COMPENSATION OF NUISANCE FACTORS FOR SPEAKER AND LANGUAGE RECOGNITION

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

[1] *Politecnico di Torino, Torino, Italy*
[2] *LOQUENDO, Torino, Italy*

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 blind channel compensation is usually performed.

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

In this paper, we evaluate the effects of the compensation of the intersession 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 with a reduced computation cost.

We report also the results of a system, based on the intersession compensation technique in the feature space that was among the best participants in the NIST 2006 Speaker Recognition Evaluation.

Moreover, we show how we obtained significant performance improvement in language recognition by estimating and compensating, in the feature domain, the distortions due to inter-speaker variability within the same language.

Index terms
Speaker Recognition; Language Recognition; Factor Analysis; Feature Compensation
EDICS: SPE-MULT Multilingual Recognition and Identification; SPE-SPKR Speaker Recognition and Characterization

I. INTRODUCTION

One of the main causes of relevant performance degradations in automatic speech recognition systems is the acoustic mismatch that occurs between training and test

environment. The variability of the channel and environment is one of the most important factors affecting the performance of text-independent speaker verification systems. Speaker variability is an additional nuisance factor for language recognition.

In this work, we evaluate the effects of the compensation of the nuisance factors obtained by means of the factor analysis approach. We propose an intersession compensation technique in the feature domain for speaker recognition, and we apply the same approach to the compensation of inter-speaker variations within the same language.

In speaker recognition, errors are due not only to the similarity among voiceprints of different speakers, but also to the intrinsic variability of different utterances of the same speaker. Moreover, the performance of a system is severely 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 or channel variability.

Cepstral Mean Subtraction (CMS) [1], can be used to mitigate the linear filtering effects of the transmission channel. RASTA processing [2] has been shown to improve the recognition performance in presence of convolutional distortions and additive noise. Another interesting proposal to contrast channel distortions that affect the distribution of features is feature warping to a standard normal distribution through short-time gaussianization [3,4].

These feature transformations do not rely on a specific model. However, this blind feature normalization does not exploit a priori knowledge of the condition as is done in the feature mapping approach [5], or in other approaches that exploit information obtained by a more detailed analysis of the variations of the speaker parameters in the acoustic space.

Feature mapping 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 does require labeled training data to identify the conditions that one wants to compensate. A data-driven feature mapping technique has

been proposed in [6] to deal with this drawback, but it still relies on a discrete number of conditions, and requires the corresponding models in testing.

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 [7] and speaker recognition [8,9], and for intersession compensation [10-13]. All these methods are generative, and use MAP adapted Gaussian Mixture Models (GMM) [14] for modeling the speakers.

Discriminative models based on Support Vector Machines (SVMs) are able to produce comparable results with respect to the state of the art GMM based systems and to improve the performance of speaker recognition systems that combine the scores produced by the generative and discriminative modeling approaches [15,16]. An approach to channel compensation in the model space of the SVMs has been proposed in [17]. It evaluates the projection of the expanded vectors in a subspace removing the nuisance dimensions that carry information not related to the speaker but only to the channel and to the environment. This approach, as well as the ones proposed in the model space of the GMMs, do not need the labeling of a discrete combination of conditions referring to the handset type, the transmission channel, the environment, and so forth, but only the speaker identity.

In this work we mainly refer to [13] for intersession compensation in the model domain. We present our modifications to this method using the NIST 2005 Speaker Recognition Evaluation data (SRE-05) [18] as a testbed.

The main objective of this work, however, was to find a solution for compensating the observation features rather than the Gaussian means. We will show that our approach based on feature domain compensation obtained similar or slightly better results with a reduced computation cost on the SRE-05 data. Moreover, we report the results of a

system, based on the intersession compensation technique in the feature space, which was one of the best participants in the NIST 2006 Speaker Recognition Evaluation.

Compensating features rather than models has the advantage that we can use the transformed parameters as observation vectors for classifiers of a different nature and complexity, and for different tasks such as language or speech recognition. We present an example of the use of feature domain compensation with a different classifier modeling the speakers by means of Support Vector Machines, and using as kernel features the polynomial expanded cepstral parameters. Although the intersession compensation is estimated in a GMM framework, the performance of a SVM classifier on the SRE-05, using compensated features, increases by 18% compared to the corresponding classifier trained and tested with the raw features. We show also that significant performance improvements can be obtained in language recognition by estimating and compensating the distortions due to inter-speaker variability within the same language.

The paper is organized as follows: the model based channel factors adaptation approach and our modifications are described in Section II, together with our proposed intersession factors feature adaptation technique. Section III summarizes the parameters of our baseline GMM systems. In Section IV, we present several results of speaker recognition experiments with different training and testing approaches. The experiments with a SVM classifier, including the use of the compensated features, are presented in Section V. Section VI illustrates the feature compensation approach for language recognition. Finally, in Section VII we present our concluding remarks.

II. ADAPTATION OF THE INTERSESSION FACTORS

Gaussian Mixture Models (GMMs) used in combination with Maximum A Posteriori (MAP) adaptation [14] represent the core technology of most of the state-of-the-art text-

independent speaker recognition systems. In these systems, the speaker models are estimated, by means of MAP adaptation, from a common GMM root model, the so-called world model or Universal Background Model (UBM). Usually, only mean vector adaptation is performed during model training. Thus, a speaker is represented by the set of mean vectors of all the Gaussians of the UBM, adapted using the speaker training data, and shares with the other speaker models the remaining UBM parameters.

A supervector that includes all the speaker specific parameters is simply obtained by appending the adapted mean value of all the Gaussians in a single stream. The same procedure allows the UBM supervector to be obtained. The speaker model can be seen as a point in a high dimensional space, whose coordinates are the supervector parameters. When some kind of mismatch, i.e. the use of different microphones or communication channels, affects the input speech, all the speaker supervector parameters are possibly modified.

The idea behind the methods proposed in [10-13], and in this paper, is that a small number of parameters - the channel factors [19] - in a lower dimensional subspace can summarize the distortions in the large supervector space.

A. Adaptation in model domain

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

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

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*.

In the approach proposed in [13], $\mathbf{x}^{(i,k)}$ is a function of speaker model *k*, session *i*, and session observation sequence $\mathbf{o}_1^T(i)$. During training, $\mathbf{x}^{(i,k)}$ and $\boldsymbol{\mu}^{(k)}$ are jointly estimated

using a small number of iterations of an EM algorithm. We refer to this approach as Speaker Dependent Intersession Compensated MAP estimation (ICM-Training). In testing, $\boldsymbol{\mu}^{(k)}$ is assumed to be known and fixed, and the session dependent supervector $\boldsymbol{\mu}^{(i,k)}$ is obtained by applying (1) after estimating the channel factors $\mathbf{x}^{(i,k)}$.

B. Training of the intersession subspace

The intersession 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 using utterances collected from different channels or conditions. Thus, the intersession factor subspace is computed 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 model. Then, the set of the differences among the supervectors of the same speaker is collected for all the available speakers [12]. Finally, the matrix **U** is obtained performing Principal Component Analysis (PCA) using as features these difference supervectors. The set of the supervector differences, and the dimension of the supervectors are very large, but we are interested only in the subspace spanned by the N leading eigen-transformations. Thus, the estimate of **U** can be effectively done by using an EM training algorithm [20]. The number of columns *N* of matrix **U**, which defines the channel subspace dimension, is usually less than 50.

It is worth noting that the corpus used for the estimation of the matrix **U** must differ from the one used for training the target speaker model. Otherwise, the intersession factors would be biased to that specific speaker set, the recognition performance would be over optimistic, and it would not generalize to different speaker sets.

C. Estimation of the channel factors

To perform intersession adaptation through (1), vector **x** must be estimated for each utterance. A Maximum Likelihood Eigen-Decomposition (MLED) solution to a similar problem has been proposed for speaker adaptation in [7]. In this approach, speaker adaptation factors are estimated using an EM-algorithm to maximize the probability of the session observations with respect to a subspace representing the so-called eigenvoices. A slight variant of this solution proposed in [9] for speaker verification is:

$$\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b} \tag{2}$$

where the elements of matrix **A** and vector **b** are:

$$a_{k,j} = \sum_{m=1}^{M} \left(\sum_{t=1}^{T} \gamma_m \left(o_t \right) \right) \cdot \frac{\mathbf{U}'_{k,m} \cdot \mathbf{U}_{j,m}}{\boldsymbol{\sigma}_m^2} \tag{3}$$

$$b_k = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_m \left(o_t \right) \cdot \frac{\mathbf{U}'_{k,m} \cdot \left(\mathbf{o}_t - \boldsymbol{\mu}_m \right)}{\boldsymbol{\sigma}_m^2} \tag{4}$$

In these equations, *T* is the number of observation frames, *M* is the number of Gaussians in the supervector, $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ are the mean and the diagonal covariance of the UBM respectively, and $\gamma_m \left(o_t \right)$ represents the posterior probability of Gaussian *m* at time *t* given the compete observation sequence. In addition, in [9] a technique called Probabilistic Subspace Adaptation (PSA), which uses MAP estimation of **x** has been presented. It assumes that the a priori distribution of **x** is Gaussian with zero means and diagonal covariance matrix **E** including the *N* leading eigen-values of the subspace **U**. In this approach **x** is obtained by the iterative application of:

$$\mathbf{x}_{MAP} = \left(\mathbf{A} + \mathbf{E}^{-1} \right)^{-1} \cdot \mathbf{b} \tag{5}$$

To perform intersession compensation, not only we use a different matrix **U**, but also we apply two variants to the approach in [9]. In testing, the estimation of the channel factors is performed using (3), (4), and (5), but the mean vector $\boldsymbol{\mu}_m$ in (4) refers to the speaker GMM, rather than to the UBM. In training, speaker dependent intersession compensation is more complex because it requires alternating between the estimation of **x** and MAP

reestimation of $\boldsymbol{\mu}_m$ according to a procedure illustrated in [13]. In our experiments, we initially set $\boldsymbol{\mu}_m$ to the UBM.

D. Simplified intersession compensation in model domain

The approach that we use for model domain compensation is similar to the one proposed in [13] with the difference that we do not perform channel compensation during training but we apply (1) only at testing time. The speaker supervector $\boldsymbol{\mu}^{(k)}$ is obtained by the usual MAP speaker adaptation, without any additional computation. Moreover, we perform a single estimation of $\mathbf{x}$ for the Probabilistic Subspace Adaptation through (5), rather than iterate the application of (5) and (1).

At testing time, the verification scores are computed as the log-likelihood ratio of the test utterance, using compensated speaker and UBM means. This produces good performance improvements even without any normalization of the raw scores.

As a further simplification, we tried to drop the model dependency of the channel factor vector. Since $\mathbf{x}^{(i,k)}$ should account for the distortions produced in the supervector space by the intersession variability, we expect that $\mathbf{x}^{(i,k)}$ depends on the utterance *i*, but only slightly on the speaker model *k*. To verify this hypothesis we ran several tests estimating the intersession factors $\mathbf{x}$ by using the UBM rather than the GMM *k* to compute the posterior probability $\gamma_m\left(o_t\right)$ in (3) and (4).

Thus, we set:

$$\mathbf{x}^{(i,k)} = \mathbf{x}^{(i)} \qquad \forall k \tag{6}$$

and we apply the same compensation:

$$\boldsymbol{\mu}^{(i,k)} = \boldsymbol{\mu}^{(k)} + \mathbf{U} \cdot \mathbf{x}^{(i)} \tag{7}$$

for each model *k* that must be scored against utterance *i*.

As we report in Section IV-C, the obtained results are almost equivalent to the ones obtained with speaker-model dependent estimation using (1), but with significant saving of computation time, in particular when score normalization is performed by T-Norm [21], which would require the estimation of a different $\mathbf{x}^{(i,k)}$ for every impostor model *k*. Moreover, the good results obtained with this last simplification, suggest the possibility of applying the intersection compensation directly in feature domain.

E. Adaptation in feature domain

Intersession adaptation in the model domain has proved to improve the performance of GMM speaker recognition systems. However, it is not readily applicable to GMMs with a different number of Gaussians, to other types of classifiers, like SVM or Artificial Neural Networks (ANNs), or to other tasks requiring more complex models, for example HMMs for speech recognition.

It is possible to perform feature domain intersession compensation by projecting every observation feature $\mathbf{o}^{(i)}(t)$ towards the session independent space. The method that we propose exploits the nuisance factors to map the compensation supervector $\mathbf{U} \cdot \mathbf{x}^{(i)}$ to the acoustic features. We rely on the hypotheses that led to (7), i.e. 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 (7) for each Gaussian component *m* of a supervector as:

$$\boldsymbol{\mu}_m^{(i)} = \boldsymbol{\mu}_m + \mathbf{U}_m \cdot \mathbf{x}^{(i)} \qquad \forall m \tag{8}$$

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*, $\hat{\mathbf{o}}^{(i)}(t)$, is obtained by subtracting from the observation feature a weighted sum of the intersession compensation offset values:

$$\hat{\mathbf{o}}^{(i)}(t) = \mathbf{o}^{(i)}(t) - \sum_m \gamma_m(t) \cdot \mathbf{U}_m \cdot \mathbf{x}^{(i)} \quad (9)$$

where $\gamma_m(t)$ is the Gaussian posterior probability, and $\mathbf{U}_m \cdot \mathbf{x}^{(i)}$ is the intersession compensation offset related to the *m-th* Gaussian of the UBM model. In the actual implementation, the right side summation of (9) is limited, for the sake of efficiency, to the first best contributions only (1 to 5 in our experiments). Feature domain compensation is performed both in training and in testing.

Equation (8) allows adapted feature vectors to be obtained, suitable as front-end parameters to any further classification process. We verified the quality of the transformed features using them for different classifiers and tasks as reported in Sections IV, V, and VI.

This approach, referred to in the following as Feature Domain Intersession Compensation (FDIC), has been introduced in [22]; a similar approach for feature normalization was independently developed for speech recognition in [23], with the goal of removing both speaker and channel effects from individual utterances.

III. SYSTEM DESCRIPTION

Two GMM systems have been trained and tested in this work: a Phonetic GMM (PGMM), and a classical GMM. The system based on the PGMM has been used to produce the results submitted to the SRE-05 evaluations [18]. A simple linear combination of the results of two systems - a PGMM and a GMM - was our primary system for the SRE-06 evaluation.

A. The Phonetic GMM system

The PGMM system decodes the speaker utterance, both in enrollment and in verification, producing phonetic labeled segments. The decoder is a hybrid Hidden Markov Model – Artificial Neural Network (ANN) model trained to recognize 11 language independent broad phone classes: silence, liquids, nasals, fricatives, affricates, voiced and unvoiced plosives, diphthongs, front, central, and back vowels. 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 evaluations – the Macrophone [24] for US English, and the SpeechDat(II) corpora [25] for (Dutch, French, German, Greek, Italian, Portuguese, Spanish, Swedish, and UK English).

The UBM and the voiceprints consist of a set of phonetic GMMs: each state of a phone class has an associated GMM. For each state, the maximum number of (diagonal covariance) Gaussians per mixture is 64, and the total number of Gaussians of this system is 1954. This gender– independent, and almost language–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, for each static parameter stream, on a 3 seconds sliding window excluding silence frames [3,4]. Each observation frame includes 36 parameters obtained by discarding the C_0 cepstral parameter, and computing the usual delta parameters on a symmetric window of 5 frames. Session compensation in either the model or the feature domain is performed using these features and the UBM supervector including 1954 Gaussians.

B. The GMM system

The second system is based on the classical GMM approach [14]. The GMM system is characterized by a reduced set of mixtures (512), and features (the first 12 cepstral and

their delta parameters). Again, a gender independent UBM has been trained with the same database that was employed for training the PGMM world model, but using the reduced set of features.

IV. EXPERIMENTS

The techniques described in this paper have been evaluated on the NIST SRE-05 data, and then applied to the SRE-06 evaluation. All the experiments refer to the NIST defined core test condition, including all trials in the enrollment and verification lists. There are 2771 true speaker and 28472 impostor trials for the SRE-05 evaluation set, whereas the SRE-06 evaluation set includes 3612 true speaker and 47836 impostor trials. The core test condition consists of four wire conversations lasting approximately five minutes. Each side of a conversation is recorded on a separate channel, and silences are not removed from the recordings.

The evaluation has been performed with and without score normalization. First, the raw scores are speaker-normalized by means of Z-norm [21]. The Z-norm parameters for each speaker model have been estimated using a subset of speaker samples included in the NIST SRE-04 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 [21]. A fixed set of impostor models was selected among the voiceprints enrolled with data belonging to the SRE-04 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 was measured in terms of Equal Error Rate (EER) and minimum normalized Detection Cost Function (DCF) as defined by NIST [18]. The Equal Error Rate (EER) is the error of the speaker recognition system when the detection threshold is set so that the probability of False Alarms equals the Miss Detection

probability. The DCF is the main performance measure in NIST evaluations; it is a weighted sum of the Miss Detection and False Alarm rates.

$$DCF = (C_{Miss} * P_{Miss|Target} * P_{Target}) + C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * (1 - P_{Target}) \quad (10)$$

where the evaluation parameters, $C_{Miss} = 10$, $P_{Target} = 1$, and $C_{FalseAlarm} = 0.01$ have been selected as reasonable values for a possible application. A normalized DCF is actually used, i.e. the DCF cost is normalized by the best cost that could be obtained without processing the input data.

Comparative results are also illustrated by means of the Detection Error Tradeoff curves (DET) [26]. In the DET plots, the False Alarm and Miss Detection probabilities are plotted using a normal deviate scale for each axis. A DET curve shows the performance of the system on the test set for all the operating points, i.e. for different detection thresholds.

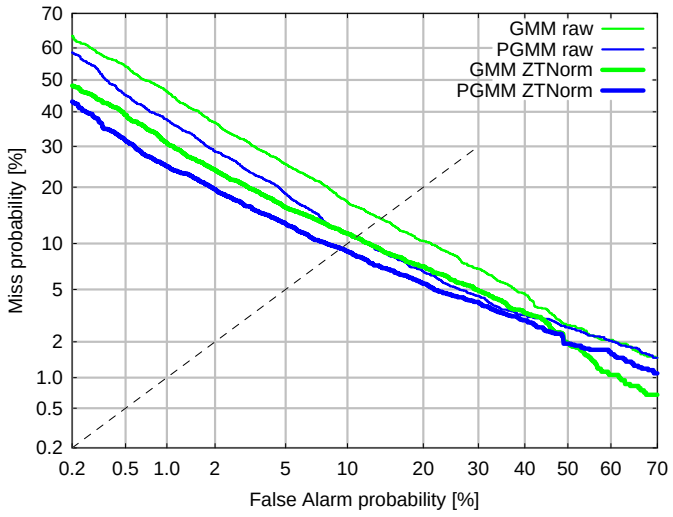

Fig. 1. DET plots for GMM and PGMM baseline systems, with and without score normalization, on NIST SRE-05 core test data.

Table I
EER (%) AND MINIMUM DCF FOR GMM AND PGMM BASELINE SYSTEMS, WITH AND WITHOUT SCORE NORMALIZATION, ON NIST SRE-05 CORE TEST DATA

System	Normalization	Train	Compensation	EER	DCF
GMM	No	MAP	No	13.8	0.548
PGMM	No	MAP	No	10.8	0.468

GMM	ZT-Norm	MAP	No	10.7	0.404
PGMM	ZT-Norm	MAP	No	9.2	0.343

A. GMM versus PGMM

Figure 1 shows the DET plots of the GMM baseline systems. The figure includes the results of the standard (GMM) and of the Phonetic (PGMM) Gaussian Mixture recognizers, obtained without and with ZT-Norm score normalization. The related scores, in terms of Equal Error Rate and minimum normalized DCF are given in Table I.

The difference in accuracy of the two systems depends on the several diversities they present. The number of acoustic features is 36 for the PGMM and 24 for the GMM, and the number of Gaussians is 1954 versus 512 respectively. Moreover, the two systems use different procedures for removing silences. The PGMM exploits the silence classification of the ANN. In the GMM system, instead, speech-silence discrimination is performed collecting the histogram of the energy values in the utterance, fitting two Gaussians, and using them for the classification.

All the results of the experiments that will be presented in the following sections have been obtained with ZT-normalized scores.

B. Simplified intersession compensation in model domain

In this section, we show that our modifications to standard training lead to similar results with a reduced computation costs. We compare the results of the simplified approach with the ones obtained with an Intersession Compensated MAP (ICM) training approach similar to the one presented in [13].

In the simplified compensation, as said in section II-D, $\mathbf{\mu}^{(k)}$ and $\mathbf{x}^{(i,k)}$ are not jointly estimated in training. Instead, the supervector $\mathbf{\mu}^{(k)}$ is trained by classical MAP, without any additional computation. Moreover, in our experiments, we perform a single iteration of the PSA estimation, obtaining one vector $\mathbf{x}^{(i,k)}$ for each tuple {test utterance *i*, speaker

model $k\}$ in (1). Intersession compensation is performed in testing only. We refer to this approach as MAP training and Model Domain, Speaker Dependent compensation.

The intersession subspace has been estimated using all the data in the NIST SRE-04 database. It includes 308 speakers and ~8 sessions per speaker. We computed all the differences among the supervectors of the same speaker from different sessions. From these 10315 vectors we derived an intersession subspace with dimension 20.

Figure 2 shows the DET plots corresponding to the standard and to the simplified approach. With ZT-Norm score normalization, the two techniques perform similarly. The computation requirements of the Intersession Compensated MAP approach do not seem to justify its use.

The simplified procedure was also effective even for larger amounts of training data, such as the ones provided by the 3 or 8 conversation sides of other SRE-05 trials.

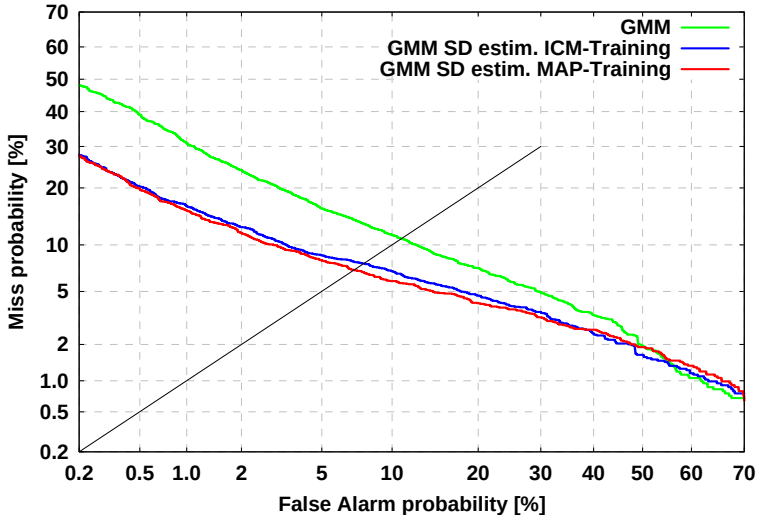

Fig. 2. DET plots for the GMM system with speaker dependent (SD) channel factors compensation, in model domain. Training MAP compared with Intersession Compensated MAP (ICM) on NIST SRE-05 core test data.

TABLE II
EER (%) AND MINIMUM DCF FOR THE GMM AND THE PGMM SYSTEMS WITH SPEAKER DEPENDENT (SD) MODEL DOMAIN COMPENSATION. TRAINING MAP COMPARED WITH INTERSESSION COMPENSATED MAP (ICM) ON NIST CORE TEST SRE-05 DATA

System	Train	Compensation	Estimation of **x**	EER	DCF
GMM	MAP	No	No	10.7	0.404
GMM	ICM	Model Domain	Speaker Dep.	7.49	0.244

GMM	MAP	Model Domain	Speaker Dep.	7.02	0.240
PGMM	MAP	Model Domain	Speaker Dep.	6.82	0.231

Similar results were obtained using the PGMM system. Table II summarizes the scores of the GMM and PGMM systems with intersession compensation in model domain. The PGMM system with MAP training and intersession compensation is only slightly better than the equivalent GMM system, reducing the appreciable gap on the uncompensated systems.

C. Feature domain intersession compensation

In this section, we show the improvement obtained with the Feature Domain Intersession Compensation (FDIC), applied to GMM and PGMM systems.

The intersession factors were computed using the UBM and kept fixed for all the speaker models verified against a given test trial, according to (7). Figure 3 shows the DET curves of the baseline GMM system without compensation, compared with the GMMs obtained by applying the UBM intersession factors both in model domain (MD) and in feature domain (FD).

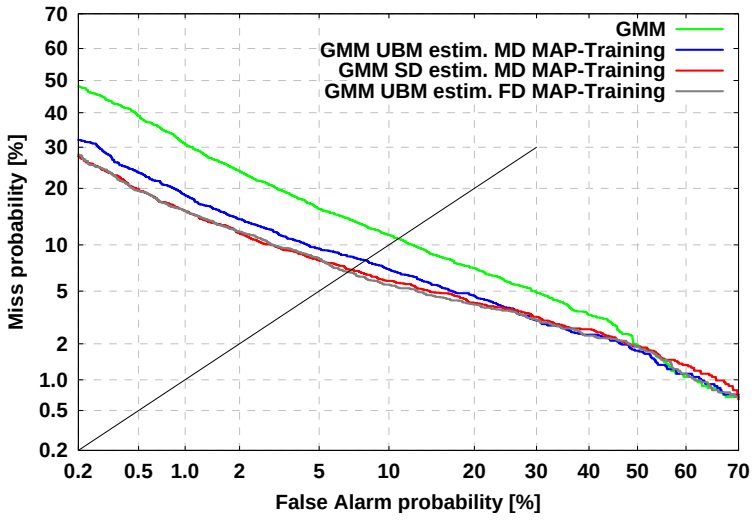

Fig. 3. DET plots for the GMM system with Speaker Dependent (SD) or UBM based compensation, in model (MD) and in feature (FD) domain, on NIST SRE-05 core test data.

Table III
EER (%) AND MINIMUM DCF FOR THE GMM SYSTEM WITH SPEAKER DEPENDENT OR UBM BASED COMPENSATION, IN MODEL AND IN FEATURE DOMAIN, ON NIST SRE-05 CORE TEST DATA

System	Train	Compensation	Estimation of $\mathbf{x}$	EER	DCF
GMM	MAP	No	No	10.7	0.404
GMM	MAP	Model Domain	UBM	8.07	0.280
GMM	MAP	Model Domain	Speaker Dep.	7.02	0.240
GMM	MAP	Feature Domain	UBM	6.80	0.241
PGMM	MAP	Model Domain	Speaker Dep.	6.82	0.231
PGMM	MAP	Feature Domain	UBM	6.79	0.231

Comparing the second and third rows of Table III it can be observed that performing the intersession compensation in model domain, the Speaker Dependent estimation of $\mathbf{x}^{(i,k)}$ is more expensive but slightly better than the UBM based Speaker Independent estimation of $\mathbf{x}^{(i)}$. However, as shown in the fourth row of Table III, feature domain compensation, estimated on the UBM, recovers this gap and achieves the same performance as speaker dependent model domain compensation. An explanation for this behavior is that in feature domain the same adaptation is performed both in enrollment and in verification. In the model domain, instead, intersession compensation is performed only in testing, while the models are trained using the conventional MAP adaptation, because no improvement was obtained by including channel factors compensation in training as reported in Table II.

Similar results were obtained using the PGMM system. It is worth noting, however, that the GMM and the PGMM give complementary results. The simple linear combination, with equal weights, of the results of the GMM and PGMM systems with feature domain compensation shown in Table III, gives a significant performance improvement both of the EER (5.94% versus 6.79%) and of the DCF (0.202 versus 0.231). This is due to the differences between the two systems that have been listed in section IV-A.

The excellent performance of our feature domain compensation technique has been confirmed in the 2006 NIST evaluation. The results shown in Figure 4 and Table IV refer again to the core test – all trials. The first result in Table IV, labeled PGMM-2005 also in Figure 4, was obtained with our "mothballed" system used for the NIST 2005 evaluation. This system did not include intersession compensation. The use of "mothballed" systems is encouraged by NIST to have a comparison of the performance improvement achieved in successive evaluations. As shown in the next two rows of Table IV, the feature

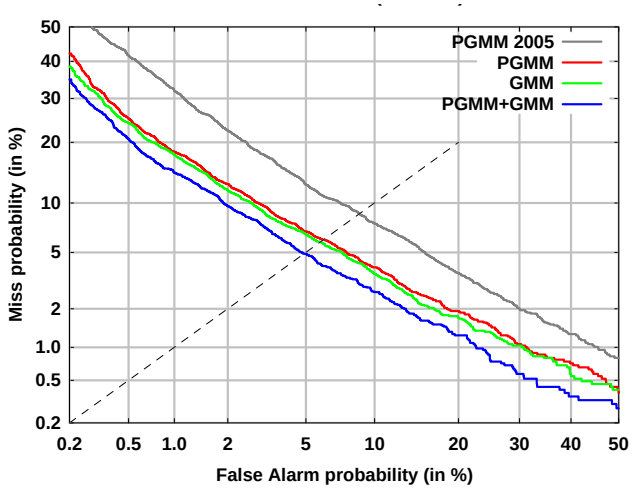

Fig. 4. DET plots for the PGMM and GMM systems using FDIC on NIST SRE-06 core test evaluation.

Table IV
PERFORMANCE OF THE PGMM AND GMM SYSTEMS USING FDIC ON NIST SRE-06 CORE TEST EVALUATION

System	Compensation	EER (%)	DCF
PGMM-2005	No	8.72	0.406
PGMM	FDIC	6.06	0.277
GMM	FDIC	5.90	0.271
PGMM+GMM	FDIC	4.96	0.236

compensated GMM and PGMM systems produce almost equivalent results, with a significant 30% reduction of the Equal Error Rate and of the Detection Cost Function.

Applying again a simple linear combination, with equal weights, of the results of the PGMM and GMM systems we have a further 10% reduction of the ERR and of the DCF.

The results of the fusion of these two systems were the ones we submitted for the 15 combinations of train and test conditions defined in the SRE-06 evaluation plan [18]. In all conditions, our system was among the best participants in this evaluation.

V. Channel compensation using SVM classifiers

The channel compensated features of (9) can be readily used as observation vectors for SVM classifiers. Our work draws on the results of the generalized linear discriminant sequential (GLDS) kernel approach of [15]. However, since for computational reasons the autocorrelation matrix **R** in [15] is usually approximated by its diagonal elements, it is possible to feed the SVM with polynomial features where each component is properly normalized by its standard deviation.

In particular, the channel factors $\mathbf{x}^{(i)}$ are estimated for each test or training utterance *i*, of target or impostor speakers. Using $\mathbf{x}^{(i)}$, every frame of the utterance is channel compensated according to (9). 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 mean vector of an utterance – variance-normalized – is the channel compensated pattern for the SVM classifiers.

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

TABLE V
EER (%) AND MINIMUM DCF FOR A GLDS SVM CLASSIFIER WITH FDIC, ON NIST SRE-05 CORE TEST DATA

SYSTEM	EER	DCF
SVM	9.81	0.362
SVM FD COMP	8.65	0.299
SVM + PGMM FD COMP.	6.18	0.211

The gender independent impostor set required to train the SVM models includes the utterances of 1619 speakers obtained from the train splits of the NIST SRE-00 and SRE-04 databases.

Table V shows the EER and minimum DCF rates of the SVM system with and without feature compensation on the same evaluation data used for the PGMM and GMM tests. Although the intersession factors are estimated in the framework of the GMMs, the feature domain compensation reduces the EER and the minimum DCF, by 12% and 18% respectively. The linear combination of the SVM and PGMM systems improves the performance of the latter by 9%. However, since our SVM approach, either alone or in combination with the PGMM system, was less accurate than the GMM system using the same parameters, we did not employ the SVM system for the SRE-06 evaluation.

VI. Nuisance compensation in language recognition

To verify the quality of the nuisance compensated features in a different task, we performed a set of experiments on language recognition.

All the experiments have been performed on the NIST 1996, 2003, and 2005 Language Recognition Evaluation (LRE) data according to NIST evaluation rules [27]. The first two test corpora include 12 target languages: American English, Arabic, Canadian French, Farsi, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese. Russian has been used as the out-of-language in the 2003 tests. In these evaluations there are three duration settings: 3, 10, and 30 seconds. The 1996 evaluation database consists of 1503, 1501 and 1492 sessions of 3, 10, and 30 seconds, respectively. The 2003 evaluation has 1280 trials for each duration setting.

The LRE-05 corpus includes seven languages and two dialects: English-American, English-Indian, Hindi, Japanese, Korean, Mandarin-Mainland, Mandarin-Taiwan, Spanish, and Tamil. The evaluation data consists of 3662 trials for each duration setting.

In all the experiments we used the Shifted Delta Cepstral (SDC) features [28]. The SDC coefficients are computed, for a cepstral frame at time *t*, according to:

$$\begin{gathered}\Delta c_n(t,i) = c_n(t+iP+d) - c_n(t+iP-d) \\ n = 0, N-1 \qquad i = 0, k-1\end{gathered} \tag{11}$$

where *n* is the *n-th* MFCC, *d* is the lag of the delta coefficients, *P* is the distance between successive delta computations, and $i = 0, k-1$, is the SDC block number. The final feature vector is the concatenation of *k* blocks of *N* parameters. The configuration 7-1-3-7 for *N-d-P-k* is normally used for language recognition. We append these features to the first 7 static cepstral coefficients obtaining a 56 parameter observation vector.

A. Language recognition with channel compensated GMMs

In the first set of experiments, we compared the performance of a gender dependent GMM classifier [14] using either the original or the channel compensated features. The UBM and the language GMMs consist of mixtures of 512 Gaussians.

To compensate channel distortions, we reused the same list of files that was employed for estimating the intersession subspace in the speaker recognition experiments. The $\mathbf{U}$ matrix, although computed on a different set of parameters - 7 cepstral plus 49 SDC, rather than 12 cepstral parameters plus their derivatives - represents the same channel subspace, which can be considered task independent.

We then created a gender-dependent model for each of the 12 target languages in the NIST corpora using the training and development sets of the CallFriend corpus [24] including a total of 1174 speakers. The number of factors used for frame compensation was fixed to 20.

During testing, the UBM gender model that produces the best likelihood for the current utterance is selected, together with the set of its corresponding gender-dependent GMM language models. The final score for each language includes both the T-normalization, performed on the *L*-1 alternative language GMMs, and the log-likelihood normalization [29]:

$$\tilde{s}_l = \log\left(\frac{1}{L-1}\cdot\frac{e^{s_l}}{\sum_{k\neq l} e^{s_l}}\right) \tag{12}$$

where $l = 1, \cdots, L$ and s_l are the index and the log-likelihood score of the $l-th$ language GMM respectively.

Comparing the results, shown in the first two rows of Table VI, obtained with the raw and channel compensated models respectively, it is clear that channel compensated features provide significant performance improvement, increasing with the duration of the utterances, up to 57.6% for the 30 seconds trials of the 1996 evaluation.

This result not only shows that the channel compensation approach in feature space can be effectively applied to other tasks modeled within the GMM framework, but also that the channels subspace is fairly task and language independent.

B. Language recognition with speaker compensated GMMs

For improving language recognition, however, we are interested in compensating not only the channel nuisances, but also the inter-speaker variations within the same language. Thus, we estimate another inter-speaker subspace matrix $\mathbf{U}_S$ with a large set of differences between models generated using different speaker utterances of the same language. In particular, we trained a gender independent $\mathbf{U}_S$ matrix reusing the same CallFriend database of 1174 speakers employed for training the two gender dependent UBMs. The differences between the supervectors were computed among speakers of the same gender, but matrix $\mathbf{U}_S$ was estimated by pooling all the difference supervectors (limited to 25000) in the same set. Since there are few different sessions for the same

TABLE VI
EER (%) AND RELATIVE IMPROVEMENT (%), IN PARENTHESES,
FOR A GMM CLASSIFIER WITHOUT AND WITH NUISANCE COMPENSATION
ON THE NIST LRE TASKS

CORPUS	1996			2003			2005		
NUISANCE COMPENSATION	DURATION			DURATION			DURATION		
	3s	10s	30s	3s	10s	30s	3s	10s	30s
NO	17.34	9.00	5.10	19.33	11.25	7.02	22.05	14.87	11.57
CHANNEL	15.94	6.13	2.16	17.50	8.41	4.08	21.72	13.91	9.64

	(8.1)	(31.9)	(57.6)	(9.5)	(25.2)	(41.9)	(1.5)	(6.5)	(16.7)
SPEAKER	15.35	5.47	2.01	16.75	7.85	3.60	21.25	12.60	8.19
	(11.5)	(39.2)	(60.6)	(13.3)	(30.2)	(48.7)	(3.6)	(15.3)	(29.2)

speaker, the main compensation is inter-speaker, but it possibly includes session variations.

Compensating the features for the inter-speaker variations, we obtained a 30% to 60% reduction of the EER in the 30s duration tests, depending on the database, as shown in the last row of Table VI.

C. Language recognition with speaker compensated GMM-SVMs

Finally, we used speaker compensation in a discriminative language classifier approach that has been proposed in [30] for speaker recognition, rather than for language recognition. We adapted, from a gender independent UBM, a GMM for each utterance in every language. The adapted mean values of these GMMs, properly normalized, can be used as the supervector kernels for SVM based language recognition. This approach is referred to in the following as GMM-SVM.

The results of this approach are shown in the first row of Table VII. Compared with the corresponding performance of the GMM classifier (last row of Table VI) it can be noticed that the GMM-SVM system obtains far better results for the tests of 30s duration. For shorter durations, however, the estimation of the utterance GMMs is not robust enough, due to the lack of data compared with the number of parameters of the GMMs. Thus, the GMM system gives better results in these conditions.

D. Language recognition with phonetic models

The systems presented so far exploit acoustic features only. Since the combination of acoustic based systems with phonetic systems produced excellent performances in the last formal NIST evaluations [31,32], this section briefly describes the contribution to language recognition of our phonetic system. The system we use is based on the Parallel

Phone tokenizer-SVM recognition approach that has been proposed in [33], again for speaker recognition. It uses phone sequences provided by multiple recognizers in different languages. We used 6 different phone recognizers for the following languages: Catalan, German, Italian, Spanish, Swedish, and UK English.

Our phone recognizer is the standard Loquendo-ASR decoder, which has been employed in combination with a phone-class loop grammar, in the first step of the PGMM speaker recognition system, presented in section III-A. Using a phone-loop grammar with diphone transition constraints, this recognizer produces either the best-decoded phone string or a phone lattice for each utterance. In these experiments only the best decoded string was used to collect the statistics of the *n*-gram phone occurrences of each utterance (*n* limited to 3).

TABLE VII
EER (%) OF THE ACOUSTIC GMM-SVM AND OF THE PHONETIC-SVM CLASSIFIERS, COMPARED WITH "STATE-OF-THE-ART" PERFORMANCE ON THE NIST LRE TASKS

Corpus	1996			2003			2005		
Model	Duration			duration			duration		
	3s	10s	30s	3s	10s	30s	3s	10s	30s
GMM-SVM	19.29	6.60	1.20	21.16	8.32	2.16	23.17	12.40	5.92
PHONETIC-SVM	14.91	4.32	0.94	18.15	6.08	1.99	20.19	10.00	5.22
Fusion	12.46	3.07	0.53	15.63	4.50	1.07	17.06	7.77	3.97
MMI-GMM	-	-	-	14.8	5.5	2.1	17.2	8.6	4.6
PHONETIC-LM	-	-	-	18.8	6.6	1.8	21.4	10.7	5.3
Fusion	-	-	-	11.8	3.0	0.8	14.1	6.4	2.9

The results of the test on the NIST evaluation data, reported in the second row of Table VII, show that the phonetic-SVM approach produces better performance than the GMM-SVM in all tests. The linear combination of the two systems with equal weight allows the overall performance to be increased in most tests by more than 20%.

The three rows in the bottom half of Table VII report the results given in [32] and [34] for the LRE-03 and LRE-05 evaluation data respectively. In these experiments, the acoustic system is based on GMMs trained by Maximum Mutual Information Estimation (MMIE), and the phonetic system exploits three phoneme recognizers and a lattice based language model. In our knowledge, these results are among the best achieved so far on these data. Although our work was not focused on language recognition, the reader may have an indication of the quality of our models by comparison with these "state of the art" results. We expect that better performance can be achieved by training larger GMM-SVM models and by exploiting the information in the phonetic lattices, rather than the best hypotheses, produced by our phonetic decoders.

VII. CONCLUSION

The main contribution of this work is the observation that the distortions produced in the supervector space by the intersession variability should depends on the current utterance, but only slightly on the speaker model. Based on this assumption, we have shown that the acoustic space distortion can be estimated using the UBM rather than the speaker dependent models without accuracy loss. This result allowed us to propose a feature domain intersession compensation approach (FDIC) that projects every observation feature towards the session independent space.

This nuisance compensation approach in feature domain gives the same benefits as the model domain approach, but with reduced computation costs. Moreover, the transformed parameters can be used as observation vectors for classifiers of a different nature, and for different tasks. This has been confirmed using the FDIC approach as a front-end for Support Vector Machine classifiers, and for compensating inter-session distortions and the inter-speaker variations in language recognition experiments.

All these experiments show that performing variability compensation as a front-end process for speaker and language recognition systems is always beneficial, even if the

compensation is estimated in a GMM framework and the new features are given as input to another classifier. In the GMM framework, it is also possible to use a model for feature compensation and another one, with a larger number of parameters, for recognition, but the best performance is achieved if variability estimation and classification is performed using the same GMM.

Two techniques based on SVMs that have been proposed for speaker recognition - the GMM-SVM and the Parallel Phone tokenizer-SVM - have been shown to produce good performance in language recognition as well, the former taking advantage of inter-speaker frame compensation.

Compared with our model, the model in [19] is richer, and the joint factor analysis estimation procedure proposed by the same authors for intersession compensation in model domain in successive papers is more elegant and general because it takes into account both speaker and channel variability. Future work will be devoted to compare the effectiveness of the two approaches.

References

[1] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” Journal of the Acoustical Society of America, Vol. 55, pp.1304-1312, 1974.

[2] H. Hermansky and N. Morgan , “RASTA processing of speech,”, IEEE Trans. Speech Audio, Vol. 2, n. 4, pp. 578-589, 1994.

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

[4] B. Xiang, U.V. Chaudhari, J. Navratil, G.N. Ramaswamy, and R.A. Gopinat, “Short-time Gaussianization for Robust Speaker Verification,” in Proc. ICASSP 2002, Vol. 1, pp. 681-684, 2002.

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

[6] M. Mason, R. Vogt, B. Baker, and S. Sridharan,”Data-driven Clustering for Blind Feature Mapping in Speaker Verification,” in Proc. Interspeech 2005, pp. 3109-3112, 2005.

[7] 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.

[8] O. Thyes, R. Kuhn, P. Nguyen, and J.C. Junqua, “Speaker Identification and Verification Using Eigenvoices,” in Proc. ICSLP 2000, pp. 242-245, 2000.

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

[10] P. Kenny, M. Mihoubi, and P. Dumouchel, “New MAP Estimators for Speaker Recognition,” in Proc. EUROSPEECH 2003, pp. 2964-2967, 2003.

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

[12] N. Brümmer, “The Spescom Data Voice NIST SRE 2004 System,” presented at NIST SRE 2004 Evaluation Workshop, Toledo, Spain, 2004.
[13] R. Vogt, B. Baker and S. Sridharan, “Modelling Session Variability in Text-independent Speaker Verification,” in Proc. INTERSPEECH 2005, pp. 3117-3120, 2005.
[14] 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.
[15] W. M. Campbell, “Generalized Linear Discriminant Sequence Kernels for Speaker Recognition,” in Proc. ICASSP 2002, pp. I-164-167, 2002.
[16] S. Kajarekar, “Four Weightings and a Fusion: a Cepstral-SVM System for Speaker Recognition,” in Proc. 2005 Automatic Speech Recognition and Understanding Workshop, pp. 17- 22 2005.
[17] A. Solomonoff., W.M. Campbell and I. Boardman, “Advances In Channel Compensation For SVM Speaker Recognition,” in Proc. ICASSP 2005, pp. I-629-632, 2005.
[18] National Institute of Standards and Technology, “NIST speech group website,” http://www.nist.gov/speech/tests/spk/index.htm, 2005.
[19] P. Kenny, P. Dumouchel, “Disentangling Speaker and Channel Effects in Speaker Verification,” in Proc. ICASSP 2004, pp. I-37-40, 2004.
[20] M. E. Tipping and C. M. Bishop, “Mixtures of Probabilistic Principal Component Analysis,” Neural Computation, vol.11, no.2, pp. 443-482, 1999.
[21] 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.
[22] 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.
[23] 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.
[24] J. Bernstein, K. Taussig, and J. Godfrey, “MACROPHONE,” available at http://www.ldc.upenn.edu
[25] Available at http://www.speechdat.org/SpeechDat.html
[26] 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.
[27] National Institute of Standards and Technology, “NIST Speech Group Website,” http://www.nist.gov/speech/tests/lang/index.htm, 2005.
[28] 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,” in Proc. ICSLP 2002, pp. 90-93, 2002.
[29] W.M. Campbell, J.R. Campbell, D.A. Reynolds, E. Singer and P.A. Torres-Carrasquillo, “Support Vector Machines for Speaker and Language Recognition”, in Computer Speech and Language, Vol. 20, pp. 210-229, 2006.
[30] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, “Support Vector Machines Using GMM Supervectors for Speaker Verification,” IEEE Signal Processing Letters, Vol. 13, no. 5, pp. 308-311, 2006.
[31] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, “Advanced Language Recognition using Cepstra and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation," IEEE Odyssey Speaker and Language Recognition Workshop, Puerto Rico, 2006.
[32] L. Burget, P. Matejka, and J. Cernocky, “Discriminative Training Techniques for Acoustic Language Identification,” in Proc. ICASSP 2006, Vol. I, pp. 209-212, 2006.
[33] W.M. Campbell, J.R.Campbell, D.A. Reynolds, D.A. Jones and T.R. Leek, “High-level Speaker Verification with Support Vector Machines”, in Proc. ICASSP 2004, Vol I, pp. 73-76, 2004.
[34] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, “Brno University of Technology System for NIST 2005 Language Recognition Evaluation,” IEEE Odyssey Speaker and Language Recognition Workshop, Puerto Rico, 2006.