Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition

Marc Ferras, Cheung-Chi Leung, Claude Barras, and Jean-Luc Gauvain, *Member, IEEE*

***Abstract*—In the last years the speaker recognition field has made extensive use of speaker adaptation techniques. Adaptation allows speaker model parameters to be estimated using less speech data than needed for maximum-likelihood (ML) training. The maximum *a posteriori* (MAP) and maximum-likelihood linear regression (MLLR) techniques have typically been used for adaptation. Recently, MAP and MLLR adaptation have been incorporated in the feature extraction stage of support vector machine (SVM)-based speaker recognition systems. Two approaches to feature extraction use a SVM to classify either the MAP-adapted Gaussian mean vector parameters (GSV-SVM) or the MLLR transform coefficients (MLLR-SVM). In this paper, we provide an experimental analysis of the GSV-SVM and MLLR-SVM approaches. We largely focus on the latter by exploring constrained and unconstrained transforms and different choices of the acoustic model. A channel-compensated front-end is used to prevent the MLLR transforms to adapt to channel components in the speech data. Additional acoustic models were trained using speaker adaptive training (SAT) to better estimate the speaker MLLR transforms. We provide results on the NIST 2005 and 2006 Speaker Recognition Evaluation (SRE) data and fusion results on the SRE 2006 data. The results show that using the compensated front-end, SAT models and multiple regression classes bring major performance improvements.**

***Index Terms*—Constrained MLLR (CMLLR), Gaussian supervectors, Gaussian mixture model (GMM), maximum-likelihood linear regression (MLLR), speaker recognition, support vector machine (SVM).**

I. Introduction

CURRENT state-of-the-art systems for text-independent speaker recognition use cepstral coefficients as base features. Although popular and successful, cepstral features are not optimal for speaker recognition tasks, since they result from the interaction of several information sources such as the message, acoustic context, channel and speaker, the latter factor exhibiting the lowest variability [1]. From this view, the speaker information seems to be buried underneath other sources of variability. Modeling the undesired variability, e.g., channel or text-dependency, to remove its harmful factors has been widely used to address this problem. Several channel and session compensation techniques, e.g., Feature mapping (FM) [2], Factor analysis (FA) [3] or nuisance attribute projection (NAP) [4] have been successfully applied and are being used in state-of-the-art systems. Session and channel mismatch have also been addressed using score normalization techniques such as T-norm or H-norm [5].

Adaptation techniques have long been used in speech recognition to improve robustness with respect to speaker variability. State-of-the-art large vocabulary continuous speech recognition (LVCSR) systems use speaker-adapted models. The goal of adaptation techniques is to turn speaker-independent models into speaker-dependent ones using much less data than would be needed for full speaker-dependent training. In speaker recognition, speaker adaptation was first used in the GMM-UBM paradigm [6], where a universal background model (UBM) is trained on data from many speakers in an attempt to model the whole set of observable speakers. The UBM is adapted to each speaker via maximum *a posteriori* (MAP) estimation [7] using the enrollment data. This allows a detailed model to be trained when little data is available, which is often the case when a large number of parameters are estimated. In recent years, eigenchannel [8] and joint factor analysis (JFA) [9], [10] MAP adaptation have given excellent results in scenarios with large inter-session variability. These techniques use more or less complex models to separate the speaker and channel variabilities during adaptation.

Recently, two other successful approaches to speaker recognition have used adaptation techniques to obtain features that are classified using support vector machines (SVMs). A first approach uses the mean vectors of a speaker-adapted GMM, obtained via MAP adaptation of a UBM, as features. A Gaussian supervector is formed by stacking all mean vectors of this model and is classified using a SVM. We refer to this approach as Gaussian supervectors or GSV-SVM [11]. In a second approach, the hidden Markov models (HMM) of an automatic speech recognition (ASR) system are adapted using maximum-likelihood linear regression (MLLR) and the transform coefficients used as features. MLLR transforms a speaker-independent model into a speaker-dependent one, capturing information that is specific to the speaker. The use of MLLR transform coefficients as features has been addressed in [12]–[14] and, when classified using a SVM, it is referred

Manuscript received November 10, 2008; revised June 13, 2009. First published October 09, 2009; current version published July 14, 2010. This work was supported in part by OSEO under the Quaero program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

M. Ferras, C. Barras, and J.-L. Gauvain are with LIMSI-CNRS, 91403 Orsay, France (e-mail: ferras@limsi.fr; barras@limsi.fr; gauvain@limsi.fr).

C.-C. Leung is with the Human Language Technology Department, Institute for Infocomm Research (I2R), Singapore 138632 (e-mail: ccleung@i2r.a-star.edu.sg).

Digital Object Identifier 10.1109/TASL.2009.2034187

to as MLLR-SVM. A purely acoustic variant using constrained MLLR (CMLLR) and a universal background model (UBM) in a speaker adaptive training (SAT) [15] framework has been presented in [16].

This paper presents an in-depth exploration of MAP and MLLR adaptation in the context of GSV-SVM and MLLR-SVM systems. Given the relevance of session compensation in speaker recognition, two widely used compensation techniques are considered, i.e., feature mapping at the cepstral level and NAP at the SVM feature level. For the MLLR-SVM systems, the type of transform (MLLR versus CMLLR), the model (GMM versus phonemic HMM) and the front-end (ASR versus SID cepstral normalizations) are studied. This last point is specially meaningful in the context of the recent NIST Speaker Recognition Evaluation (SRE) campaigns, focusing on channel mismatch. Using a channel-compensated front-end allows MLLR adaptation to focus on the speaker components of cepstra rather than both speaker and channel components.

The remainder of the paper is organized as follows. Section II reviews adaptation methods as well as speaker adaptive training. Section III provides a quick overview of support vector machines for the speaker recognition tasks. Section IV presents the evaluation protocols and task used in these experiments. Section V describes the architectures developed for this work, starting with the cepstral front-ends, then the LVCSR acoustic models, and finally the configuration of the SVM-based systems targeted in this study. In Section VI, the acoustic speaker recognition systems used as an experimental baseline are described. In Section VII, we present and discuss the results for GSV-SVM and MLLR-SVM systems individually as well as the fusion results for the NIST 2005 and 2006 Speaker Recognition Evaluations. Conclusions are given in Section VIII.

II. Speaker Adaptation

Speaker adaptation techniques seek to obtain a speaker-dependent model given a speaker-independent model and some speech data belonging to a specific speaker. The speaker-independent model is typically trained using speech data from many speakers. The adaptation procedure transforms the model parameters to optimize a certain criterion, e.g., maximizing posterior probability or likelihood. This section presents three techniques for Gaussian mean adaptation, namely MAP, i.e., standard Bayesian adaptation, and MLLR and constrained MLLR under the maximum-likelihood criterion. The use of CMLLR in SAT is described in the last part of the section.

A. Maximum a Posteriori

A Gaussian mixture model (GMM) for a random multivariate variable $\mathbf{x}$ can be formulated as

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{N} \lambda_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where λ_i is the weight for the ith Gaussian, $\mathcal{N}()$ is the Gaussian probability density function and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance matrix for Gaussian i.

MAP estimation [6], [7] maximizes the *a posteriori* distribution of the adaptation data $\mathbf{X}$ given the a priori model parameters $\boldsymbol{\Theta}$, that is, using the Bayes formula

$$\arg\max_{\boldsymbol{\Theta}} p(\mathbf{X}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \tag{2}$$

where $p(\mathbf{X}|\boldsymbol{\Theta})$ is the likelihood function of $\mathbf{X}$ given the model parameters and the prior distribution for the mean vectors are assumed to be Gaussian.

The re-estimation formulas are derived using the expectation–maximization (EM) algorithm, which balances the new estimates on the adaptation data and the prior knowledge. Given that mean vectors are placed at the most likely points of each Gaussian component, an efficient way of changing the overall statistical distribution is by shifting them. Thus, a simple form of MAP adaptation is mean adaptation[1] which moves the Gaussian mean vectors according to

$$\hat{\boldsymbol{\mu}}_i = \alpha_i E_i\{\mathbf{x}\} + (1 - \alpha_i)\boldsymbol{\mu}_i \tag{3}$$

where $\hat{\boldsymbol{\mu}}_i$ is the adapted mean vector for the ith Gaussian, $E_i\{\mathbf{x}\}$ the expected mean feature vector for the adaptation data, $\boldsymbol{\mu}_i$ its prior mean vector, $\mathbf{x}$ a random feature vector, and α_i the adaptation factor

$$\alpha_i = \frac{n_i}{n_i + \tau} \tag{4}$$

which weights the old and new estimates via the relevance factor τ. Given a specific sequence of adaptation data $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_t}, \ldots, \mathbf{x_T})$ with $1 \leq t \leq T$, the effective number of frames assigned to Gaussian i, n_i is estimated as

$$n_i = \sum_{t=1}^{T} p(i|\mathbf{x}_t) \tag{5}$$

and $E_i\{\mathbf{x}\}$ as

$$E_i\{\mathbf{x}\} \approx \frac{1}{n_i} \sum_{t=1}^{T} p(i|\mathbf{x}_t)\mathbf{x}_t \tag{6}$$

where $p(i|\mathbf{x}_t)$ is the occupancy probability for the ith Gaussian, defined as

$$p(i|\mathbf{x}_t) = \frac{\lambda_i \mathcal{N}(\mathbf{x_t}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{M} \lambda_j \mathcal{N}(\mathbf{x_t}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{7}$$

B. Maximum-Likelihood Linear Regression

MLLR [17], [18] adapts the observation probability of a HMM in a parametric way, by finding a transform that maximizes the likelihood of the adaptation data given the transformed Gaussian parameters, i.e., $p(\mathbf{X}|\boldsymbol{\Theta})$. As opposed to standard MAP adaptation which adapts only the observed Gaussian components, MLLR adapts all of the components in a set of Gaussians, a so-called regression class. In mean

[1]We present mean adaptation only since these parameters are commonly used in speaker recognition. Please refer to [6] and [7] for the weight and covariance re-estimation formulas.

adaptation, Gaussian mean vectors $\boldsymbol{\mu}$ of the model are adapted using an affine transform with parameters $\mathbf{A}$ and $\mathbf{b}$ as

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{8}$$

where $\hat{\boldsymbol{\mu}}$ is the adapted mean vector. Using the resulting mean-adapted model, covariance matrices can be also adapted as

$$\hat{\boldsymbol{\Sigma}} = \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^T \tag{9}$$

at the expense of estimating the additional linear transform $\mathbf{H}$. $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ are, respectively, the non-adapted and adapted covariance matrices for the adapted Gaussian. As in mean adaptation, covariance matrices are also adapted in the maximum-likelihood sense using the EM algorithm. Details on the estimation procedure and MLLR variants can be found in [18].

MLLR transforms are typically estimated across a set of Gaussians, a regression class, that shares the same transformation parameters.[2] Using the acoustic models of a LVCSR system, it is relatively easy to define a fixed number of regression classes based on the phonetic similarity of tri-phone models. More sophisticated approaches use knowledge-based or data-driven decision trees that dynamically determine the number regression classes based on the observation probability similarity and taking into account the amount of available adaptation data per class [19]. Therefore, each of the regression classes results in a separate MLLR transform that is used to adapt a subset of the Gaussian parameters in the model.

C. Constrained MLLR

A main concern of MLLR adaptation is how to reliably estimate the regression coefficients using the available training data. It is common to simplify the regression model by using diagonal or block-diagonal covariance matrices [18] thereby reducing the number of parameters in the linear regression model or to share the mean and variance transforms. CMLLR [20] as described in this section falls into the latter category, using the same transform for mean vector and covariance matrix adaptation. For an arbitrary Gaussian component in a regression class, its parameters are transformed as

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{10}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \tag{11}$$

where the linear transform $\mathbf{A}$ is used for adaptation of both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. A main difference from MLLR adaptation of the Gaussian mean vectors is that, using the same number of parameters, the covariance matrices are also adapted. The algorithm used for MLLR adaptation can also be used to estimate the CMLLR transforms. Sufficient statistics are computed given the current estimates of $\mathbf{A}$ and $\mathbf{b}$ in the expectation step and the likelihood function is maximized with respect to these parameters in the maximization step.

When only one regression class is used, adaptation can be performed in the model-space, as in (10), or alternatively in the feature-space by transforming the features so that the likelihood function with respect to the speaker-independent model is maximized. The feature-space transform is

$$\mathbf{x_t} = \mathbf{A^{-1}}\hat{\mathbf{x}}_\mathbf{t} - \mathbf{A^{-1}}\mathbf{b} \tag{12}$$

where $\mathbf{x_t}$ is the speaker-independent feature vector at time t and $\hat{\mathbf{x}}_\mathbf{t}$ is the corresponding speaker-dependent feature vector. This property is particularly useful in SAT, used in the feature extraction scheme presented in [16] and described next.

[2]Note that MLLR adaptation of a single Gaussian is equivalent to ML retraining of the Gaussian.

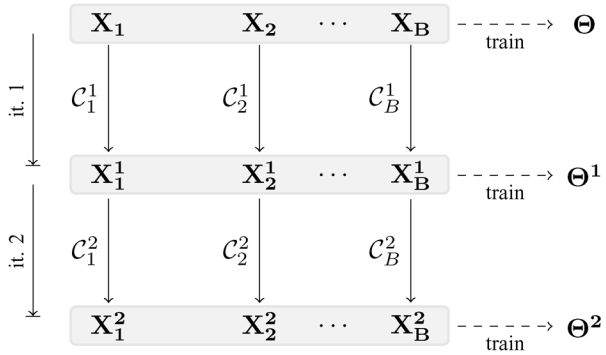


Fig. 1. Block diagram of two iterations of SAT.

D. Speaker Adaptive Training

A common use of feature-space CMLLR is SAT [15] which seeks to jointly estimate a set of CMLLR transforms, one per speaker, and a speaker-independent model in the transformed feature space. Such a speaker-independent model captures the overall feature distribution of a large number of speakers. Given a set of B speakers and their corresponding adaptation cepstra $\mathbf{X_i}$ for $1 \leq i \leq B$, SAT optimizes the maximum likelihood criterion on a per-speaker basis as

$$\arg\max \boldsymbol{\Theta}, \mathcal{C}_i \prod_{i=1}^{B} p\left(\mathcal{C}_i(\mathbf{X_i})|\boldsymbol{\Theta}\right) \tag{13}$$

where the individual speaker-dependent transforms $\mathcal{C}_i$ and the model parameters $\boldsymbol{\Theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N)$ are jointly estimated. Such an optimization is commonly done in two steps by, first, estimating the feature-space CMLLR transforms $\mathcal{C}_i$ that project the speaker-dependent features onto a speaker-independent space and, second, retraining the speaker-independent model $\boldsymbol{\Theta}$ using those features. This process, illustrated in Fig. 1, can be iterated several times in an EM manner, obtaining a speaker-independent model with lower inter-speaker variability, at each iteration.

III. Support Vector Machines

The systems explored in this work use discriminative modeling of speakers based on SVMs, introduced in speaker verification a few years ago. Such classifiers are capable of successfully discriminating high-dimensional and sparse feature spaces where other modeling approaches fail to generalize. SVMs [21] are binary classifiers which use a weighted sum of kernel functions as the discriminant function. For a set of input–output pairs

of training samples $(\mathbf{x_l}, t_l)$ with $1 \leq l \leq N$ and $t_l \in \pm 1$ for positive and negative classes

$$f(\mathbf{x}) = \sum_{l=1}^{N_{SV}} \alpha_l t_l k(\mathbf{x_l}, \mathbf{x}) + b \tag{14}$$

where $\sum_{l=1}^{N_{SV}} \alpha_l t_l = 0$, $\alpha_l > 0$ and b is an offset. In this expansion, the N_{SV} support vectors $\mathbf{x_l}$, the training data points lying on the separation margin, as well as α_l are obtained so as to maximize the classification margin. The soft-margin variant further minimizes the number of classification errors so that it can deal with nonlinear separable data sets. The kernel function satisfies the Mercer condition, i.e., $k(\cdot,\cdot)$ must be positive semi-definite. This condition implies that k can be written as

$$k(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^T \phi(\mathbf{x_j}) \quad 1 \leq i, j \leq N \tag{15}$$

which is a regular dot product on a possibly infinite vector space mapped from the input space by the function $\phi(\cdot)$.

IV. Task and Evaluation

The speaker verification systems explored in this study were evaluated using conversational telephone speech data following the NIST 2005 and 2006 Speaker Recognition Evaluation (SRE)[3] protocols. A speaker verification system is asked to decide whether speech from a given target speaker is present in a particular speech segment. We used the SRE 2005 English-only core-condition data for system development and the SRE 2006 English-only core-condition data for system evaluation. These data consist of five-minute-long segments containing about two minutes of speech per conversation side.[4] A total of 646 (274 male/372 female) target speaker segments are available for model training in SRE 2005 and 816 (354 male/462 female) for SRE 2006. 2117 test speaker segments (907 male/1210 female) and 3735 (1606 male/2129 female) are available for SRE 2005 and SRE 2006, respectively. The ratio of impostor to true access trials is about ten in both cases and all trials involve speakers with the same gender, known *a priori*.

The primary performance measure for the NIST speaker detection task is the detection cost function (DCF) defined as the weighted sum of the false alarm and miss error probabilities $DCF_{\text{Norm}} = P_{\text{Miss}} + 9.9 \times P_{\text{FalseAlarm}}$. We also report the minimal DCF (MDC) value obtained *a posteriori* for the best possible detection threshold. Since this operating point favors false alarms, we provide the equal error rate (EER) as an alternative performance measure. The detection error tradeoff (DET) curves [22] are used to assess system behavior over the full range of operating points. The DET curve is comparable to the receiver operating characteristics (ROC) curve but uses a nonlinear axis, which results in a linear curve for a normal distribution, improving its readability.

[3]The NIST 2005 and 2006 SRE evaluation plans, http://www.nist.gov/speech/tests/spk/.

[4]The core conditions involve the two conversation sides.

V. System Description

The systems explored in this paper use the adaptation methods described in Section II to extract base features that are particularly relevant for SVM-based speaker recognition. All of them use SVM classifiers, differing only in the base feature vectors, and have the same postprocessing steps. The details of the systems are given in the following sections.

A. Front-End

We use two different cepstral front-ends as a side effect of using the previously trained models of the LVCSR system for the MLLR and CMLLR transform computation:

- **Speech Recognition (PLP12)**: This is the front-end used by the previously trained LVCSR system. It uses 39 cepstral features with 12 MEL-PLP coefficients and the log-energy along with their corresponding Δ and $\Delta\Delta$ coefficients extracted every 10 ms using a 30-ms window on the 0–3.8 kHz bandwidth. Mean and variance normalization are applied to each segment of interest. When used in the LVCSR-based systems, only the frames assigned to the speech states of the acoustic models are used. When used with the other systems, speech activity detection (SAD) is performed based on the voicing features as produced by the ESPS get_f0[5] pitch extraction algorithm.
- **Speaker Recognition (PLP15N)**: This front-end uses feature-level channel compensation and feature Gaussianization as is commonly done for speaker recognition. The configuration was optimized for use in past NIST SRE evaluations. We use 15 MEL-PLP coefficients along with their Δ, $\Delta\Delta$ coefficients, and the Δ and $\Delta\Delta$ energies for a total of 47 features. The features are extracted every 10 ms using a 30-ms window on the 0–3.8 kHz bandwidth. For the LVCSR-based systems, only the frames assigned to the speech states of the acoustic models are used. For the other systems, the voiced frames are determined by the ESPS get_f0[5] pitch extraction algorithm. We apply gender-specific feature mapping [2] to compensate for channel distortion using segments from the test speakers in previous NIST SRE test sets 1997 to 2002 (24 769 segments, 6 hours/gender) as training data. The resulting features are Gaussianized using feature warping [23] with a 3-s window.

B. LVCSR

We use several acoustic model setups to compute both the phonetic alignment and to estimate the MLLR transforms. The acoustic models and a pronunciation dictionary[6] are used to align the provided word-level transcripts with the audio data. We explore three acoustic model configurations, two based on the PLP12 and PLP15N front-ends and one trained using SAT:

- The **PLP12 AM** system is based on the LIMSI SWB speech-to-text system [24]. It uses gender-independent continuous density HMM with Gaussian mixtures for acoustic modeling. The acoustic models are tied-state,

[5]KTH Software, http://www.speech.kth.se/software.

[6]We used the pronunciation dictionary of the LVCSR system [24] and manually added missing entries.

context-dependent triphones, where the tied-states are automatically found by means of a decision tree. A total of about 6400 tied-states are used, with 32 Gaussians per state. This system uses the PLP12 front-end, i.e., 39 cepstral features, and it was trained on Switchboard I (4862 conversation sides), Switchboard II (2348 sides), Callhome (240 sides) and Fisher (6127 sides) corpora, for a total of 13577 conversation sides involving about 650 hours of data.

- The **PLP15N AM** system is the same as PLP12 AM except that it uses the PLP15N front-end, with the speaker-recognition-specific normalizations. Switching to the PLP15N front-end required re-training the acoustic models. For this purpose, exactly the same training data was used as for PLP12 AM training. Since both front-ends result in time-aligned cepstra, the alignments produced with the PLP12 AM for the training data were also used when training the new PLP15N acoustic models. The PLP12 AM and PLP15N AM are therefore directly comparable.
- For the **PLP15N+SAT AM** system, the PLP15N AM acoustic models were used as seed models for one iteration of SAT re-estimation [15]. We computed one CMLLR transform per speaker using all of his/her speech data. The acoustic models were retrained using the CMLLR-transformed cepstra. In this case, we used a slightly different clustering threshold optimized for these features. We obtained a total of 6100 tied-states, a number comparable to the 6400 states in the PLP12 AM and PLP15 AM.

C. SVM-Based Systems

The SVM-based systems differ in the strategies used to obtain the base supervectors, one per speaker and per session. They share the same postprocessing and SVM setup in order to ease the comparison of the different features. The training data and tuning parameters were set to maximize the SRE 2005 cross-validation performance.

Nuisance attribute projection (NAP) [4], [25] inter-session variability compensation is applied to the base supervectors, prior to normalization. NAP finds a linear transform that removes the subspace exhibiting the largest inter-session variability in the feature space.[7] The NAP transform is obtained using NIST SRE 2004 training data, which is known to potentially have a high inter-session variability.[8] We set the session subspace dimension to 50 which was experimentally found to be almost optimal for all systems described in this paper.

The resulting supervectors are normalized by means of min–max component scaling. Every feature is fit into the range $[-1/\sqrt{M}, 1/\sqrt{M}]$, where M is the number of features of the vector. This forces the SVM to deal with dot products with a maximum magnitude of 1. The resulting mean value of the features is expected to be 0, so any offset before normalization is removed. Min–max statistics are collected from the impostor speaker set described below. In preliminary experiments, this method was found to outperform mean and variance normalization as well as rank normalization for several SVM-based acoustic systems.

The impostor speaker data consists of 2243 speech segments[9] from the NIST SRE 2004 training data plus 4854 speech segments[10] from the Switchboard I (SWB1) corpus, all in the English language with a minimum and an average effective duration of 10 seconds and 2 minutes[11] respectively. Transcripts are available for all of the segments. The SRE 2004 transcripts were obtained automatically using the RT'03 BBN speech recognition system and they were provided by NIST for the SRE 2004 evaluation. The SWB1 data were manually transcribed (LDC Corpus 93T4). All SVM-based systems share the same impostor data, since transcripts are needed for some MLLR systems but not for other acoustic systems.

The SVM classifier uses a linear kernel and it is trained using gender-dependent impostor speaker data. We used the SVMTorch[12] package developed at the IDIAP laboratory, without score normalization as it resulted in a performance loss.[13]

D. GSV-SVM System

The Gaussian supervector (GSV) approach [11] uses the mean vectors of a speaker-dependent GMM as features, where these are obtained via standard MAP adaptation[14] [7] of a previously trained GMM-UBM estimated using speech data from many speakers. Assuming N Gaussian components in the GMM, the mean vectors $\boldsymbol{\mu_i} = [\mu_{i1}, \ldots, \mu_{iC}]^T$ for $1 \leq i \leq N$ are arranged as

$$\mathbf{m} = \left[\boldsymbol{\mu_1^T}, \boldsymbol{\mu_2^T}, \ldots, \boldsymbol{\mu_N^T}\right]^T \tag{16}$$

resulting in a Gaussian mean supervector $\mathbf{m}$ of dimension NC, where N is the number of Gaussians and C the dimension of the cepstral features. For a speaker of interest one vector is used as the base supervector.

For the PLP15N features, we use two gender-dependent UBMs with diagonal covariance matrices trained on about 120 hours of speech data per gender, the same impostor speaker data that is used for SVM training, i.e., NIST SRE 2004 and Switchboard I data. We use five iterations of maximum likelihood training with 1% of the global variance as the variance floor. The number of Gaussians used ranges from 64 up to 1024 depending on the configuration. To obtain the speaker-specific models we use three iterations of standard MAP mean adaptation with a relevance factor of 10.

[7]An orthonormal set of vectors spanning the maximal inter-session variability subspace can be obtained from the eigenvectors $\mathbf{E} = (\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_{D_{NAP}}})$ corresponding to the largest eigenvalues of the inter-session covariance matrix. Based on $\mathbf{E}$, we use the projection matrix $\mathbf{I} - \mathbf{E}\mathbf{E}^T$ to remove session variability from a feature vector.

[8]Most of the 310 speakers have more than ten sessions per speaker involving several channel conditions

[9]About 60 hours of speech, after speech activity detection.

[10]For a total of about 170 hours of speech excluding silence segments.

[11]For homogeneity with train and test data, which have an average duration of 2 minutes as well.

[12]SVMTorch, a support vector machine for large-scale regression and classification problems http://www.idiap.ch/learning/SVMTorch.html.

[13]We found both gender-independent and gender-dependent T-norm to be harmful for several SVM-based systems. We believe this might be related to the highly skewed score distributions, far from a Gaussian shape, output by the SVM. Scores gather around -1 roughly ranging from -0.8 to -1.1, which seems to be due to the strong imbalance of the training data, i.e., 1 true speaker against 7000 impostor speakers.

[14]eigenchannel [8] or joint factor analysis [9] are alternative methods which can be used to obtain inter-session compensated supervectors.

Given the high dimensionality of the supervectors used, reaching tens of thousands of components for the best performing configurations, the feature dimensionality can become larger than the number of training samples. We use a soft-margin SVM for classification since, in such degenerate situations, it successfully avoids overfitting by balancing machine complexity versus training performance.

We use the SVM configuration described in Section V-C. The linear kernel is derived from an approximation of the Kullback–Leibler (KL) divergence, a measure of dissimilarity between the distributions given by the GMM of two speakers, described in [25]. Given two models for segments s^a and s^b, the distance can be expressed as

$$k(s^a, s^b) = \sum_{i=1}^{N} \left(\sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-\frac{1}{2}}\boldsymbol{\mu}_i^{\boldsymbol{a}}\right)^T \left(\sqrt{\lambda_i}\boldsymbol{\Sigma}_i^{-\frac{1}{2}}\boldsymbol{\mu}_i^{\boldsymbol{b}}\right) \tag{17}$$

where we keep the notation used in (1). The covariance matrices are the same for both segments since only the means are adapted. This kernel satisfies Mercer's condition since it is linear. A regular dot product using the normalized Gaussian supervectors

$$\mathbf{m}' = \left[\sqrt{\lambda_1}\frac{\mu_1}{\sigma_1}, \sqrt{\lambda_2}\frac{\mu_2}{\sigma_2}, \ldots, \sqrt{\lambda_M}\frac{\mu_M}{\sigma_M}\right]^T \tag{18}$$

where $M = NC$, N is the number of Gaussians and C is the dimension of the cepstral feature vectors, is equivalent to (17). μ_i and σ_i for $1 \leq i \leq M$ are the scalar mean and variance parameters of the corresponding cepstral and Gaussian components. We prefer this second form since the normalized supervectors can be post-processed arbitrarily, e.g., for inter-session variability compensation.

E. MLLR-SVM Systems

The MLLR-SVM systems use the MLLR regression coefficients arranged in a vector form as the base supervectors and a SVM as classifier. We use two MLLR-SVM variants in our experiments, MLLR$_h$-SVM and MLLR$_g$-SVM where either the acoustic models of a LVCSR system or a GMM-UBM are used to align cepstra and compute MLLR transforms.

1) $MLLR_h$-SVM: This system is based on the MLLR-SVM system proposed in [12]. Given the orthographic transcription of a speech segment, we use the acoustic models of a LVCSR system described earlier in Section V-B and the pronunciation dictionary to align the corresponding speech data to the transcripts. This alignment is used to assign each frame to a regression class, covering a part of the acoustic space. One MLLR transform is computed per regression class using the same acoustic models used for alignment. The coefficients of transform r are stacked as a supervector of the form

$$\mathbf{m_r} = [\mathbf{A^r_{11}}, \ldots, \mathbf{A^r_{1C}}, \ldots, \mathbf{A^r_{C1}}, \ldots, \mathbf{A^r_{CC}}, \mathbf{b^r_1}, \ldots, \mathbf{b^r_C}]^T \tag{19}$$

with $\mathbf{A^r}$ and $\mathbf{b^r}$ being the matrix and the offset,[15] i.e., only mean vectors are adapted, and C the dimension of the cepstral feature vectors. The supervectors for all transforms are concatenated together in one vector

$$\mathbf{m} = \left[\mathbf{m_1^T}, \ldots, \mathbf{m_R^T}\right]^T \tag{20}$$

assuming a total of R regression classes. We use such a vector as the base supervector for every speaker of interest.

The number of transforms used depends on the amount of speech data available for adaptation. Using many classes results in a finely represented phonetic space but less speech data is available for each class-dependent transform. We force a full-matrix[16] transform regardless of the amount of adaptation data assigned to the corresponding class. Three static regression-class configurations involving only speech[17] are used in these experiments.

- One transform (**1t**), speech only.
- Two transforms (**2t**), vowels and consonants.
- Three transforms (**3t**):
  - — fricative and stop consonants;
  - — nasal consonants, semivowels and back vowels;
  - — front vowels.

The MLLR supervectors rapidly end up with thousands of features[18] that are classified using a SVM, as described in Section V-C. The linear kernel reduces to computing a regular dot product of the MLLR supervectors as

$$k(s^a, s^b) = (\mathbf{m}^a)^T\mathbf{m}^b \tag{21}$$

where $\mathbf{m}^a$ and $\mathbf{m}^b$ are the MLLR supervectors, as defined in (20), corresponding to speech segments s_a and s_b.

2) $MLLR_g$-SVM: A large-vocabulary ASR system needs huge amounts of speech and text data, as well as substantial computational resources for training. This makes the implementation of such a system not accessible to everyone. A simple and cost-effective alternative is to replace the acoustic HMM by a GMM-UBM, hence MLLR$_g$-SVM. The cepstra are now aligned against a single HMM state with a global Gaussian mixture observation probability. The phonetic-class alignment is no longer straightforward.[19] However, a GMM-based system can be used for any language since no transcripts or ASR hypotheses are required. Another advantage of a GMM-based approach is that any cepstral front-end with any kind of normalization, including session and channel compensation, can be used. Multiple SAT iterations can also be performed as CMLLR computation and training are faster for a GMM than for the acoustic models of a LVCSR system.

We use two gender-dependent GMM-UBM trained using the impostor data, i.e., SRE 2004 and Switchboard I. These are the same GMM-UBMs used by the GSV-SVM system. A single MLLR transform is computed and the corresponding

[15]Offset coefficients are always included as they directly compensate for convolutional distortion in the cepstral features.

[16]According to experiments that are not included in this paper, backing off to a diagonal MLLR transform when lacking data for reliable estimation resulted in increased error rates for the 2t and 3t classes.

[17]The non-speech class involving silence and breath is dropped as it is assumed to carry no speaker information.

[18]For the PLP12 front-end, each MLLR transform has $39 \cdot 39 + 39 = 1560$ coefficients, $47 \cdot 47 + 47 = 2256$ for PLP15N. This dimension is multiplied by the number of transforms.

[19]Explicit assignment of Gaussians has been used as an alternative in [26]

supervector is normalized and classified as in the MLLR_h-SVM system. We refer to this simplified approach as MLLR_g-SVM.

F. CMLLR-SVM Systems

The CMLLR-SVM systems follow the same strategy as the MLLR-SVM systems. We also explore two variants of the CMLLR-SVM approach depending on whether the acoustic models of a LVCSR system or of a GMM-UBM are used to compute the CMLLR transforms, resulting in the CMLLR_h-SVM and CMLLR_g-SVM systems, respectively.

1) $CMLLR_h$-SVM: This system uses the acoustic HMM of an LVCSR system for alignment and to estimate the feature-space CMLLR transforms. We use one transform per speaker segment corresponding to the speech class only, as is often performed in LVCSR systems. In our experiments, this also allows comparison with a purely acoustic approach based on a GMM-UBM, which uses a single transform for the whole model. We compute one feature-space transform per segment given by the parameters $(\mathbf{A^{-1}}, -\mathbf{A^{-1}b})$ which are inverted to obtain the model-space transforms $(\mathbf{A}, \mathbf{b})$ whose parameters are actually used for classification. These latter parameters are the features used in the CMLLR_h-SVM system. Other than using CMLLR transforms, all steps are the same as in the MLLR_h-SVM system.

2) $CMLLR_g$-SVM: The CMLLR_g-SVM system estimates a single feature-space CMLLR transform using a GMM-UBM. In principle, this approach is thought to work together with SAT, since the transforms used for model training and those used for feature extraction become homogeneous, i.e., both use CMLLR. A main difference of this approach with respect to SAT as used in speech recognition is that we compute one feature-space CMLLR transform per speaker segment, resulting in a speaker- and session-dependent transform. In the training phase, these transforms remove both the inter-speaker and inter-session variabilities in the GMM-UBM.

We use the gender-dependent GMM-UBMs used in the GSV-SVM system, i.e., trained using SRE 2004 and Switchboard I data. When using SAT, we perform one retraining iteration only so that GMM-based and LVCSR-based systems use the same number of iterations, which eases comparison of the systems using SAT. For feature extraction purposes, the feature-space CMLLR transforms are inverted as in CMLLR_h-SVM to obtain the model-space transform parameters. Further processing is the same as in the MLLR_h-SVM system.

VI. Baseline Systems

Although the main aim of this paper is to compare systems using different adaptation methods, it is also interesting to test their behavior in combination with other systems given that, currently, fusing systems integrating some degree of diversity is a major source of system performance improvement. In this section we describe the two acoustic state-of-the-art systems used as baseline systems in these experiments either for individual or fused system comparison.

A. PLP-SVM System

The PLP-SVM system is based on the generalized linear discriminant sequence (GLDS) kernel [27] and uses PLP15N features, explicit polynomial mapping and a SVM classifier. Cepstra are expanded by concatenating first, second- and third-order monomial expansions forming as many supervectors as cepstral vectors. These are normalized to have a unity variance within the speech segment and finally averaged, obtaining 20 824 features per speaker segment. This expansion can be seen as estimating first, second- and third-order statistical moments of the cepstral vectors. We use a kernel principal component analysis (KPCA) [28] with a second-order polynomial kernel to extend the polynomial features to orders higher than three[20] while reducing the dimensionality of the feature space. We used 2917 session in the SRE 2004 data to train the KPCA transform. Taking all of the eigenvectors, we obtain 2917 output features. These vectors are kept as base features for the PLP-SVM following NAP compensation, normalization and SVM classification as in all SVM-based systems.

B. PLP-GMM System

The PLP-GMM system [29] is based on the GMM-UBM paradigm [6] using hybrid-domain eigenchannel compensation based on a factor analysis model of utterance variability [30]. The front-end is based on the PLP15N analysis bypassing feature mapping, since it showed a negative interaction with factor analysis compensation. We use two gender-dependent UBMs with 1536 mixtures each, trained on about 24 hours of speech from SRE 2000 and SRE 2001 development and training data. Covariance matrices are diagonal and a variance floor threshold of 1% of the global variance is applied at each training iteration. Speaker models are obtained using three iterations of eigenchannel adaptation with the ALIZE toolkit [31], thus performing model-domain session compensation for the target speaker segments. The channel factor-loading matrix was trained using the SRE 2004 training data, the same data as used for NAP compensation in the SVM-based systems, and a channel-space dimension of 40. Test segments are compensated in the feature-domain and scored using standard log-likelihood ratios, taking the 20 top-scoring Gaussians. We use gender-dependent T-norm [5] for score normalization based on 250 male and 250 female segments taken from the SRE 2004 training data.

C. System Fusion

SVM-based systems obtain scores by projecting test segment supervectors against the supervector obtained for the target speaker during training. These scores should be in the range $[-1, +1]$, although the large number of impostor speaker segments used for training highly biases their distribution towards -1. The PLP-GMM system outputs log-likelihood-ratio scores, i.e., a target model versus the UBM likelihood scores.

We use forward–backward scoring for all of the systems [32], which aims at making the train and test phases symmetric. The forward system uses a conventional approach where the test speech is scored against the target speaker model. In the backward system, we score the training speech against the test

[20]Explicit polynomial mapping is untractable for such orders.

TABLE I
MDC AND EER OF GSV-SVM SYSTEMS ON THE SRE 2005 AND SRE 2006 EVALUATION DATA. COLUMN F SHOWS FORWARD SCORES AND COLUMN FB SHOWS AVERAGED FORWARD AND BACKWARD SCORES WITH A WEIGHT OF 0.5. THE BEST SCORES IN EACH COLUMN ARE SHOWN IN BOLDFACE

	SRE 2005				SRE 2006			
	MDC		EER (%)		MDC		EER (%)	
System	F	FB	F	FB	F	FB	F	FB
GSV 64g	.0226	.0214	5.32	5.24	.0207	.0201	4.82	4.59
GSV 128g	.0192	.0190	5.11	5.03	**.0181**	.0174	4.03	3.91
GSV 256g	**.0177**	**.0172**	**4.66**	4.45	.0182	**.0174**	**3.54**	**3.26**
GSV 512g	.0186	.0179	4.91	**4.40**	.0200	.0193	4.09	3.77
GSV 1024g	.0199	0.187	5.03	4.62	.0218	.0184	4.26	3.67

speaker model. Therefore, we obtain two scores per system and per trial. Each score is considered individually for system fusion.

We use a logistic regression model[21] for score fusion, which outputs normalized log-likelihood-ratio scores. The SRE 2005 data was used for training the model and the SRE 2006 data was used for test, so only performance for the latter are shown in the results. Scores for empty segments were excluded from training.

VII. RESULTS

A. Individual Systems

We conducted two series of experiments, one to evaluate MAP adaptation in the GSV-SVM systems and one to evaluate MLLR adaptation in the MLLR-SVM systems, with a focus on the latter. We give results for forward and forward–backward fused systems on both the NIST SRE 2005 and SRE 2006 data. Improvements are always relative unless otherwise stated.

1) GSV-SVM Systems: The GSV-SVM performance was first assessed for several configurations differing in the number of Gaussians used in the GMM speaker models. We tested from 64 to 1024 Gaussians in exponential steps. The relevance factor τ was set to 10, the same value obtained by optimization on the PLP-GMM system. MDC and EER values for forward and forward–backward averaged systems are shown in Table I. Decreasing error rates can be observed as more Gaussians are used in the speaker models up to 256 Gaussians, slightly lower than the optimal number reported in other studies [33]. For 512 and more Gaussians performance drops again, probably due to having to estimate too many parameters in the speaker models for the amount of data actually used for adaptation. This trend is seen for both SRE 2005 and SRE 2006 data.

Forward–backward system fusion brings improvements in all cases, with gains dependent on the system and the evaluation data. Overall, the relative gains are in the range of 3% to 15% for MDC and 1.5% to 13% for EER, thus exhibiting a large variability.

2) MLLR-SVM Systems: The MLLR-SVM systems allow exploration of a wide variety of adaptation schemes each with their pros and cons. CMLLR performs mean and variance adaptation at the cost of reduced adaptation capability. GMM do not require transcripts to perform adaptation, but modeling is less precise than LVCSR acoustic models. Using a channel-compensated front-end allows MLLR adaptation to focus on modeling only the speaker components but requires retraining of the acoustic models. Several acoustic classes can be used for more precise adaptation when enough speech data are available. In these experiments, we explored front-end type, transform type, model type and training technique and number of transforms used for acoustic model adaptation. Given the large number of different configurations, acronyms are introduced to ease further discussion. The naming conventions designate systems by capital letters indicating the type of MLLR transform (CMLLR or MLLR), model type (GMM or HMM) and the number of cepstral coefficients in the front-end (12 for PLP12 or 15 for PLP15N). For the MLLR systems using HMM, the number of transforms is also specified. Eventually, if speaker adaptive training is used, the SAT term is added too. Table II shows the system names along with their respective configurations.

[21]We used the FoCal toolkit http://www.dsp.sun.ac.za/nbrummer/focal/index.htm.

TABLE II
SYSTEM NAMING CONVENTION FOR CMLLR-SVM AND MLLR-SVM SYSTEMS. COLUMNS SPECIFY SYSTEM ACRONYM, TYPE OF TRANSFORM (CMLLR VERSUS MLLR), MODEL TYPE (GMM VERSUS HMM), FRONT-END TYPE (PLP12 VERSUS PLP15N), SAT ($\surd$) OR STANDARD ML ($\times$) MODEL TRAINING, AND NUMBER OF TRANSFORMS (1 TO 3)

System	Front-end	Model	SAT	Type	Transforms
CG12	PLP12	GMM	$\times$	CMLLR	1
CG12 SAT	PLP12	GMM	$\surd$	CMLLR	1
CH12	PLP12	HMM	$\times$	CMLLR	1
CG15	PLP15N	GMM	$\times$	CMLLR	1
CG15 SAT	PLP15N	GMM	$\surd$	CMLLR	1
CH15	PLP15N	HMM	$\times$	CMLLR	1
CH15 SAT	PLP15N	HMM	$\surd$	CMLLR	1
MG12	PLP12	GMM	$\times$	MLLR	1
MG12 SAT	PLP12	GMM	$\surd$	MLLR	1
MH12 1t	PLP12	HMM	$\times$	MLLR	1
MH12 2t	PLP12	HMM	$\times$	MLLR	2
MH12 3t	PLP12	HMM	$\times$	MLLR	3
MG15	PLP15N	GMM	$\times$	MLLR	1
MG15 SAT	PLP15N	GMM	$\surd$	MLLR	1
MH15 1t	PLP15N	HMM	$\times$	MLLR	1
MH15 1t SAT	PLP15N	HMM	$\surd$	MLLR	1
MH15 2t	PLP15N	HMM	$\times$	MLLR	2
MH15 2t SAT	PLP15N	HMM	$\surd$	MLLR	2
MH15 3t	PLP15N	HMM	$\times$	MLLR	3
MH15 3t SAT	PLP15N	HMM	$\surd$	MLLR	3

As shown in [16], convergence of the SAT re-estimation process in a CMLLR/GMM system is fast. One or two SAT iterations already provide a significant gain while keeping the computational cost at a reasonable level. For this reason, we use one SAT iteration in these experiments. This allows a fair comparison with the PLP15N+SAT acoustic models of the LVCSR system which used one iteration to keep computational resources at a reasonable level.

Table III compares results for several CMLLR-SVM systems with different front-ends and models, including SAT. The absolute MDC and EER are higher for the SRE 2005 data compared to the SRE 2006 data, suggesting structural differences in the two databases, e.g., the proportion of native speakers. This difference could be partly explained by the use of mostly native

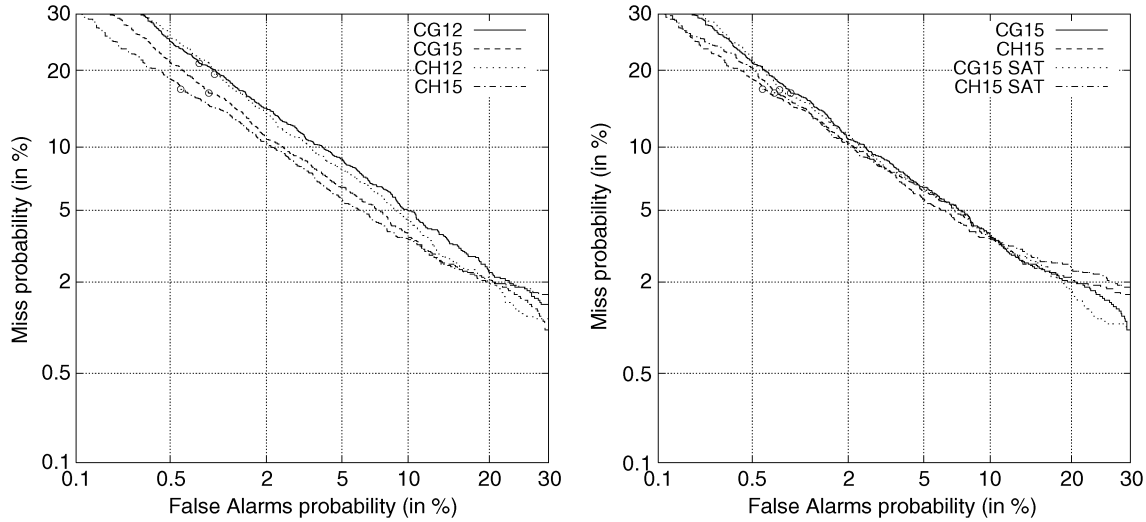


Fig. 2. DET curves for the CMLLR systems on the SRE 2006 evaluation data: varying the front-end (left) and model (right).

TABLE III
MDC AND EER OF CMLLR-SVM SYSTEMS ON THE SRE 2005 AND SRE 2006 EVALUATION DATA. COLUMN F SHOWS FORWARD SCORES AND COLUMN FB SHOWS AVERAGED FORWARD AND BACKWARD SCORES WITH A WEIGHT OF 0.5

	SRE 2005				SRE 2006			
	MDC		EER (%)		MDC		EER (%)	
System	F	FB	F	FB	F	FB	F	FB
CG12	.0342	.0308	7.57	8.15	.0290	.0265	6.88	6.71
CG12 SAT	.0326	.0293	7.82	7.72	.0287	.0253	6.53	6.20
CH12	.0329	.0298	7.53	7.24	.0292	.0258	6.70	6.39
CG15	.0303	.0276	7.93	7.86	.0255	.0241	5.96	5.61
CG15 SAT	.0292	.0265	7.27	6.99	.0246	.0241	5.88	5.61
CH15	**.0264**	**.0237**	**6.28**	**6.28**	**.0230**	**.0216**	**5.46**	**5.24**
CH15 SAT	.0286	.0244	6.32	6.36	.0236	.0220	5.84	5.56

English speaker data for training, resulting in a significant phonetic mismatch between train and test data.

The choice of the front-end has a large influence in performance, probably because of the specific speaker recognition normalizations,[22] namely feature mapping and feature warping, used for the PLP15N features.[23] Systems using PLP15N features outperform their PLP12-based counterparts. Using forward scoring, a relative gain of around 10%–14% in MDC, and up to 13% in EER are obtained for the GMM-based systems (CG15 versus CG12, CG15 SAT versus CG12 SAT) and from 11% to 20% in MDC or EER using LVCSR acoustic models (CH15 versus CH12). Forward–backward scoring, which makes scores less dependent on the target speaker[24] was found to improve performance by 5%–20% in MDC and 3%–18% in EER. Overall, the MDC gains are slightly higher for SRE 2006 while the EER reductions are larger for SRE 2005. Fig. 2 (left) shows DET curves for systems using PLP12 and PLP15N front-ends on the SRE 2006 data. A consistent improvement is seen for almost all operating points for both the GMM-based and HMM-based systems.

Concerning the type of model used to compute CMLLR transforms, an HMM is clearly advantageous with the PLP15N features, this is not the case with the PLP12 features. For the former, we observe relative gains of 8%–20% in MDC or EER for the forward systems and 6%–20% for the forward–backward fused systems (CH15 versus CG15). The DET curves in Fig. 2 (left) show rather consistent gains for these systems for most of the operating points. Gains are in general slightly lower for the PLP15 SAT acoustic models, reaching 13% but also as small as 1%. Using the PLP12 acoustic models results in very small improvements, with GMM-based systems eventually outperforming HMM-based systems. This can be seen in the left part of Fig. 2 (left) around the MDC, shown with the circle.

Using SAT models turns out to be slightly beneficial for GMM-based systems, but performance decreases for the HMM-based systems. Forward systems using GMM (CG15 SAT versus CG15) show relative improvements of over 3% in MDC and 1%–8% in EER. Gains are slightly larger for the forward–backward systems on the SRE 2005 data but no gain is observed on the SRE 2006 data. As for the HMM-based systems, the results suggest that there is a bad interaction of using the CMLLR transforms together with SAT, as this approach always leads to performance loss for both the forward and the forward–backward scored systems.

As for the MLLR-based systems, we assessed the effects of the front-end, model, number of transforms and SAT. Table IV shows results for the most relevant experiments. We discuss the most important points in the following.

The use of the PLP15N front-end, with specific speaker recognition normalizations, results in large performance improvements over PLP12, using mean and variance normalization. Relative gains of around 25% MDC and EER were obtained for most of the experiments (MH15 versus MH12), with both forward and forward–backward scoring. These gains seem to be independent of the number of transforms used.

[22]Note that all experiments use NAP inter-session compensation so, regarding the channel, the results actually show the interaction of channel mapping and NAP together. We found that NAP always brought a performance gain.

[23]Although the number of coefficients and the use of the log-energy coefficient also changes from PLP12 to PLP15N front-ends.

[24]Forward–backward scoring can be thought of as a rough form of per-trial T-norm using the test speaker as the only cohort speaker.

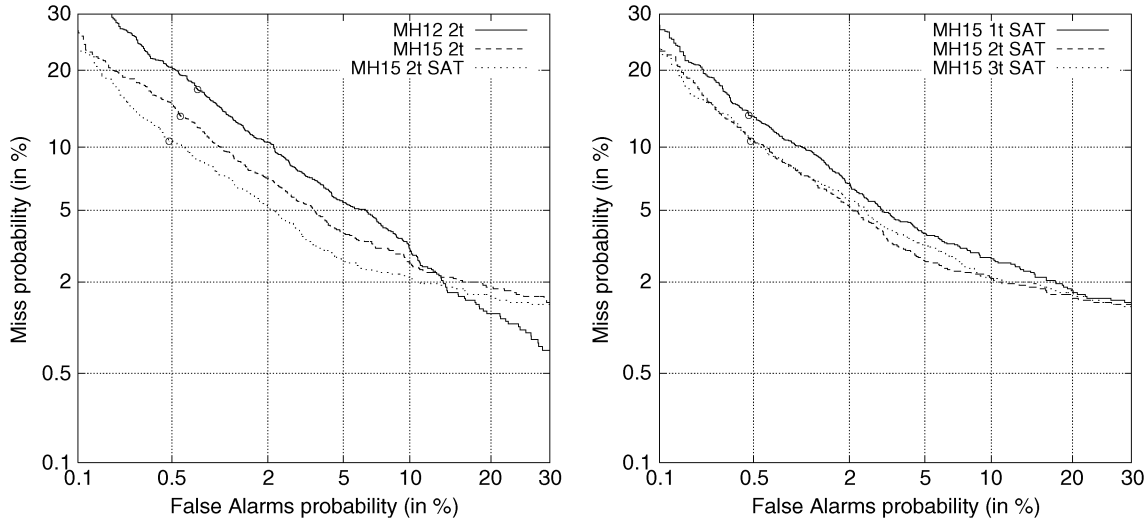


Fig. 3. DET curves for the MLLR systems on the SRE 2006 evaluation data: per-model curves using two transforms (left) and per-class curves using the PLP15N front-end and SAT models (right).

TABLE IV
MDC AND EER OF MLLR-SVM SYSTEMS ON THE SRE 2005 AND SRE 2006 EVALUATION DATA. COLUMN F CORRESPONDS TO FORWARD SCORES ONLY. COLUMN FB SHOWS AVERAGED FORWARD AND BACKWARD SCORES WITH A WEIGHT OF 0.5

	SRE 2005				SRE 2006			
	MDC		EER (%)		MDC		EER (%)	
System	F	FB	F	FB	F	FB	F	FB
MG12	.0384	.0363	9.81	9.36	.0301	.0290	7.40	7.13
MG12 SAT	.0326	.0310	8.07	7.94	.0272	.0248	6.16	5.79
MH12 1t	.0348	.0310	7.98	7.64	.0281	.0250	5.92	5.79
MH12 2t	.0310	.0268	6.94	6.65	.0245	.0206	5.37	5.05
MH12 3t	.0298	.0262	6.94	6.78	.0261	.0223	5.28	5.01
MG15	.0367	.0342	8.90	8.86	.0304	.0282	7.72	7.58
MG15 SAT	.0278	.0264	7.20	7.28	.0244	.0226	5.70	5.43
MH15 1t	.0254	.0232	5.74	5.70	.0207	.0189	5.10	4.82
MH15 2t	.0222	.0201	5.86	5.74	.0192	.0171	4.27	4.15
MH15 3t	.0219	.0201	5.45	5.41	.0191	.0171	4.27	4.23
MH15 1t SAT	.0199	.0180	4.86	4.57	.0183	.0164	4.32	4.31
MH15 2t SAT	**.0180**	**.0159**	**4.53**	**4.33**	.0155	.0143	**3.45**	**3.57**
MH15 3t SAT	.0185	.0167	4.95	4.87	**.0154**	**.0136**	3.77	3.68

Although the PLP12 and PLP15N front-ends also differ in the number of PLP coefficients, we believe that most of the gain is obtained by using feature mapping and feature warping.[25] Fig. 3 (right) shows DET curves for the MH12 2t and MH15 2t systems on the SRE 2006 data. The improvement obtained using the PLP15N front-end is consistent in the low false-alarm probability region covering the MDC and the EER operating points.

Using the acoustic models of the LVCSR system instead of a GMM-UBM improves system performance significantly, even though we restricted here to only one transform. The gains for systems using the PLP12 features are half those of the systems using the PLP15N features, keeping in mind that GMM-based and HMM-based systems differ in the speech activity detection used, i.e., voicing level versus alignment, and the SAT approach used, i.e., per-session SAT versus per-speaker SAT respectively. PLP12 experiments (MH12 versus MG12) show overall relative gains of 10% in MDC and EER regardless of the scoring approach and evaluation corpus. PLP15N systems using HMM (MH15 versus MG15, MH15 SAT versus MG15 SAT) show enormous gains compared to those using a GMM, in the range of 28% to 37% both in MDC or EER. These results are stable regardless of the scoring approach and the use of SAT. They highlight the importance of precisely modeling speech in a text-independent speaker recognition task. Its combination with speaker-specific acoustic-level normalizations seems specially fruitful. Their interaction with NAP or the number of coefficients has not been explored in these experiments.

SAT models bring significant gains compared to regular ML training and are additive to those obtained using PLP15N features. GMM-based systems using SAT and PLP12 features (MG12 SAT versus MG12) obtain relative improvements of around 15% in MDC and EER with even larger gains, in the range 17%–28% in MDC and EER when PLP15N features (MG15 SAT versus MG15) are used. Systems using CMLLR transforms (CG12 SAT and CG15 SAT) obtain similar gains. Note, however, that these systems outperform MLLR counterparts when standard maximum likelihood training is used for GMM (CG12 versus MG12, CG15 versus MG15). For HMM-based systems, SAT results in large performance gains too, in the range of 15%–26% in MDC and 9%–28% in EER on both the SRE 2005 and SRE 2006 data. Considering the cumulative gain from systems using PLP12 acoustic models to those using PLP15N SAT acoustic models, improvements over 40% are achieved. Fig. 3 (left) shows DET curves for PLP12, PLP15 and PLP15 SAT acoustic models for systems using 2 transforms on the SRE 2006 data. The cumulative improvement is constant for a wide range of operating points.

The number of MLLR transforms has a considerable impact on system performance. There is an overall trend for lower error rates as more classes are used in the system. Since each MLLR transform specializes in one acoustic space split, data are better

[25]We observed in past experiments that using 12 to 16 PLP coefficients with feature mapping and warping resulted in similar performance, although 15 was found to be optimal.

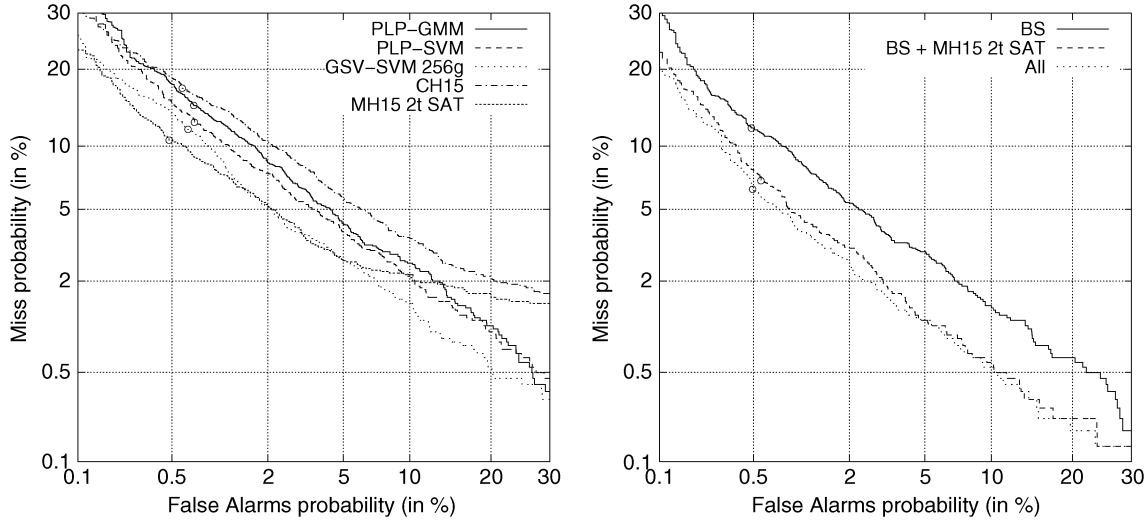


Fig. 4. DET curves for individual systems on the SRE 2006 evaluation data (left) and fused systems (right).

fit to the linear regression model, affecting performance correspondingly. However, the results show that using three transforms is not always beneficial.[26] The amount of data used to estimate each transform plays an important role for the segment lengths we deal with: the more classes the less data are available per class. Along these lines, the PLP12 and PLP15N features result in 1560 and 2256 regression coefficients per transform, respectively. About 30% more parameters must be estimated for the latter. We note, for instance, that the diagonal MLLR back-off rates dramatically rose using the PLP15N features, although we forced full MLLR matrices in all of the presented experiments. Fig. 3 (right) shows DET curves for the MH15 SAT systems using from one to three transforms. Going from one to two transforms brings a consistent improvement along almost the entire operating range, while going from two to three transforms does not improve the performance anywhere.

TABLE V
MDC AND EER OF INDIVIDUAL AND FUSED SYSTEMS ON THE SRE 2005 AND SRE 2006 EVALUATION DATA. COLUMN F SHOWS FORWARD SCORES AND COLUMN FB SHOWS AVERAGED FORWARD AND BACKWARD SCORES WITH A WEIGHT OF 0.5

	SRE 2005				SRE 2006			
	MDC		EER (%)		MDC		EER (%)	
System	F	FB	F	FB	F	FB	F	FB
(a) PLP-GMM	.0287	.0202	5.82	4.74	.0218	.0177	4.69	3.72
(b) PLP-SVM	.0211	.0204	4.82	4.48	.0198	.0189	4.41	4.14
(c) GSV-SVM 256g	**.0177**	.0172	4.66	4.45	.0182	.0174	3.54	**3.26**
(d) CH15	.0264	.0237	6.28	6.28	.0230	.0216	5.46	5.24
(e) MH15 2t SAT	.0180	**.0159**	**4.53**	**4.33**	**.0155**	**.0143**	**3.45**	3.57
(a+b) Baseline (bl)	—		—		.0169	.0155	3.45	3.35
(bl+c)	—		—		.0149	.0147	3.13	3.12
(bl+d)	—		—		.0154	.0148	3.12	3.03
(bl+e)	—		—		.0126	.0118	2.57	**2.43**
(bl+c+d)	—		—		.0142	.0140	2.68	2.85
(bl+c+e)	—		—		.0114	.0114	2.25	2.57
(bl+c+d+e) All	—		—		**.0113**	**.0114**	**2.20**	2.48

B. System Fusion

The best performing GSV-SVM, CMLLR-SVM and MLLR-SVM systems, i.e., GSV-SVM 256 g, CH15, and MH15 2t SAT, were selected from Tables I, III, and IV for combination with two other standard cepstral systems, PLP-GMM and PLP-SVM, previously described in Section V-C. Table V shows individual system results for the SRE 2005 and SRE 2006 data and fusion results for SRE 2006 only. The SRE 2005 data was used to train the fusion model and excluded from the evaluation.

The GSV-SVM 256 g system is the best performing of the non-MLLR-based individual systems. The MH15 2t SAT system outperforms the rest of the individual system overall, with relative gains of at least 15% in MDC. In EER terms, the gains are more variable, from 2% to 20%, with GSV-SVM eventually outperforming MH15 2t SAT. A large difference in performance is observed for the PLP-GMM system using forward and forward–backward scoring, the latter improving around 20% in MDC and EER. DET curves of all the individual systems on the SRE 2006 data are shown in Fig. 4 (left). Performance of the CH15 system lies far away from the rest of the individual systems, while MH15 2t SAT outperforms all the systems in the low false-alarm rate region. Similar performance is obtained for the GSV-SVM and MH15 2t SAT systems for operating points around the EER.

As for fusion, the baseline system the combination of the PLP-GMM and PLP-SVM systems. In global terms, adding any one of the three MLLR-based systems to the baseline improves performance. The GSV-SVM 256 g system and the CH15 one bring slight improvements to the baseline, while fusing the baseline with the MH15 2t SAT system brings a relative gain of over 23% in MDC and EER regardless of the scoring approach. This suggests that Gaussian supervectors are somewhat redundant with respect to the baseline system, given that the PLP-GMM already uses the GMM and the PLP-SVM uses the SVM. On the CMLLR side, although the CH15 system is much less performant than GSV-SVM, the fusion of the baseline with any of

[26]The choice of the classes may have an effect on performance as well. Consonants and vowels are used in 2t systems based on a priori phonetic criteria whereas semiautomatic clustering is used for 3t systems.

these systems results in similar improvements. This can be interpreted as the CMLLR transforms providing complementary information. The combination of the baseline with the GSV-SVM 256 g and CH15 systems brings small gains, especially in EER, while including the MH15 2t SAT system dominates performance, once again obtaining relative gains over 20%. This effect is clearly shown in Fig. 4 (right) where fusing all the individual systems performance leads to only a slight improvement of performance.

VIII. Conclusion

We studied two approaches to feature extraction for speaker recognition based on two speaker adaptation techniques, namely Gaussian supervectors using MAP adaptation and MLLR transforms. Our experiments showed that 1) an approach using MLLR transform features classified using a SVM is an actual alternative to current state-of-the-art acoustic systems. Using features optimized for speaker recognition, the MLLR-SVM systems outperformed all other acoustic systems at the MDC operating point, including a likelihood-ratio-based GMM-UBM system using hybrid factor analysis inter-session compensation and a system using Gaussian supervector features and SVM classification. The channel-compensated front-end seems to prevent the transform coefficients from capturing the channel variability. Speaker adaptive training and multiple regression classes were found to improve performance for the most advanced adaptation schemes. 2) For the most simple setups, systems based on CMLLR were found to outperform those based on MLLR, with speaker-recognition features and SAT bringing large improvements. 3) The use of phonemic HMM instead of a GMM for adaptation results in interesting gains in performance. However, these gains should be balanced against the increase of training complexity and resources. 4) The GSV-SVM system outperforms both the PLP-GMM and PLP-SVM systems, showing the effectiveness of including both GMM and SVM classification into one system. 5) Fusion improvements are dominated by the baseline and MLLR-SVM system performances. The GSV-SVM and CMLLR-SVM systems bring about the same improvement after fusion while obtaining very different performance individually. This suggests that MLLR transform coefficients involve information that is complementary to that of GMM mean vectors. Including the GSV-SVM system in a fusion already using GMM and MLLR transform coefficients does not bring any further gain.

References

[1] S. Kajarekar, N. Malayath, and H. Hermansky, "Analysis of speaker and channel variability in speech," in *Proc. IEEE Workshop Speech Recognition and Understanding*, Dec. 1999, pp. 59–62.

[2] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. 53–56.

[3] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proc. IEEE Speaker Odyssey Workshop*, Jun. 2004, pp. 219–226.

[4] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc. IEEE Speaker Odyssey Workshop*, 2004, pp. 57–62.

[5] P. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process.*, vol. 10, pp. 42–54, 2000.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[7] J. L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Guassian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[8] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Lisbon, Portugal, 2005, pp. 3117–3120.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[10] N. Dehak, R. Dehak, P. Kenny, and S. Sridharan, "Comparison between factor analysis and GMM support vector machines for speaker verification," in *Proc. IEEE Speaker Odyssey Workshop*, Stellenbosch, South Africa, 2008, paper 009.

[11] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[12] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. Eur. Conf. Speech Commun. Technol. (EUROSPEECH)*, Sep. 2005, pp. 2425–2428.

[13] A. Stolcke, L. Ferrer, and S. Kajarekar, "Improvements in MLLR-transform-based speaker recognition," in *Proc. IEEE Speaker Odyssey Workshop*, Jun. 2006, pp. 1–6.

[14] M. Ferras, C. C. Leung, C. Barras, and J.-L. Gauvain, "MLLR techniques for speaker recognition," in *Proc. IEEE Speaker Odyssey Workshop*, Jan. 2008, pp. 21–24.

[15] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Philadelphia, PA, 1996, vol. 2, pp. 1137–1140.

[16] M. Ferras, C. C. Leung, C. Barras, and J.-L. Gauvain, "Constrained MLLR for speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2007.

[17] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[18] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, no. 4, pp. 249–264, Oct. 1996.

[19] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA Spoken Lang. Technol. Workshop*, 1995, pp. 110–115.

[20] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

[21] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 1, pp. 121–167, 1998.

[22] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Commun. Technol. (EUROSPEECH)*, 1997, vol. 4, pp. 1895–1898.

[23] J. Pelecanos and S. Sridharan, "Feature warping for speaker verification," in *Proc. IEEE Speaker Odyssey Workshop*, 2001, pp. 213–218.

[24] J. L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. 212–215.

[25] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 97–100.

[26] Z. N. Karam and W. M. Campbell, "A multi-class MLLR kernel for SVM speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2008, pp. 4117–4120.

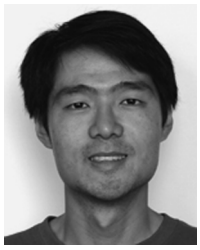
[27] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, pp. 161–164.

[28] B. Scholkopf, A. Smola, and K. R. Muller, "Kernel principal component analysis," in *Advances in Kernel Methods-Support Vector Learning*. Cambridge, U.K.: MIT Press, 1999.

[29] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, April 2003, pp. II-49–II-52.
[30] D. Matrouf, N. Scheffer, B. Fauve, and F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2007, pp. 1242–1245.
[31] J.-F. Bonastre, F. Wals, and S. Meigner, "Alize, a free toolkit for speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2005, pp. 737–740.
[32] N. Brummer, "The Spescom DataVoice and University of Stellenbosch NIST SRE 2005 System," in *NIST Speaker Recognition Workshop*, Jun. 2005.
[33] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2072–2084, Sep. 2007.

Marc Ferras received the B.S. degree in computer science, the M.S. degree in telecommunications, and the European Masters in language and speech from the Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, in 1999 and 2005, respectively. He is currently pursuing the Ph.D. degree at the Spoken Language Processing Group, LIMSI-CNRS, Orsay, France.

He has researched mobile robotics at the Automatic Control Department, ESAII, UPC, single-microphone speech enhancement techniques at the Signal Theory and Communications Department, ETSETB, UPC, and multiple-microphone de-reverberation techniques while visiting the International Computer Science Institute (ICSI), Berkeley, CA, within the AMI program. His current research interests are speaker recognition, speech recognition and machine learning.



Cheung-Chi Leung received the B.Eng. degree from the Hong Kong University of Science and Technology in 1999, and the M.Phil. and Ph.D. degrees from the Chinese University of Hong Kong in 2001 and 2004, respectively.

He then was a Postdoctoral Researcher at the Spoken Language Processing Group, CNRS-LIMSI, Orsay, France. He is currently a Research Fellow at the Human Language Technology Department, Institute for Infocomm Research (I2R), Singapore.

Claude Barras graduated from the Ecole Supérieure d'Electricité Supélec, Gif-sur-Yvette, France, and received the Ph.D. degree in computer science from the University Pierre-et-Marie-Curie, Paris, France, in 1996.

Since 2001, he has been an Associate Professor in the University Paris-Sud, Paris, France, and works in the Spoken Language Processing Group at LIMSI-CNRS, Orsay, France. His research fields include speech recognition, speech annotation, and speaker recognition.

Jean-Luc Gauvain received the Ph.D. degree in electronics from the University of Paris XI, Paris, France, in 1982.

He is a senior researcher at the CNRS where he is head of the LIMSI Spoken Language Processing Group, Orsay, France. His primary research centers on large-vocabulary continuous speech recognition and audio indexing. His research interests also include conversational interfaces, speaker identification, language identification, and speech translation. He has participated in many speech-related projects both at the French National and European levels and has led the LIMSI participation in the DARPA/NIST organized evaluations since 1992, most recently for the transcription of broadcast news data and of conversational speech. He has over 220 publications.

Dr. Gauvain received the 1996 IEEE Signal Processing Society Best Paper Award in Speech Processing and the 2004 ISCA Best Paper Award for a paper in the *Speech Communication* journal. He was a member of the IEEE Signal Processing Society's Speech Technical Committee from 1998 to 2001, and has been coeditor-in-chief of the *Speech Communication* journal from 2005 to 2008.