors affecting the performance of text-independent speaker verification systems.

The best techniques for channel compensation are model based. Most of them have been proposed for Gaussian Mixture Models, while in the feature domain typically blind channel compensation is performed.

The aim of this work is to explore techniques that allow more accurate channel compensation in the domain of the features. Compensating the features rather than the models has the advantage that the transformed parameters can be used with models of different nature and complexity, and also for different tasks.

In this paper we evaluate the effects of the compensation of the channel variability obtained by means of the channel factors approach. In particular, we compare channel variability modeling in the usual Gaussian Mixture model domain, and our proposed feature domain compensation technique. We show that the two approaches lead to similar results on the NIST 2005 Speaker Recognition Evaluation data.

Moreover, the quality of the transformed features is also assessed in the Support Vector Machines framework for speaker recognition on the same data, and in preliminary experiments on Language Identification.

1. INTRODUCTION

In speaker recognition, errors are due not only to the similarity among speaker voiceprints, but also to the intrinsic variability of different utterances of the same speaker. Moreover, performance is heavily affected when a model, trained in a set of conditions, is used to test speaker data collected from different microphones, channels, and environments. In this paper we will refer to all these mismatching conditions as intersession variability or simply as channel variability.

Several proposals have been made to contrast these effects by means of feature transformations [1] [2]. Since some feature based transformations, such as feature warping [1], do not rely on a specific model, they can be used as an additional front-end processing step for any recognition system that takes advantage of this compensation technique. However, this blind feature normalization does not exploit a priori knowledge of the condition as in [2], or other information that can be obtained by a more detailed analysis of the variations of the speaker parameters in the acoustic space.

Feature mapping [2] uses 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 requires labeled training data that identify the conditions that one wants to compensate.

Thus, model-based techniques have been recently proposed that are able to compensate speaker and channel variations without requiring the explicit identification and labeling of different conditions. These techniques share a common background: modeling the variability of speaker utterances constraining them to a low dimensional space. This approach has proved to be effective for speaker adaptation both in speech recognition [3] and speaker verification [4], and for channel compensation, in speaker recognition [5] [6]. All these methods are generative and use MAP adapted Gaussian Mixture Models (GMM) [7] for modeling the speakers.

In this work we mainly refer to [6] for intersession compensation in the model domain. We present our modifications to this method, comparing the obtained results on the NIST 2005 Speaker Recognition Evaluation data (SRE-05) [8] and showing that our approach leads to similar results with a reduced computation cost.

The main objective of this work, however, has been to find a solution allowing compensating the observation features rather than the Gaussian means.

Compensating features rather than models has the advantage that the transformed parameters can be used as observation vectors for classifiers of different nature and complexity, and also for different tasks such as language or speech recognition.

The paper is organized as follows: the model based channel factors adaptation approach and our modifications are described in Section 2, together with our proposed channel factors feature adaptation technique. Section 3 summarizes the parameters of our baseline GMM systems. The experimental results, including the use of the compensated feature with a SVM classifier, are presented in Section 4. Some concluding remarks are given in Section 5.

2. CHANNEL FACTORS ADAPTATION

Gaussian Mixture Models (GMMs) used in combination with Maximum A Posteriori (MAP) adaptation [7] 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 speaker is, thus, represented by the set of the adapted mean vectors of all the Gaussians of the UBM.

A *supervector* that includes all the speaker specific parameters can be obtained simply appending the adapted mean value of all the Gaussians in a single stream. The same can be done for the UBM, obtaining the UBM supervector.

When some kind of mismatch affects the input speech, all the speaker supervector parameters are possibly modified. The idea behind the methods proposed in this paper 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* [9].

2.1 Model-domain adaptation

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

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

where $\boldsymbol{\mu}^{(i,k)}$ and $\boldsymbol{\mu}^{(k)}$ are the adapted and the original supervector of GMM *k* respectively. **U** is a low rank matrix projecting the channel factors 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*.

The approach that we use is similar to the formulation in [6] with the difference that we do not perform channel compensation during training but apply (1) only at testing time. The $\boldsymbol{\mu}^{(k)}$ supervectors are obtained by the classical MAP speaker adaptation, without any additional computation. The verification score is obtained computing the log-likelihood ratio of the test utterance using compensated speaker and UBM means.

Since the vector $\mathbf{x}^{(i,k)}$ should account for the distortions produced in the supervector space by the intersession variability, we would expect that $\mathbf{x}^{(i,k)}$ depends on the utterance *i*, but only weakly on the speaker model *k*.

To verify this hypothesis we run several tests estimating the parameters of **x** using the UBM**,** i.e. dropping the dependence on the GMM *k*. This is equivalent to apply the normalization:

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

for all the models *k* that must be scored against utterance *i*. As reported in Session 4.1, the obtained results were almost equivalent to the ones obtained with the speaker-model dependent estimation of (1), but with relevant saving of computation time, in particular when T-Norm score normalization [10] is applied.

2.1.1 Training of the channel factors subspace

The channel factors subspace, modeled by the low rank matrix **U**, is assumed to represent the distortion due to the intersession variability. 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 a database has been set up including a large number of speakers, each one with multiple recordings collected from different calls and channels.

An EM training algorithm has been used to compute the **U** matrix [5]. The number of columns *N* of the matrix **U** defines the channel subspace dimension and it is typically less than 50.

2.1.2 Estimation of the channel factors parameters

To perform channel adaptation through equation (1) or (2), the channel factors vector **x** must be estimated for each test utterance.

A maximum likelihood solution to this problem has been proposed in [3] for speaker adaptation. 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 one vector x(i,k) for each tuple {test utterance i, model k} in equation (1), or a single vector x(i) for a test utterance i in equation (2).

2.2 Feature-domain adaptation

The feature domain method that we propose allows exploiting the benefits of the channel factors adaptation, mapping the compensation supervector on the acoustic features.

We rely on the hypotheses that led to equation (2): we assume that the acoustic space distortion, characterized by the vector $\mathbf{x}^{(i)}$, can be estimated using the UBM rather than the speaker dependent model GMM *k*. Neglecting, for the sake of conciseness, the model index *k*, we rewrite (2) for each Gaussian component *m* of the supervector as:

$$\boldsymbol{\mu}_m^{(i)} = \boldsymbol{\mu}_m + \mathbf{U}_m\mathbf{x}^{(i)} \qquad \forall m \qquad (3)$$

where of $\boldsymbol{\mu}_m{}^{(i)}$, $\boldsymbol{\mu}_m$ and $\mathbf{U}_m$ refers to the *m-th* Gaussian of the GMM. The number of rows of the mean vectors and of the subspace matrix $\mathbf{U}_m$, is equal to the dimension of the input feature vector.

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

$$\hat{\mathbf{O}}^{(i)}(t) = \mathbf{O}^{(i)}(t) - \sum_m \gamma_m(t)\,\mathbf{U}_m \mathbf{x}^{(i)} \tag{4}$$

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 (4) is limited, for the sake of efficiency, to the first best contributes only. The experiments have been performed using the first 5 best contributions. Only negligible improvement of performance has been observed increasing the number of best contributions.

Equation (4) allows obtaining adapted feature vectors suitable as front-end parameters to any further classification process.

3. SYSTEM DESCRIPTION

A classical GMM have been used in this work for the development of the channel factors compensation approach. The system uses 13 Mel Frequency Cepstral Coefficients (MFCC). Feature warping to a Gaussian distribution is then performed, for each static parameter stream, on a 3 sec sliding window excluding silence frames [1]. 24 parameters per frame are obtained discarding the C_0 cepstral parameters and computing the usual delta parameters on a symmetric 5 frame window.

The GMM system is characterized by a set of 512 mixtures. A gender independent UBM has been trained using 20 hours of speech of 10 different languages using corpora not specifically collected for speaker recognition evaluations, mainly coming from the SpeechDat corpora. The dimension of the channel subspace, equal to the number of columns of the channel subspace matrix **U,** has been set to 20 for all the experiments.

4. EXPERIMENTS

The speaker recognition methods described in this paper were evaluated on the NIST 2005 Speaker Recognition Evaluation data (SRE05) [8]. All tests are related to the core test condition, as defined by NIST, including all trials in the enrollment and verification lists (2771 true speaker and 28472 impostor trials).

The evaluation has been carried out with and without score normalization. First the raw score are speaker-normalized by means of Z-norm. The Z-norm parameters for each speaker model have been evaluated using a subset of speaker samples included in the NIST SRE04 database [8]. Separate statistics have been collected for the female and male speakers, using 2 audio samples of 80 speakers for each gender.

Test dependent normalization is performed using T-norm [10]. A fixed set of impostor models have been selected among the voiceprints enrolled with data belonging to the SRE04 evaluation. The T-norm parameters for each test sample were estimated using the Z-normalized scores of the impostor voiceprints. We refer to the Z-Norm followed by T-Norm as ZT-Norm.

The performance of the systems proposed in this paper was evaluated in terms of Equal Error Rate (EER) and minimum normalized Detection Cost Function (DCF) (as defined by NIST [8]).

Table 1 gives the results scores obtained with and without ZT-Norm score normalization on the GMM baseline system.

System	EER	DCF
GMM raw	13.8	0.548
GMM ZT-Norm	10.7	0.404

Table 1 - EER and minimum DCF for GMM baseline system, with and without score normalization

4.1 UBM channel factors compensation

The channel factors were computed on the UBM and kept fixed for all the speaker models verified against a given speaker utterance. Table 2 shows the results of the GMM system with and without compensation, applying the UBM channel factors both in model (MD) and in feature (FD) domain respectively.

System – UBM compensation	EER	DCF
GMM MD raw	9.48	0.348
GMM FD raw	9.16	0.357
GMM MD ZT-Norm	8.07	0.280
GMM FD ZT-Norm	6.80	0.241

Table 2 - EER and minimum DCF with UBM channel factors compensation, in model (MD) and feature (FD) domain

The effectiveness of the channel factors compensation is significant both on the raw and ZT-normed scores. Moreover, better performance is obtained by the feature domain UBM compensation. This can be probably ascribed to the fact that in feature domain the same adaptation is performed both in enrollment and in verification. In the model domain, instead, channel compensation was performed only in testing, while the models were trained using the conventional MAP adaptation, because no improvement was obtained including the channel factors compensation in training (see next subsection).

4.2 Speaker-dependent channel factor compensation

Speaker dependent channel factors compensation was tested in the model domain.

Using the GMM system, we compared the results obtained by means of standard MAP training and channel factors compensated MAP training similar to [6]. During recognition, speaker dependent, channel factor adaptation is performed.

Somewhat surprisingly, our experimental results show that without score normalization, standard MAP training outperforms channel factors MAP. It is worth noting that the raw scores are typically affected by the lack of homogeneity among the speaker models, this is particularly true for channel factor compensated MAP. More significant is the comparison of the normalized scores. Since using ZT-Norm scores the two techniques give similar performance, the computation requirements of the channel factor compensated MAP don't seem to justify its use.

System – SD compensation	EER	DCF
GMM TrMAP raw	8.72	0.333
GMM TrCFM raw	11.87	0.493
GMM TrMAP ZT-Norm	7.02	0.240
GMM TrCFM ZT-Norm	7.49	0.244

Table 3 - EER and min DCF with speaker dependent (SD), model domain compensation. Training MAP (TrMAP) and channel factors compensated MAP (TrCFM)

4.3 SVM channel compensation

Discriminative SVM models of speaker recognition are attractive because they are trained to minimize the errors. Moreover they are typically smaller than the generative models trained with the same amount of data and require less computational resources both in training and testing.

Our work draws on the results of the generalized linear discriminant sequential (GLDS) kernel approach of [11]. However, since for computational reasons the autocorrelation matrix R in [11] is usually approximated by its diagonal elements, it turns out that it is possible to feed a SVM that uses a linear inner-product kernel, with polynomial features where each component is properly normalized by its standard deviation.

For SVM model space channel compensation, an original approach has been proposed in [12]. It evaluates the projection of the expanded vectors in a subspace that remove the dimensions that carry information not related to the speaker but only to the channel and the environment. We didn't follow this approach mainly because it relies on a discrete number of models of known conditions. We used, instead, the channel compensated features as observation vectors for the SVM classifiers. In particular, the channel factors $\mathbf{x}^{(i)}$ are estimated for each test or training utterance i (including the ones related to the set of impostors).

Using $\mathbf{x}^{(i)}$, every frame of that utterance is channel compensated according to (4). A polynomial expansion of the third order is then performed, and the mean and variance of every component of all the expanded vectors are evaluated. The expanded vector of an utterance – variance-normalized – is the channel compensated pattern for the SVM classifiers.

The observation vectors for the SVM classifiers are the same 24 parameters of the GMM system, and their expansion up to the third order polynomial.

The gender independent impostor set necessary to train these discriminative models includes the utterances of 1619 speakers obtained from the train splits of the NIST SRE-2000 and SRE-2004 databases.

Table 4 shows the results of the SVM system. Without score normalization the SVM and the GMM system (see Table 2) have similar accuracy but the GMM system outperforms SVMs using ZT-Norm. The score normalization does not give appreciable performance improvements to the SVM system.

Although less precise than the GMM system using the same parameters, the advantage of using SVM classifiers is not only their reduced computational cost both in training and in testing, but also their ability to produce scores that tend to be intrinsically normalized. This happens because each speaker model is trained against the same set of impostors, and both the speaker and impostor utterances are channel compensated.

System	EER	DCF
SVM raw	9.41	0.369
SVM FD Comp. raw	8.79	0.318
SVM ZT-Norm	9.81	0.362
SVM FD Comp. ZT-Norm	8.65	0.299

Table 4 - EER and minimum DCF for SVM system, and channel factors compensation in feature domain

4.4 Language identification

To verify the quality of the channel compensated features in a completely different task, we perform an experiment on language identification comparing the performance of a gender independent classifier, based on SVMs, using three sets of basic features: the 24 MFCC features, their channel compensated counterparts, and the shifted-delta parameters proposed [13]. Again, the vectors were subjected to a polynomial expansion of the third order and the SVMs trained using a linear kernel.

Basic features	ERR %
1. 12 MFCC+delta	18.16
2. Channel compensated 12 MFCC+delta	9.80
3. 49 shifted-delta	7.99
4. 2. and 3. fused	5.67

Table 5 - Language Identification ERRs

From the OGI 22 Languages database, 8 languages were selected among the ones appearing also in the OGI

Multilanguage Telephone Speech: English, German, Hindi, Italian, Korean, Mandarin, Spanish, and Tamil.
For each language, the conversations were equally split into a train and test list. The impostor set for a given language was composed of the set of conversations of all the remaining languages. Segments of 30 seconds have been used for testing.

The results, in terms of EER percentage, are shown in Tab. 5. Comparing the first and second rows we see that the feature domain channel factors compensation halves the Equal Error Rate. The fusion of the two systems, shown in row 4, is obtained by a linear combination of the scores produced by the two systems.

It is worth noting that the features were compensated using the same transformation matrix **U** computed for the speaker recognition experiments. This result not only shows that the channel compensation approach in feature space can be applied to other tasks, but also that the channels subspace is fairly task and language independent.

5. CONCLUSIONS

We have shown that the feature adaptation approach proposed in this paper has the same benefits of the channel factors model domain adaptation. Moreover it can be used with other types of classifiers, like SVM or ANNs and for other tasks.

Future research will be devoted to applying this technique decoupling the model used for feature compensation from the ones used for recognition, even within the GMM framework.

The system based on SVMs is attractive from an application point of view because the produced scores are fairly well stable when there are variations in the training and test conditions. This characteristic may avoid the burdensome task of the score normalizations.

REFERENCES

[1] J. Pelecanos, and S. Sridharan, “Feature Warping for Robust Speaker Verification”, *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”, *Proc. EUROSPEECH-2003*, pp. 2021-2024, 2003.

[5] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice Modeling with Sparse Training Data”, *IEEE Trans. on Speech and Audio Processing*, Vol.13, No.3, May. 2005, pp. 345-354.

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

[7] 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.

[8] National Institute of Standards and Technology, “NIST speech group website,” http://www.nist.gov/speech, 2005.

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

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

[11] W. M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition", *Proc. ICASSP, 2002*, pp. I-164-167, 2002.

[12] A. Solomonoff., W.M. Campbell and I. Boardman, “Advances in Channel Compensation for SVM Speaker Recognition," *Proc. ICASSP 2005*, pp. I-629-632, 2005.

[13] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features”, *Proc. ICSLP 2002*, pp. 90-93, 2002.