Speaker and Session Variability in GMM-Based Speaker Verification

P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel

EDICS Category: SPE-SPKR

***Abstract*— We present a corpus-based approach to speaker verification in which maximum likelihood II criteria are used to train a large scale generative model of speaker and session variability which we call joint factor analysis. Enrolling a target speaker consists in calculating the posterior distribution of the hidden variables in the factor analysis model and verification tests are conducted using a new type of likelihood II ratio statistic. Using the NIST 1999 and 2000 speaker recognition evaluation data sets, we show that the effectiveness of this approach depends on the availability of a training corpus which is well matched with the evaluation set used for testing. Experiments on the NIST 1999 evaluation set using a mismatched corpus to train factor analysis models did not result in any improvement over standard methods but we found that, even with this type of mismatch, feature warping performs extremely well in conjunction with the factor analysis model and this enabled us to obtain very good results (equal error rates of about 6.2%).**

Index terms: speaker verification, Gaussian mixture, factor analysis

I. Introduction

Simply stated, the basic problem in speaker verification is to decide whether two utterances have been uttered by the same speaker or by different speakers. Put another way, one has to decide whether the differences between the two utterances are better accounted for by inter-speaker variability or by inter-session variability, that is, the variability exhibited by a given speaker from one recording session to another. This type of variability is usually attributed to channel effects although this is not strictly accurate since intra-speaker variation (the speaker's health or emotional state for example) and phonetic variation are also involved.

In state of the art methods of speaker verification, speaker variability is assumed to be of primary importance but it has long been recognized that session variability is a serious problem. In face recognition, it has been found that models of intra-person variability (which capture differences in posture and illumination in different images of the same subject) are capable of good performance even when inter-person variability is not modeled at all [1]. This suggests that a systematic model of session variability could prove to be useful in speaker verification, particularly if it is integrated with an effective model of speaker variability. As a first attempt at this problem, we proposed a model of session variability in [2] which we referred to as eigenchannel MAP. In [3] we showed how this model can be integrated with standard models of speaker variability, namely classical MAP [4] and eigenvoice MAP [5], to produce a model of speaker and session variability which we refer to as *joint factor analysis*. In this article we will present an overview of the factor analysis model as it was originally formulated in [3] and we will explore how it can be applied to text-independent speaker verification. We have explored various simplifications and refinements of the factor analysis model in subsequent work [6], [7], [8], [9], [10].

Our original motivation in developing the factor analysis model was to use model adaptation techniques developed for speech recognition to perform channel adaptation of speaker models in speaker recognition. Two difficulties arise here. Firstly, very little data may be available for channel adaptation. For example, in all but the most recent NIST speaker recognition evaluations (SRE's), test utterance durations in the core condition range from 15 to 45 seconds. Secondly, model adaptation techniques developed for speech recognition conflate inter-speaker and channel variability so that, although they are usually thought of as performing speaker adaptation, they may be performing channel adaptation in some situations and speaker adaptation in others. In order to be effective for speaker recognition, model adaptation techniques must be capable of adapting speaker models to the channel effects in a test utterance *without* adapting them to the speaker in the test utterance.

We have attempted to deal with these problems by modifying the eigenvoice and EMAP (extended MAP) approaches to model adaptation that have been developed in speech recognition specifically in order to deal with situations where very small amounts of adaptation data are available (as in on-line speaker adaptation [11]). These methods have not been widely used in speaker recognition so we will give a brief description of them here. Applied to the problem of estimating speaker-dependent Gaussian mixture models (GMM's), EMAP requires specifying a prior probability distribution on the GMM mean vectors. Equivalently, it requires specifying a prior distribution on GMM supervectors where a GMM supervector is defined by concatenating the mean vectors associated with the individual Gaussians in the GMM. This supervector distribution is assumed to be Gaussian; let us denote the mean supervector by $\boldsymbol{m}$ and the supervector covariance matrix

The authors are with the Centre de recherche informatique de Montréal (CRIM); email: patrick.kenny@crim.ca, gilles.boulianne@crim.ca, pierre.ouellet@crim.ca, pierre.dumouchel@crim.ca. This work was supported in part by the Natural Science and Engineering Research Council of Canada and by the Ministère du Développement Économique et Régional et de la Recherche du Gouvernement du Québec. The authors would like to thank the anonymous reviewers whose close reading of manuscript helped to clarify many obscure points.

0000–0000/00$00.00 © IEEE

by $\boldsymbol{B}$. Given adaptation data for a speaker, the posterior distribution for the speaker's supervector can be calculated using $\boldsymbol{m}$ and $\boldsymbol{B}$ as in [5]. Because their role is to specify the prior distribution of the parameter that we want to estimate (namely the speaker's supervector), $\boldsymbol{m}$ and $\boldsymbol{B}$ are known as hyperparameters. Of course these hyperparameters also need to be estimated and this can be done by maximizing a likelihood function whose arguments are the hyperparameters [5]. This approach to hyperparameter estimation has come to be known as *maximum likelihood II* in the general machine learning literature [12].

Classical MAP corresponds to the special case where $\boldsymbol{B}$ is taken to be diagonal and estimated empirically. The advantage of EMAP over classical MAP is that it takes account of the correlations between different Gaussians in a speaker model. Thus, whereas classical MAP only adapts the Gaussians which are observed in the adaptation data, EMAP adapts all of the Gaussians even in situations where only a small fraction of them are observed.

Eigenvoice methods are based on the assumption that the supervector covariance matrix $\boldsymbol{B}$ is full but of low rank so that speaker supervectors are constrained to lie in a linear manifold of low dimension which is known as the *speaker space*. This type of constraint facilitates very rapid speaker adaptation since only a small number of free parameters need to be estimated, namely the coordinates of a speaker's supervector relative to a basis of the speaker space. (The eigenvectors of $\boldsymbol{B}$ which correspond to non-zero eigenvalues — the 'eigenvoices' — constitute such a basis.) We will refer to these free parameters as *speaker factors*.

Combining the eigenvoice assumption with EMAP gives eigenvoice MAP [5]. This type of model adaptation can be modified to tackle the problem of channel adaptation of speaker models for speaker recognition by assuming that the channel-dependent supervectors for different recordings of each speaker have a Gaussian distribution centered on the speaker's supervector. If the covariance matrices of these speaker-dependent distributions are tied across all speakers and $\boldsymbol{C}$ denotes the common value, then $\boldsymbol{C}$ can be estimated by the same methods as the supervector covariance matrix $\boldsymbol{B}$ in eigenvoice MAP. This is the basic idea in eigenchannel MAP [2]. Our experience has been that the eigenvalues of $\boldsymbol{C}$ (like those of $\boldsymbol{B}$) decay rapidly, so $\boldsymbol{C}$ can be taken to be of low rank in practice. This makes it possible to perform channel adaptation of speaker models on very short test utterances. Since the supervectors which account for inter-session variation all lie in the range of $\boldsymbol{C}$, it is natural to think of the range of $\boldsymbol{C}$ as the *channel space* and to define *channel factors* analogously to speaker factors.

The development of eigenchannel MAP in [2] was incomplete because it addressed the first of the following questions but not the second:

1) How is it possible to adapt a speaker model to the channel effects in a test utterance without performing speaker adaptation?
2) How is it possible to estimate a speaker model in a way which is immune to the channel effects in the speaker's enrollment data?

In order to provide an answer to the second question, it seems to be necessary to integrate eigenchannel MAP with a model of inter-speaker variability. The simplest possibility is to use the prior in eigenvoice MAP for this purpose; this is the basic idea underlying the factor analysis model. Thus we assume that each speaker- and channel-dependent supervector can be decomposed into a sum of two supervectors, one of which lies in the speaker space and the other in the channel space. Given an enrollment recording for a speaker we can disentangle the speaker and channel effects in the corresponding speaker- and channel-dependent supervector by calculating the joint posterior distribution of the speaker and channel factors. Suppressing the contribution of the channel factors to the supervector gives (in theory at least) an estimate of the speaker's supervector which is immune to the channel effects in the enrollment recording and hence an answer to the second question above. In formulating the factor analysis model we actually took this idea one step further by incorporating the prior for classical MAP as well as the prior for eigenvoice MAP in order to compensate for the rank deficiency problem in eigenvoice MAP [5]. Posterior calculations and maximum likelihood II training algorithms for the joint factor analysis model are worked out in detail in [3].

Note that the factor analysis model is quite similar in spirit to feature mapping. In [13], the basic assumption is that each speaker- and channel-dependent supervector is a sum of a speaker-dependent supervector and a channel-dependent supervector. The major difference is that the factor analysis model treats the channel space as a continuum whereas in [13] channel effects are quantized so that there is a discrete set of channel supervectors (one for electret handsets, another for carbon and so forth). For this approach, the second question above presents no particular difficulty since it can be tackled by applying the appropriate type of channel compensation in enrollment as well as in testing [13].

In undertaking the present work, our aim was to conduct speaker verification experiments on one or more of the NIST speaker recognition evaluation sets using a new type of likelihood II ratio statistic derived from the joint factor analysis model. This entails as a first step fitting the joint factor analysis model (using the maximum likelihood II criterion) to a large training corpus in which there are several recordings of each speaker, such as the corpora distributed by the Linguistic Data Consortium (LDC). In order to do this type of experiment properly, the training corpus and evaluation set need to be disjoint but reasonably well matched with respect to both speaker and channel characteristics. Here we ran into a difficulty which we had not anticipated, namely that the NIST evaluation sets had been collected in such a way as to make it practically impossible to fulfill this requirement prior to 2005 (as we will explain). On the other hand, for both the 2004 and 2005 evaluations, the evaluation data was drawn from a common source, namely the Mixer collection, which was designed specifically to stimulate research in channel modeling for speaker recognition [14]. It became possible to experiment properly with the factor analysis model in 2005 because, for purposes of testing on the 2005 evaluation data, the 2004 evaluation data can serve as a training corpus. So, although

the current work was done in 2004, we decided not to submit it for publication until we were in a position to produce results on the 2005 data set (in the companion paper [7]).

Our main concern in this article was to see if a large scale factor analysis model of speaker and session variability (having up to 500 speaker factors and 100 channel factors and hence several hundred times as many free parameters as in the standard GMM/UBM approach to speaker verification) could be successfully trained on corpora containing hundreds of hours of data using the maximum likelihood II prescriptions in [3]. Secondly, we wanted to see if a likelihood II ratio statistic derived from such a model could be used successfully in speaker verification. We adopted this approach to constructing a likelihood ratio statistic because it enabled us to tackle the problem of channel compensation in a much more sophisticated way than eigenchannel MAP: firstly, it takes account of channel effects in a target speaker's enrollment data as well as in a test utterance (by integrating over all possible values of the channel factors in each case, rather than by using point estimates); and secondly, it takes account of the uncertainty of the target speaker's location in the supervector space that results from the fact that the enrollment data is of limited duration (by integrating over the posterior distribution of the speaker factors, where the posterior distribution is calculated from the speaker's enrollment data).

We report the results of experiments on the NIST 1999 and 2000 evaluation sets in this article. We investigated some methods to mitigate the mismatch problem, even though these methods violate the NIST evaluation protocol. The main idea here was to try to adapt a joint factor analysis model to a given target speaker population using the enrollment data (but not the test data) for the target speakers. To facilitate this we also estimated universal background models on the enrollment data in our experiments on the 1999 evaluation set (but not in the case of the 2000 evaluation set). For these reasons, our results are not strictly comparable with those reported by other authors on these evaluation sets. (But note that the results on the NIST 2005 evaluation set in [7] were obtained without any violations of the NIST protocol.) As it turns out, the gains in performance obtained by this type of adaptation from one speaker population to another were very minor. On the other hand we found that substantial improvements (30% reductions in error rates) could be obtained by using feature warping [15], [16] in conjunction with the joint factor analysis model, and this result convinced us to continue to develop the model despite the obstacles we had encountered with it initially.

The article is organized as follows. In Section II we describe the factor analysis model and the likelihood II function using the same notation as in [3]. We briefly describe the maximum likelihood II procedures for estimating the hyperparameters from training corpora and for adapting them from one speaker population to another and we explain how enrolling a target speaker reduces to calculating the posterior distribution of the hidden variables in the factor analysis model. In Section III we explain how to construct the likelihood II ratio statistic that we used for our speaker verification experiments. Section IV explains how we chose the training corpora and evaluation sets for our experiments (and why this was problematic prior to 2005). Sections V and VI describe how we conducted our experiments and we conclude with a discussion of the results in Section VII.

II. Factor Analysis

The factor analysis model combines the priors underlying classical MAP, eigenvoice MAP and eigenchannel MAP, so we begin by reviewing these and showing how a single prior can be constructed which embraces all of them. We assume a fixed GMM structure containing a total of C mixture components. Let F be the dimension of the acoustic feature vectors.

A. Speaker and channel factors

Our basic assumption is that a speaker- and channel-dependent supervector can be decomposed into a sum of two supervectors, one of which depends on the speaker and the other on the channel, and that speaker supervectors and channel supervectors are statistically independent and normally distributed. The dimensions of the covariance matrices of these distributions are enormous $(CF \times CF)$ so we have to explain how we propose to model these covariance matrices.

Let $\boldsymbol{M}(s)$ be the speaker supervector for a speaker s and let $\boldsymbol{m}$ denote the speaker- and channel-independent supervector. (The simplest way to estimate $\boldsymbol{m}$ is to take the supervector from a Universal Background Model (UBM).) In classical MAP it is assumed that, for a randomly chosen speaker s, $\boldsymbol{M}(s)$ is normally distributed with mean $\boldsymbol{m}$ and a diagonal covariance matrix $\boldsymbol{d}^2$. It is convenient to describe this prior in terms of hidden variables as follows:

$$\boldsymbol{M}(s) = \boldsymbol{m} + \boldsymbol{d}\boldsymbol{z}(s), \tag{1}$$

where $\boldsymbol{z}(s)$ is a hidden vector distributed according to the standard normal density, $N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$. (It is easily seen that, under this assumption, the expectation of $\boldsymbol{M}(s)$ is $\boldsymbol{m}$ and its covariance is $\boldsymbol{d}^2$.)

Provided that $\boldsymbol{d}$ is non-singular, MAP speaker adaptation using this prior distribution is guaranteed to be asymptotically equivalent to maximum likelihood estimation of speaker models as the amount of adaptation data increases. However, there does not seem to be any principled reason for assuming that the covariance matrix is diagonal. Treating mixture components in a speaker model as being statistically independent has the disadvantage that MAP adaptation can only update mixture components which are observed in the adaptation data. Thus, if the number of mixture components C is large, classical MAP tends to saturate slowly in the sense that large amounts of enrollment data are needed to use it to full advantage.

Eigenvoice MAP assumes instead that there is a rectangular matrix $\boldsymbol{v}$ of dimensions $CF \times R$ where $R \ll CF$ such that, for a randomly chosen speaker s,

$$\boldsymbol{M}(s) = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s), \tag{2}$$

where $\boldsymbol{y}(s)$ is a hidden $R \times 1$ vector having a standard normal distribution. Since the dimension of $\boldsymbol{y}(s)$ is much smaller than that of $\boldsymbol{z}(s)$, eigenvoice MAP tends to saturate much more quickly than classical MAP. But this approach to speaker adaptation suffers from the drawback that, in estimating $\boldsymbol{v}$

from a given training corpus, it is necessary to assume that R is less than or equal to the number of training speakers [5], so that a large number of training speakers may be needed to estimate $\boldsymbol{v}$ properly. Thus in practice there is no guarantee that eigenvoice MAP adaptation will exhibit correct asymptotic behavior as the quantity of enrollment data for a speaker increases. No matter how much enrollment data is made available, the eigenvoice MAP estimate of the speaker's supervector is constrained to lie in the subspace spanned by the training speakers' supervectors even if the 'true' speaker supervector lies elsewhere.

The strengths and weaknesses of classical MAP and eigenvoice MAP complement each other. (Eigenvoice MAP is preferable if small amounts of data are available for speaker adaptation and classical MAP if large amounts are available.) An obvious strategy to combine the two is to assume a decomposition of the form

$$\boldsymbol{M}(s) = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s) + \boldsymbol{d}\boldsymbol{z}(s), \tag{3}$$

where $\boldsymbol{y}(s)$ and $\boldsymbol{z}(s)$ are assumed to be independent and to have standard normal distributions. In other words, $\boldsymbol{M}(s)$ is assumed to be normally distributed with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{v}\boldsymbol{v}^* + \boldsymbol{d}^2$. This is a factor analysis model in the sense of [17]. The components of $\boldsymbol{y}(s)$ are *common speaker factors* and the components of $\boldsymbol{z}(s)$ are *special speaker factors*; $\boldsymbol{v}$ and $\boldsymbol{d}$ are *factor loading matrices*. The *speaker space* is the affine space defined by translating the range of $\boldsymbol{v}\boldsymbol{v}^*$ by $\boldsymbol{m}$. If $\boldsymbol{d} = \boldsymbol{0}$, then all speaker supervectors are contained in the speaker space; in the general case ($\boldsymbol{d} \neq \boldsymbol{0}$) the term $\boldsymbol{d}\boldsymbol{z}(s)$ serves as a residual which compensates for the fact that this type of subspace constraint may not be realistic. This type of prior distribution has been used as a basis for speaker adaptation in both speech recognition [18] and speaker recognition [19].

In order to incorporate channel effects, suppose we are given recordings $h = 1, \ldots, H(s)$ of a speaker s. For each recording h, let $\boldsymbol{M}_h(s)$ denote the corresponding speaker- and channel-dependent supervector. We assume as in [2] that the difference between $\boldsymbol{M}_h(s)$ and $\boldsymbol{M}(s)$ can be accounted for by a vector of common channel factors $\boldsymbol{x}_h(s)$ having a standard normal distribution. That is, we assume that there is a rectangular matrix $\boldsymbol{u}$ of low rank (the loading matrix for the channel factors) such that

$$\left.\begin{array}{rcl} \boldsymbol{M}(s) & = & \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y}(s) + \boldsymbol{d}\boldsymbol{z}(s) \\ \boldsymbol{M}_h(s) & = & \boldsymbol{M}(s) + \boldsymbol{u}\boldsymbol{x}_h(s) \end{array}\right\} \tag{4}$$

for each recording $h = 1, \ldots, H(s)$. Note that the speaker factors and the channel factors play different roles, in that the speaker factors are assumed to have the same values for all recordings of the speaker whereas the channel factors vary from one recording to another.

Thus we are assuming that channel supervectors are contained in a low-dimensional subspace of the supervector space, namely the range of $\boldsymbol{u}\boldsymbol{u}^*$, which we refer to as the *channel space*. The rationale for this assumption is that it has invariably been our experience that the eigenvalues of $\boldsymbol{u}\boldsymbol{u}^*$ decay rapidly so there is little loss in accuracy in assuming that $\boldsymbol{u}$ is of low rank. Given a random symmetric positive definite matrix, there is no reason why its eigenvalues should decay rapidly to zero, yet this phenomenon is frequently observed in physics and engineering. A plausible explanation for this is that, since the eigenvectors produced in the Karhunen-Loève expansion of a physical signal are orthogonal, the energies in the directions of these eigenvectors are additive. The average energy in each of these directions is just the corresponding eigenvalue, so since the total energy of the signal is finite, the sum of the eigenvalues must converge. In order for this series to converge, the eigenvalues have to tend to zero rapidly.

Thus the hypothesis that the covariance matrix for channel compensation supervectors is of low rank (or, equivalently, that the channel space is of low dimension) is a plausible one. Indeed, it has been our experience that incorporating special channel factors (analogous to the diagonal term in (3)), which would result in a covariance matrix of full rank, hurts performance in speaker verification. If the channel covariance matrix were really of full rank then it would be possible to make one speaker sound like any other by varying the channel conditions. This would be very bad news for speaker recognition! Of course, there are grounds for questioning the assumption that the channel compensation supervectors are normally distributed in the channel space. The success of the method of feature mapping [13] suggests that the correct distribution may be multimodal rather than unimodal, so that a Gaussian mixture may be the most appropriate way to model it. This question is addressed in [9].

Thus, in its current form, the joint factor analysis model is specified as follows. If R_C is the number of channel factors and R_S the number of speaker factors, the model is specified by a quintuple $\boldsymbol{\Lambda}$ of the form $(\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{d}, \boldsymbol{\Sigma})$ where $\boldsymbol{m}$ is $CF \times 1$, $\boldsymbol{u}$ is $CF \times R_C$, $\boldsymbol{v}$ is $CF \times R_S$, and $\boldsymbol{d}$ and $\boldsymbol{\Sigma}$ are $CF \times CF$ diagonal matrices. To explain the role of $\boldsymbol{\Sigma}$, fix a mixture component c and let Σ_c be the corresponding block of $\boldsymbol{\Sigma}$. For each speaker s and recording h, let $M_{hc}(s)$ denote the subvector of $\boldsymbol{M}_h(s)$ corresponding to the given mixture component. We assume that, for all speakers s and recordings h, observations drawn from mixture component c are distributed with mean $M_{hc}(s)$ and covariance matrix Σ_c.

In the case $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{u} = \boldsymbol{0}$ the factor analysis model reduces to the prior for eigenvoice MAP. In the case where $\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{0}$ we obtain the prior for classical MAP.[1] If we assume that $\boldsymbol{M}(s)$ has a point distribution instead of the Gaussian distribution specified by (1) and that this point distribution is different for different speakers we obtain the prior for eigenchannel MAP.

The special speaker factors $\boldsymbol{z}(s)$ are included in the model in order to ensure that it inherits the asymptotic behavior of classical MAP, but they are costly in terms of computational complexity. The reason for this is that, although the increase

[1] If $\boldsymbol{d}^2$ is assumed to be related $\boldsymbol{\Sigma}$ by an equation of the form

$$\boldsymbol{d}^2 = \frac{1}{r}\boldsymbol{\Sigma}$$

where r is a 'relevance factor' [4], then the classical MAP estimation formulas are easily seen to be a special case of Proposition 2 in [5] (if $\boldsymbol{v}$ is replaced by $\boldsymbol{d}$). Thus, although its role is rarely spelled out explicitly, the diagonal matrix $\boldsymbol{d}$ is the key to the success of the GMM/UBM approach to speaker verification.

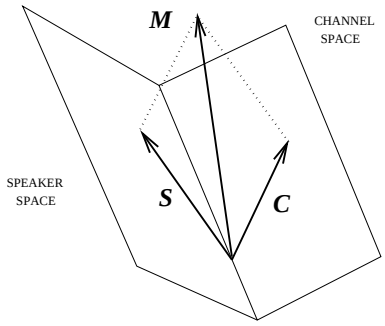

Fig. 1. *In the PCA case, a speaker- and channel-dependent supervector $\boldsymbol{M}$ can be written as a sum of two supervectors, one of which ($\boldsymbol{S}$) lies in the speaker space and the other ($\boldsymbol{C}$) lies in the channel space (in accordance with the parallelogram rule). In the general case, speaker supervectors are distributed in the neighborhood of the speaker space.*

in the number of free parameters is relatively modest since (unlike $\boldsymbol{u}$ and $\boldsymbol{v}$) $\boldsymbol{d}$ is assumed to be diagonal, introducing $\boldsymbol{z}(s)$ greatly increases the number of hidden variables. It is also a major source of complication in [3]; for example factor analysis models do not form a conjugate family unless $\boldsymbol{d} = \boldsymbol{0}$. We will use the term Principal Components Analysis (PCA) to refer to the case where $\boldsymbol{d} = \boldsymbol{0}$. The model is quite simple in this case since the basic assumption is that each speaker- and channel-dependent supervector is a sum of two supervectors, one of which is contained in the speaker space and the other in the channel space. This decomposition is actually unique since the range of $\boldsymbol{u}\boldsymbol{u}^*$ and the range of $\boldsymbol{v}\boldsymbol{v}^*$, being low dimensional subspaces of a very high dimensional space, (typically) only intersect at the origin. (The representation in Fig. 1 is slightly misleading because it suggests that the intersection of the speaker and channel spaces is of positive dimension.)

B. The factor analysis likelihood II function

Suppose that we are given a hyperparameter set $\boldsymbol{\Lambda}$ and a set of recordings for a speaker s indexed by $h = 1, \ldots, H(s)$. For each recording h, assume that each observation vector has been aligned with a mixture component as in a Viterbi alignment and let $\mathcal{X}_h(s)$ denote the collection of labeled frames for the hth recording. Let $\underline{\mathcal{X}}(s)$ be the vector obtained by concatenating the collections of labeled frames $\mathcal{X}_1(s), \ldots, \mathcal{X}_{H(s)}(s)$ for all of the speaker's recordings; these are the observable variables for the factor analysis model. Let $\underline{\boldsymbol{X}}(s)$ be the vector obtained by concatenating the hidden variables $\boldsymbol{x}_1(s), \ldots, \boldsymbol{x}_{H(s)}(s), \boldsymbol{y}(s), \boldsymbol{z}(s)$. (As in [3] we use under bars when referring to collections of recordings rather than to an individual recording.)

If $\underline{\boldsymbol{X}}(s)$ were given, we could write down $\boldsymbol{M}_h(s)$ and calculate the (Gaussian) likelihood of $\mathcal{X}_h(s)$ for each recording h, so the calculation of the likelihood of $\underline{\mathcal{X}}(s)$ would be straightforward. Let us denote this conditional likelihood by $P_{\boldsymbol{\Lambda}}(\underline{\mathcal{X}}(s)|\underline{\boldsymbol{X}}(s))$. Since the values of the hidden variables are not given, calculating the likelihood of $\underline{\mathcal{X}}(s)$ requires evaluating the integral

$$\int P_{\boldsymbol{\Lambda}}(\underline{\mathcal{X}}(s)|\underline{\boldsymbol{X}}) N(\underline{\boldsymbol{X}}|\underline{\boldsymbol{0}}, \underline{\boldsymbol{I}}) d\underline{\boldsymbol{X}}, \tag{5}$$

where $N(\underline{\boldsymbol{X}}|\underline{\boldsymbol{0}}, \underline{\boldsymbol{I}})$ is the standard Gaussian kernel

$$N(\boldsymbol{x}_1|\boldsymbol{0}, \boldsymbol{I}) \cdots N(\boldsymbol{x}_{H(s)}|\boldsymbol{0}, \boldsymbol{I}) N(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{I}) N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}).$$

We denote the value of this integral by $P_{\boldsymbol{\Lambda}}(\underline{\mathcal{X}}(s))$. A closed form expression for this integral is given in Theorem 3 in [3] in terms of the Viterbi statistics of the various utterances.[2]

It is quite common in speaker recognition to use Viterbi-type approximations (particularly if the number of Gaussians in the UBM is large), but strictly speaking, this is not really satisfactory — the correct procedure would be to sum over all possible alignments of observations with mixture components. However, summing over all possible alignments in evaluating the factor analysis likelihood function would be computationally intractable. In the case where $\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{0}$ the problem can be solved by dynamic programming [20], [21] but this approach is not computationally feasible in the general case (unless the number of speaker factors and channel factors are constrained to be unrealistically small). We dealt with this problem in our implementation by substituting Baum-Welch statistics for Viterbi statistics in evaluating (5). (We used the same expedient in [5].)

C. Speaker-independent hyperparameter estimation

If we are given a training corpus in which each speaker is recorded in multiple sessions, the hyperparameters $\boldsymbol{\Lambda}$ can be estimated by EM algorithms which guarantee that the total likelihood of the training data increases from one iteration to the next. (The total likelihood of the training data is $\prod_s P_{\boldsymbol{\Lambda}}(\underline{\mathcal{X}}(s))$ where s ranges over the speakers in the training corpus. This is a likelihood II function since its arguments are the hyperparameters $\boldsymbol{\Lambda}$.) We refer to these as speaker-independent hyperparameter estimation algorithms (or simply as *training* algorithms) since they consist in fitting (3) to the entire collection of speakers in the training data rather than to an individual speaker. These algorithms are described in Theorems 4, 5 and 7 of [3].

One estimation algorithm, which we will refer to simply as maximum likelihood estimation, can be derived by extending Proposition 3 in [5] to handle the hyperparameters $\boldsymbol{u}$ and $\boldsymbol{d}$ in addition to $\boldsymbol{v}$ and $\boldsymbol{\Sigma}$. Another algorithm can be derived by using the divergence minimization approach to hyperparameter estimation introduced in [22]. This seems to converge much more rapidly but it has the property that it keeps the orientation of the speaker and channel spaces fixed so that it can only be used after maximum likelihood estimation has already been carried out. Minimum divergence estimation seems to produce better eigenvalue estimates than maximum likelihood estimation. This is to be expected, since the only freedom it has is to rotate the eigenvectors in the speaker and channel spaces and scale the eigenvalues.

[2]For a given sequence of observation vectors $Y_1, \ldots, Y_T$, the first and second order Viterbi statistics for each mixture component c are defined as

$$\sum_t Y_t$$
$$\sum_t Y_t Y_t^*$$

where the sums extend over all observations aligned with the given mixture component.

Admittedly, the literature on eigenvoice methods is unclear as to whether very precise estimates of the eigenvalues are needed in practice. Evidence presented in [11] suggests that, in extreme situations where utterances are very short or the number of eigenvoices is large, precise estimates may be helpful, but in most implementations of eigenvoice methods other than [11], [5] the eigenvalues are ignored altogether in speaker adaptation. (This is tantamount to treating all of the non-zero eigenvalues as infinite in eigenvoice MAP.) The most widely used procedure to estimate eigenvoices and eigenvalues relies on MLLR or classical MAP (rather than maximum likelihood estimation) to produce supervectors for the speakers in the training corpus. This tends to produce noisy estimates of the eigenvalues so that it is not always clear which eigenvalues should be treated as being effectively 0. Thus some authors have concluded that eigenvalue-based dimensionality reduction is better avoided altogether [23]; but note that this is only feasible if the number of speakers in a training corpus is quite limited. In eigenchannel MAP on the other hand, dimensionality reduction cannot be avoided, so careful estimation of the eigenvalues certainly seems to be desirable.

D. Adapting from one speaker population to another

The effectiveness of the speaker-independent hyperparameter estimation algorithms in estimating a joint factor analysis model depends critically on the availability of a training corpus in which there are multiple recordings of each speaker — it seems very unlikely that speaker and session effects can ever be broken out using a training corpus in which there is just one recording for each speaker, such as the enrollment data provided by NIST for one of the restricted data SRE's. So in order to test our model on, say, the NIST 1999 SRE data, we need an ancillary training corpus such as the union of Switchboard II, Phases 1 and 2, to train the joint factor analysis model. Thus in practice there may be a mismatch between the training speaker population and the target speaker population.

This raises an issue which seems to be of basic importance for the factor analysis model namely, whether the speaker and session components of the model can be successfully estimated on different training corpora. Theorems 8 and 9 in [3] present two hyperparameter estimation algorithms which attempt to address this problem: one using the maximum likelihood approach and the other using the minimum divergence approach. We will present the results of experiments with both of these algorithms in this article. In these experiments, we first estimate a full set of hyperparameters $\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{d}$ and $\boldsymbol{\Sigma}$ on the ancillary training corpus and then, holding $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$ fixed, re-estimate $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{d}$ on the enrollment data (but not the test data) for the target speakers (Fig. 2). In other words, we keep the hyperparameters associated with channel space fixed and re-estimate only the hyperparameters associated with the speaker space. It turns out to be important to use the divergence minimization rather than the maximum likelihood approach in this situation. That is, it is necessary to keep the orientation of the speaker space fixed as well as that of the channel space, rather than change it to fit the target speaker

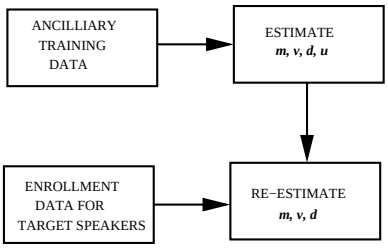

Fig. 2. *A data set such as the enrollment data provided by NIST for one of the restricted data evaluations is not adequate to train a factor analysis model. We estimate the speaker-independent hyperparameters on a much larger ancillary training corpus that contains multiple recordings for each speaker (such as one or more of the Switchboard corpora). In most of our experiments this is followed by adapting the hyperparameters that model inter-speaker variability (namely $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{d}$) to the target speaker population; we assume that channel effects are invariant so we keep $\boldsymbol{u}$ fixed.*

population (in order to avoid over training on the limited amount of enrollment data in a NIST evaluation set). Although this type of adaptation to the target speaker population violates the NIST evaluation protocol, we decided to explore it because of its intrinsic scientific interest.

E. Speaker-dependent hyperparameter estimation

In order to construct the likelihood ratio statistic that we used in our speaker verification experiments, we also need a speaker-dependent hyperparameter estimation algorithm. Recall that in the factor analysis model we have hyperparameters $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{d}$ whose role is to model the distribution of speaker supervectors. The idea in speaker-dependent hyperparameter estimation is that if we are given some enrollment data for a particular speaker s, we can use this data to calculate the posterior distribution of the hidden variables $\boldsymbol{y}(s)$ and $\boldsymbol{z}(s)$ which specify the speaker's supervector $\boldsymbol{M}(s)$, and use the hyperparameters $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{d}$ to model the posterior distribution of $\boldsymbol{M}(s)$ instead of modeling the distribution of speakers in the population at large (Fig. 3). This is broadly analogous to the way speaker models are derived by MAP adaptation from a universal background model in the GMM/UBM approach. In our experiments, the procedure for enrolling a target speaker consists in carrying out this type of speaker-dependent hyperparameter estimation.

Thus we assume that, for a given speaker s and recording h,

$$\left.\begin{array}{rcl} \boldsymbol{M}(s) & = & \boldsymbol{m}(s) + \boldsymbol{v}(s)\boldsymbol{y}(s) + \boldsymbol{d}(s)\boldsymbol{z}(s) \\ \boldsymbol{M}_h(s) & = & \boldsymbol{M}(s) + \boldsymbol{u}\boldsymbol{x}_h(s) \end{array}\right\}. \quad (6)$$

That is, we make the hyperparameters $\boldsymbol{m}$, $\boldsymbol{v}$ and $\boldsymbol{d}$ speaker-dependent but we continue to treat $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$ as speaker-independent. (The rationale here is that channel effects should not vary from one speaker to another.)

In order to estimate the speaker-dependent hyperparameters $\boldsymbol{m}(s), \boldsymbol{v}(s)$ and $\boldsymbol{d}(s)$, we find the distribution of the form $\boldsymbol{m}(s) + \boldsymbol{v}(s)\boldsymbol{y} + \boldsymbol{d}(s)\boldsymbol{z}$ (where, as usual, $\boldsymbol{y}$ and $\boldsymbol{z}$ have standard normal distributions) which is closest to the posterior distribution of $\boldsymbol{M}(s)$ in the sense that the Kullback-Leibler divergence is minimized. Thus $\boldsymbol{m}(s)$ is an estimate of the speaker's supervector when channel effects are abstracted

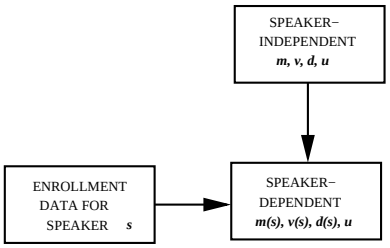

Fig. 3. *For a target speaker s, the speaker-dependent hyperparameters $\boldsymbol{m}(s), \boldsymbol{v}(s)$ and $\boldsymbol{d}(s)$ model the posterior distribution of the speaker's supervector, $\boldsymbol{M}(s)$. This posterior is calculated using the speaker-independent hyperparameters and the speaker's enrollment data.*

and $\boldsymbol{d}(s)$ and $\boldsymbol{v}(s)$ measure the uncertainty in this estimate. This enrollment procedure is just the minimum divergence estimation algorithm applied to a single speaker in the case where $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$ are held fixed. Note that minimum divergence rather than maximum likelihood is the right criterion to use here since, in enrolling a speaker, we want to impose the constraint that the estimate of the speaker's supervector lies in the speaker space. Details are given in Theorem 10 of [3]. For each target speaker s, we will denote the speaker-dependent hyperparameter set $(\boldsymbol{m}(s), \boldsymbol{u}, \boldsymbol{v}(s), \boldsymbol{d}(s), \boldsymbol{\Sigma})$ by $\Lambda(s)$.

III. Likelihood Ratio Statistics

The likelihood II function described in Section II-B can be used to construct likelihood ratio statistics for speaker verification in various ways. We will describe two of these statistics, which we call the *batch* and *sequential* likelihood ratios, in this section. Both of these statistics are similar to the Bayes factors in [20], [21], [24] in that they make allowances for the fact that a target speaker's location in supervector space is uncertain (due to limited enrollment data) but, unlike these Bayes factors, they incorporate mechanisms to compensate for channel effects in the target speakers' enrollment data and in test utterances.

We assume that we are given a collection of one or more enrollment utterances for a target speaker s and a test utterance and that we wish to test the hypothesis that the speaker in the test utterance is s against the null hypothesis that the speaker in the test utterance is somebody else.

The batch likelihood ratio is most easily described in the case where there is just one enrollment recording (the so-called speaker comparison problem), but the extension to multiple enrollment recordings is straightforward. Under the alternative hypothesis, the speaker factors for the enrollment and test recordings are the same and the joint likelihood of the pair of recordings is given by evaluating the integral (5) with $H(s) = 2$. Under the null hypothesis, the two recordings are statistically independent; the likelihood of each recording is given by evaluating the integral (5) with $H(s) = 1$ in each case, and the joint likelihood of the two recordings is just the product of these two integrals. Thus the two hypotheses give rise to different ways of evaluating the joint likelihood of the enrollment and test recordings. The batch likelihood ratio is just the quotient of the two values, and the larger the ratio, the stronger the evidence in favor of the alternative hypothesis. Note that there is no notion of a target speaker model in this construction. All that is needed is to understand how to evaluate the integral (5), but this computation can be prohibitively expensive if the number of speaker factors is large. (A large number of speaker factors is desirable for discriminating between speakers.)

The sequential likelihood ratio statistic is constructed in a more traditional way. Using enrollment recordings for the target speaker s we estimate a speaker-dependent version of the factor analysis model (Section II-E) which models the posterior distribution of the speaker-dependent supervector for the given speaker. This gives us two ways of evaluating the likelihood of the test utterance, namely with the speaker-independent hyperparameters $\boldsymbol{\Lambda}$ and the speaker-dependent hyperparameters $\boldsymbol{\Lambda}(s)$, and hence another way of constructing a likelihood ratio statistic for deciding between the two hypotheses. If $\mathcal{X}$ denotes the collection of labeled observations in the test utterance (each observation being labeled by the corresponding mixture component), this statistic is given in the log domain by

$$\ln \frac{P_{\boldsymbol{\Lambda}(s)}(\mathcal{X})}{P_{\boldsymbol{\Lambda}}(\mathcal{X})}, \tag{7}$$

where the numerator and the denominator are evaluated as in (5) (with $H(s) = 1$ in each case). We refer to this as the sequential log likelihood ratio because it lends itself naturally to progressive speaker adaptation if the speaker-dependent hyperparameter estimation algorithm is applied recursively. (That is, whenever a new recording for a given speaker s becomes available, we update the speaker-dependent hyperparameters $\boldsymbol{m}(s), \boldsymbol{v}(s)$ and $\boldsymbol{d}(s)$ by using the current estimates of these hyperparameters rather than the speaker-independent hyperparameters as the starting point for speaker-dependent estimation.) It is well known that progressive speaker adaptation can improve the performance of speaker recognition systems dramatically and recent NIST SRE's have permitted this avenue to be explored. (We have taken up this question in [10].)

It can be shown that the two likelihood ratios are identical in the case where $\boldsymbol{d} = \boldsymbol{0}$, but in the general case the sequential likelihood ratio is only an approximation to the batch likelihood ratio. The reason for this is that the family of joint factor analysis models is not a conjugate family unless the condition $\boldsymbol{d} = \boldsymbol{0}$ is satisfied (see the discussion preceding Theorem 10 in [3]). For the same reason, the results of applying the speaker-dependent hyperparameter estimation algorithm recursively in evaluating the sequential likelihood ratio (i.e. progressive speaker adaptation) are exactly correct only if $\boldsymbol{d} = \boldsymbol{0}$.

However it is preferable to use the sequential likelihood ratio in practice, because it is easy to find computationally tractable approximations for evaluating it; accordingly, this is the only likelihood ratio that we used in the experiments reported here. To understand the computational issues, consider first the numerator of (7). By Theorem 3 in [3], the computation needed to evaluate the numerator essentially boils down to calculating the Cholesky decomposition of a matrix of dimension $(R_S + R_C) \times (R_S + R_C)$ constructed from $\boldsymbol{v}(s)$

and $\boldsymbol{u}$. (Recall that $\boldsymbol{u}$ is a $CF \times R_C$ matrix and $\boldsymbol{v}$ is a $CF \times R_S$ matrix.) Reducing the rank of $\boldsymbol{v}(s)$ to manageable proportions by discarding the minor eigenvalues of $\boldsymbol{v}(s)\boldsymbol{v}^*(s)$ alleviates the computational burden. Since $\boldsymbol{v}(s)$ captures the uncertainty in the values of the speaker factors after enrollment, most of the eigenvalues of $\boldsymbol{v}(s)\boldsymbol{v}^*(s)$ are small if the amount of enrollment data for the speaker is reasonably large. Even so, this calculation is still computationally demanding so that we were led to use cruder approximations in our subsequent work [7], [8]. We are not in a position to say whether these approximations are deleterious in situations where the amount of enrollment data is not 'reasonably large' (as in the 10 sec enrollment conditions in the NIST SRE's). Although our methods perform well under this condition, we have found that the uncertainty concerning the location of a target speaker in supervector space is very large in this situation — typically about 50% of the variance of the speaker population as a whole. Thus is may be necessary to explore other approximations in the future.

We are confronted by the same problem in evaluating the denominator of (7) as in evaluating the numerator, but this can be avoided altogether by using t-norm score normalization [25]. The idea here is to enroll a collection of t-norm speakers and, at verification time, to calculate the log likelihood ratio (7) for each t-norm speaker in addition to the target speaker s. Let μ and σ be the mean and standard deviation of the log likelihood ratios for the t-norm speakers. The normalization consists in standardizing (7) by subtracting μ and dividing by σ. It is clear that the denominator in (7) drops out of this calculation.

IV. Training Corpora and Evaluation Sets

In order to experiment with the factor analysis model in speaker recognition, we need a large *training* corpus in which each speaker is recorded under a suitable variety of channel conditions for estimating the speaker-independent hyperparameters $\boldsymbol{\Lambda}$, and an evaluation set such as NIST provides in the annual SRE's. These evaluation sets consist of *enrollment* data for each of the target speakers and *test* data which is used to measure speaker verification performance.[3] Naturally, the training corpus and the evaluation set should be disjoint but well matched with respect to both speaker and channel characteristics. Unfortunately, because of the way the NIST evaluation sets were collected in previous years, it was impossible to use them as testbeds for experimenting with factor analysis modeling prior to 2005 without violating this requirement.

[3]A note on terminology: In the general speaker recognition literature the terms training and enrollment are used almost interchangeably. In the context of this work and [7] we use the word training solely to refer to speaker-independent hyperparameter estimation (as described in Section II-C), just as one speaks of training UBM's in the GMM/UBM approach. The terms enrollment and adaptation also tend to be used interchangeably in the GMM/UBM approach (since in that situation, a target speaker model is derived from the UBM by classical MAP adaptation). In this work we use the term adaptation to refer solely to adapting speaker-independent hyperparameters from one population to another (as described in Section II-D). We use the term enrollment to refer to speaker dependent hyperparameter estimation (as described in Section II-E).

An egregious example of what can go wrong if the training corpus is not chosen carefully can be found in [26]. This was our first attempt at implementing the factor analysis model (using a small number of speaker factors and channel factors). We did a series of experiments to address some basic questions such as: Which of the two likelihood ratio statistics constructed with the likelihood II function is the more effective? Is t-norm effective? How many Gaussians should be used? Is a gender-dependent joint factor analysis more effective than a gender-independent joint factor analysis? Is silence detection necessary? These experiments were conducted using the LDC release of Switchboard Cellular Part I as the training corpus and the NIST 2001 SRE cellular data for testing [27]. The results were extraordinarily good but we discovered after the fact that the NIST 2001 cellular evaluation data which was described in [27] as 'drawn from the Switchboard-II Corpus, Phase 4' was actually entirely contained in Switchboard Cellular Part I. Thus our preliminary experiments were flawed and we were forced to abandon this testbed.

Since 2004, the NIST evaluation sets have all been drawn from the Mixer collection, which was designed specifically to stimulate research in channel modeling [14]. In particular, the evaluation sets for 2004 and 2005 were drawn from the same source (without recycling). Thus, by using the 2004 evaluation data as a training corpus (or part thereof), it is possible to experiment properly with the factor analysis model on the 2005 evaluation set — 'properly' in the sense that the training corpus and evaluation set are well matched. The results reported in the companion paper [7] were all obtained on the 2005 evaluation set.

However, prior to 2004 (when the present work was done), each of the NIST evaluation sets was drawn from a different Switchboard corpus (except in cases where data was recycled from one year to the next). Each Switchboard corpus was designed to cover a particular dialect of American English or a particular type of transmission channel (e.g. GSM or CDMA). Thus it was impossible at that time to experiment with the factor analysis model using a NIST evaluation set as a testbed and one or more of the Switchboard corpora for training without encountering a mismatch between training and evaluation conditions. For example, using the NIST 2002 or 2003 SRE data for testing and Switchboard Cellular Part I for training would not be appropriate since the test data consists principally of CDMA transmissions and there would be essentially no CDMA transmissions in the training data. (Switchboard Cellular Part I consists mostly of GSM transmissions.)

So we decided that we would just have to live with the mismatch problem until the 2005 evaluation set was made available. For our first experiments in this article we chose the Switchboard II, Phases 1 and 2 corpora for training factor analysis models and the 1999 evaluation data (which is extracted from the Switchboard II, Phase 3 corpus) for testing. The Switchboard II corpora consist of land line data with roughly equal proportions of 'same number' and 'different number' calls, so there should be no mismatch where channel characteristics are concerned. However, there is a mismatch between the training and target speaker populations because

the training speakers are from the American Midwest and Northeast and the target speakers from the South. We found that, under these conditions, speaker verification with the sequential likelihood ratio statistic yielded results which are as good as (but no better than) the best results that have been attained with standard GMM likelihood ratios and handset detection using unwarped cepstral coefficients as acoustic features.

Our efforts to improve on these results by adapting the joint factor analysis model to the target speaker population using the strategy outlined in Section II-D were largely unrewarded, but we did find that much better performance could be achieved by using Switchboard II, Phase 3 as the training corpus. Of course, this is an unfair experiment because, in this case, the evaluation set is contained in the training corpus. However it suggested that it would be useful to design an experiment with a well matched training corpus and evaluation set.

We used the NIST 2000 evaluation data for this purpose. The 2000 SRE data set is unusually large since it involves a thousand target speakers. These speakers were drawn from both Phase 1 and Phase 2 of Switchboard II and the test utterances were all extracted from different number calls. We constructed the evaluation set for our experiment on this data set by discarding every second target speaker in the 2000 evaluation and we constructed a training corpus using speakers in Switchboard II, Phases 1 and 2 which were not included in our evaluation set. We found that the joint factor analysis model performed well in this situation, which led us to conclude that the mismatch problem was indeed of critical importance. Nonetheless, we were eventually able to achieve very good results on the NIST 1999 data by using feature warping in conjunction with factor analysis modeling (error rate reductions of about 30%).

V. Implementation issues

Using large numbers of speaker and/or channel factors creates problems in speaker-independent estimation of the hyperparameters. The principal computational bottleneck here is in calculating the posterior distribution of the hidden variables (that is, $P_{\mathbf{\Lambda}}(\underline{\boldsymbol{X}}(s)|\underline{\mathcal{Y}}(s))$ for each training speaker s). If the number of recordings of the speaker is large (the ideal situation) and there are large numbers of speaker factors and channel factors, then this calculation may not be practically feasible unless $\boldsymbol{d} = \mathbf{0}$. (Recall that we use the term PCA to refer to this case.) In our preliminary experiments on the NIST 2001 cellular evaluation data using the Switchboard Cellular Part I corpus for training [26], we did not encounter any difficulties in this case because we used only 40 speaker factors and 40 channel factors, but for most of our experiments on the NIST 1999 and NIST 2000 test data we did have to impose the restriction $\boldsymbol{d} = \mathbf{0}$.[4] Calculating the posterior distribution of the hidden variables is only problematic in training the joint factor analysis model on large training sets; there is no difficulty in introducing $\boldsymbol{d}$ in adapting the speaker-independent hyperparameters from the training speaker population to a NIST target speaker population (since the enrollment data for each target speaker consists of just one or two recordings).

As we explained in Section II-B and Section III, evaluating the likelihood II ratio statistic for a given test utterance and a given set of hypothesized target speakers requires extracting the first and second order statistics from the test data using a Viterbi or Baum-Welch alignment. We used Baum-Welch alignments and gender-dependent UBM's for this purpose. An obvious advantage of using only UBM's for alignment is that a given test utterance only needs be aligned once (rather than once for each hypothesized speaker as required by the usual GMM approach). Thus the number of Gaussians in the UBM is not really a major computational issue for us.

As for signal processing, speech data was sampled at 8 kHz and 12 liftered mel frequency cepstral coefficients and an energy parameter were calculated at a frame rate of 10 ms. The acoustic feature vector consisted of these 13 parameters together with their first derivatives. For most of our experiments, we did not perform cepstral mean subtraction or normalize the energy feature. Although it is contrary to tradition in speaker recognition, there are several reasons for not doing mean normalization in the early phases of investigation: there is no difficulty in capturing linear channel effects with channel factors; there is evidence that cepstral mean subtraction reduces speaker variability as well as channel variability (albeit to a lesser extent) [28] so that it may be an impediment to discriminating between speakers; and an early experiment in combining eigenvoice and eigenchannel modeling indicated that cepstral mean subtraction could be deleterious [2]. Since we wanted to be able to compare our results on the 1999 and 2000 evaluation sets with those published by other authors, we did not use feature warping except in our final experiments (where it proved to be extremely effective).

Silences were excised from the NIST 1999 and NIST 2000 enrollment and test data so we used a silence detector to prepare the training data for our experiments on these evaluation sets. Traditionally, NIST defines a 'primary condition' for the annual evaluations. In 1999 and 2000 the primary conditions were defined in such a way that only the electret portions of the test sets were used in the evaluations. We did not impose the primary condition restrictions in our experiments since we were interested to see how the joint factor analysis model would perform on a mixture of electret and carbon data. In particular, we did not use handset detection (except to break down some of our results).

VI. Experiments

In this section we will first report results obtained on the NIST 1999 evaluation data (which is extracted from Switchboard II, Phase 3) by training PCA models (i.e. $\boldsymbol{d} = \mathbf{0}$) on Switchboard II, Phases 1 and 2. As we mentioned in Section IV, the training corpus and evaluation set for these experiments are mismatched.

In order to see how well a PCA model is capable of performing if the training corpus and evaluation set are well

[4]Even with this restriction training can still be time consuming. For example, we found that training a PCA model with 300 speaker factors and 100 channel factors on a Switchboard database takes almost one half real time per EM iteration on a 2.4 GHz Xeon CPU. We have since found a satisfactory approximation which enables the training procedure to be speeded up substantially [6].

matched, we carried out experiments on a subset of the NIST 2000 evaluation data (which was extracted from both Phases 1 and 2 of Switchboard II) where we used a subset of the NIST 2000 target speaker population for testing and a disjoint set of speakers from Switchboard II, Phases 1 and 2 for training.

For our final experiment with feature warping we used the 1999 evaluation set.

A. Data

For each target speaker in the 1999 evaluation set, the enrollment data consisted of about one minute of speech from each of two different conversations conducted over the same phone line. We used one of the two enrollment recordings for each of the target speakers (5 hours of data in the female case, 4 in the male) to train two gender-dependent UBM's each having 2,048 Gaussians. (This violates the NIST protocol.) We used the Switchboard II, Phases 1 and 2 corpora to construct two training sets (one male and one female) for training PCA models. We excluded all speakers who were included in the 1999 evaluation.[5] This left 625 female training speakers and 528 male. In our first set of experiments on the 1999 data (described in Section VI-C) we restricted ourselves to a subset of the available training data for computational reasons (128 hours after excising silences in the female case, 94 hours in the male case). However, we used all of the available training data for our second set of experiments (described in Section VI-E), namely 230 hours in the female case and 180 hours in the male case.

In order to design the training corpus and evaluation set for our experiment using the NIST 2000 evaluation data, we began by excluding every second target speaker as well as all speakers who were not in the Switchboard II, Phases 1 and 2 corpora. This gave us a set of 453 target speakers (203 male and 250 female). We selected a disjoint set consisting of 341 male and 385 female speakers from Switchboard II, Phases 1 and 2 for training. For each training speaker we used up to 20 conversation sides, giving 121 hours of training data in the female case and 96 hours in the male case. Finally, we constructed our evaluation set from the NIST 2000 evaluation data by excluding trials which involved speakers other than our 453 target speakers or test utterances which had been extracted from one of the conversation sides that we used in training. This left us with 27,438 verification trials (13,516 male and 13,922 female) involving 5,207 test utterances. Of these, 24% were recorded with carbon button handsets (compared with 23% of 6,052 test utterances in the test set as a whole).[6] Thus the results we will report on our evaluation set can be compared with results obtained by other authors on the NIST 2000 evaluation set as a whole (that is, without the primary condition restriction).

[5]Care is needed in making this determination because some speakers have aliases. This situation arises when two sets of enrollment data, one carbon and one electret, are supplied for a speaker.

[6]Handset detection was performed automatically so these figures may not be accurate.

B. A toy experiment

The most promising strategy for dealing with the mismatch between the training and target speaker populations in our experiments on the 1999 evaluation set seems to be to use a large number of speaker factors in the hope of generating a speaker space which is large enough to accommodate the target speakers as well as the training speakers. (This is borne out by experiments reported in [7].) However, if we use a large number of speaker factors then computational considerations force us to take $\boldsymbol{d} = \mathbf{0}$ in training the factor analysis model (as we mentioned in Section V) and to make an approximation in evaluating the sequential likelihood ratio statistic (7) (as we mentioned in Section III). In order to see if our model was capable in principle of performing well under these conditions, we performed an experiment on the female subset of the 1999 evaluation data where we trained a PCA model on the female portion of Switchboard II, Phase 3. (This is a 'toy' experiment since the evaluation data is a subset of the training corpus in this situation.)

In estimating the hyperparameters we limited ourselves to 10 conversation sides per speaker (this gave us 63 hours of female training data after excising silences) and we estimated a PCA model with 300 common speaker factors and 50 channel factors using the maximum likelihood training algorithm. In testing we used the sequential likelihood ratio statistic and t-norm score normalization with 50 t-norm speakers per test utterance. In order to evaluate the likelihood ratio statistic (7), we reduced the rank of $\boldsymbol{v}(s)$ from 300 to 50 for each target speaker s. Under these conditions we obtained a DCF (that is, the value of the NIST dectection cost function) of 0.016 and an equal error rate (EER) of 4.8% on the female portion of the evaluation set. This indicates that a PCA model together with the approximation used in evaluating the sequential likelihood ratio statistic can indeed perform very well, at least if the training corpus and evaluation set are perfectly matched. However it turns out that, as in the case of conventional GMM systems, the performance on different number trials is much poorer than on same number trials. This is apparent from Table I, where we have broken down the results in the same way as in [4].

TABLE I

Breakdown of results of the toy experiment on the female portion of the 1999 evaluation set obtained by training on Switchboard II, Phase 3. SNST = same number, same type (electret or carbon); DNST = different number, same type; DNDT = different number, different type.

	DCF	EER
SNST	0.006	2.2%
DNST	0.016	4.8%
DNDT	0.033	10.0%
overall	0.016	4.8%

C. Results on the 1999 evaluation set

In our first (non-toy) experiments on the 1999 evaluation set, we used 500 common speaker factors, 50 channel factors and we did not use feature warping. As in the toy experiment,

we reduced the rank of $\boldsymbol{v}(s)$ to 50 for each target speaker s in evaluating the likelihood ratio statistic (7) and we used 50 t-norm speakers for each test utterance. We first performed a series of experiments on the female portion of the evaluation set using the female training corpus extracted from Switchboard II, Phases 1 and 2 (as described in Section IV) to train PCA models. Our aim was to investigate the effectiveness of the two types of hyperparameter estimation algorithm (maximum likelihood and minimum divergence) in training the PCA models and in adapting speaker independent hyperparameters to the target speaker population (see Sections II-C and II-D). As we mentioned in Section V, it is only in adapting the speaker independent hyperparameters to the target speaker population that we relax the condition $\boldsymbol{d} = \boldsymbol{0}$. Thus most of our results were obtained with Principal Components Analysis (PCA) models rather than Factor Analysis (FA) models.

These results are summarized in Lines 1–5 of Table II. Line 1 gives the results obtained with maximum likelihood training of the PCA model without any adaptation to the target speaker population. These results are reasonably good, but not nearly as good as in our toy experiment which suggests that the mismatch between the training and target speaker populations may be a problem. The performance (as measured by the DCF) with minimum divergence training (Line 4) and with minimum divergence adaptation (Lines 3 and 5) was essentially the same as in Line 1.

TABLE II

Results on the female portion of the 1999 evaluation set obtained using maximum likelihood (ML) and minimum divergence (MD) algorithms for training speaker independent hyperparameters and adapting them to the target speaker population. PCA = principal components analysis, FA = joint factor analysis, — indicates that no adaptation to the target speaker population was performed. Feature warping was not used.

	Training	Adaptation	DCF	EER
1	ML	—	0.037	9.9%
2	ML	ML PCA	0.055	15.2%
3	ML	MD PCA	0.037	10.8%
4	ML + MD	—	0.036	10.4%
5	ML + MD	MD PCA	0.036	10.6%
6	ML + MD	MD PCA + $\boldsymbol{d}$	0.031	9.2%
7	ML + MD	MD FA	**0.030**	**9.2**%

All of these results were obtained by imposing the condition $\boldsymbol{d} = \boldsymbol{0}$. For the experiment reported in Line 6 we took the adapted PCA model from Line 5 and turned it into a joint factor analysis model simply by setting

$$\boldsymbol{d}^2 = \frac{1}{16}\boldsymbol{\Sigma} \tag{8}$$

as [4] would suggest. (As we mentioned, in the classical MAP case the 'relevance factors' in [4] are the diagonal entries of $\boldsymbol{d}^{-2}\boldsymbol{\Sigma}$.) This turned out to be a good choice; using the minimum divergence adaptation algorithm mentioned in Section II-D to adapt a joint factor analysis model initialized in this way to the target speaker population gave essentially the same results (Line 7).

Comparing the results in Lines 6 and 7 with those in Line 5 shows that joint factor analysis can outperform principal components analysis, that is, the diagonal term $\boldsymbol{dz}(s)$ in (3) can indeed play a useful role even if the number of common speaker factors is very large. The approach taken in Line 7 has the advantage that relevance factors do not have to be estimated empirically, but comparing Line 6 with Line 7, Line 4 with Line 5 and Line 1 with Line 3 shows that our attempts to mitigate the mismatch problem by adapting to the target speaker population using the approach indicated in Section II-D were unsuccessful.

Furthermore, the results in Line 2 show that adapting to the target speaker population can actually be harmful if maximum likelihood estimation is used in place of minimum divergence estimation for this purpose (as we mentioned in Section II-D). The reason for this is clear: minimum divergence adaptation preserves the orientation of the speaker space found in training, whereas maximum likelihood adaptation forgets this information and re-orients the speaker space to fit the target speaker population. Since the amount of enrollment data for the target speakers is relatively small (compared to the training corpus), maximum likelihood adaptation suffers from overfitting whereas minimum divergence adaptation does not.

We replicated the experiment in Line 7 on the male portion of the 1999 evaluation set, obtaining a DCF of 0.029 and an EER of 10.2%. Pooling male and female trials gave worse results than keeping them separate, namely a DCF of 0.033 and an EER of 10.9%, so it appears that using common thresholds for male and female trials is suboptimal for the joint factor analysis model. It is quite likely that this is due to the disparity in the sizes of the male and female training sets. (There are generally more females than males in the Switchboard corpora.)

For comparison, results on the 1999 evaluation set without the primary condition restriction are reported in [4] where an EER of 10% was obtained.

D. Results on the 2000 evaluation set

We used the NIST 2000 evaluation data to design some experiments to see how well the joint factor analysis model could perform in situations where the training corpus is well matched with the evaluation set. As we have already explained, we used a subset of the NIST 2000 evaluation data for testing and a disjoint subset of Switchboard II, Phases 1 and 2 for training in these experiments. They were carried out in the same way as those on the 1999 evaluation set except that we did not use the enrollment data in estimating the UBM's.

We used the entire training corpus to estimate gender-dependent UBM's with 2,048 Gaussians and gender-dependent PCA models with 300 speaker factors and 100 channel factors. We reduced the rank of $\boldsymbol{v}(s)$ to 100 for each target speaker s in evaluating the likelihood ratio statistic (7) and we used 50 t-norm speakers for each test utterance. On the female portion of our evaluation set, the PCA model gave a DCF of 0.030 and an EER of 8.1%. These results can be compared with the results reported in Line 5 of Table II.

The experiment reported in Line 7 of Table II showed that it was possible to compensate to some extent for the mismatch between the training and target speaker populations in the

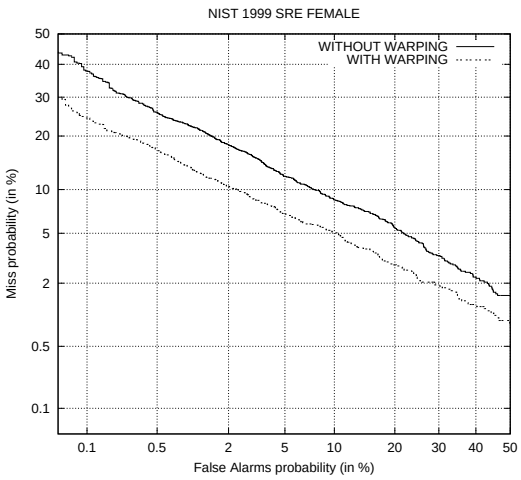

Fig. 4. *DET curves illustrating the effectiveness of feature warping in conjunction with the joint factor analysis model. Female portion of the NIST 1999 evaluation set, carbon data as well as electret. Without feature warping: DCF = 0.030, EER = 9.2% (upper curve). With feature warping: DCF = 0.021, EER = 6.2% (lower curve).*

experiments on the 1999 data by converting a PCA model to a joint factor analysis model. By using the same strategy here, we were able to obtain a DCF of 0.028 and an EER of 7.5% on the female portion of our test set. So this strategy is effective even if there is no evident mismatch although, as one would expect, the improvement in this case is smaller. Replicating this experiment on the male portion of our evaluation set, we obtained a DCF of 0.027 and an EER of 6.4%. Pooling the results (using a gender-independent decision threshold) gave a DCF of 0.028 and an EER of 7.2%.

There are few published results to compare these figures with, but it is generally recognized that, although it consists entirely of different number trials, the 2000 evaluation set is of about the same degree of difficulty as the 1999 evaluation set and that DCF's of roughly 0.037 and EER's of about 10% can be achieved by the methods in [4]. For our purposes, the important thing to note is that we obtained much better results on the 2000 test data than on the 1999 test data, and this difference can be attributed to using a training corpus which is well matched with the evaluation set.

E. The effect of feature warping

Since we failed to obtain good results on the 1999 evaluation data by compensating for the mismatch problem, we decided to try another technique. In all of the experiments reported so far we used unnormalized cepstral coefficients as acoustic features. It turns out however that using feature warping [15], [16] gives much better results. We replicated the experiment on the female portion of the 1999 evaluation set which gave us our best results (Line 7 of Table II) using feature warping, reducing the number of common speaker factors from 500 to 300 and using all of the training data available in the Switchboard II, Phases 1 and 2 corpora. Under these conditions we obtained greatly improved results, namely a DCF of 0.021 and an EER of 6.2%. The corresponding DET curves are shown in Fig. 4.

The results in the male case are similar, namely a DCF of 0.020 and an EER of 5.9%. To our knowledge these results are a good deal better than any that have been reported on the 1999 evaluation set (but recall that we have departed from the NIST evaluation protocol in some respects).

VII. Discussion

In this article we have shown how large scale factor analysis models of speaker and channel variability can be trained on large corpora using the maximum likelihood and minimum divergence algorithms described in [3] which optimize a likelihood II objective function, that is, a likelihood function whose arguments are hyperparameters. We have also shown how to construct a new type of likelihood II ratio statistic for speaker verification using such factor analysis models and how this approach overcomes the major limitation of eigenchannel MAP by taking account of channel effects at enrollment time as well as at verification time.

Our first results on the NIST 1999 evaluation set (obtained without feature warping) were worse than the results of the toy experiment in Section VI-B might lead one to expect. This suggested that the dialectical mismatch between the training and target speaker populations might be to blame. That there is some truth to this supposition is clear from the much better results we obtained in our experiments on the NIST 2000 evaluation data which were conducted with a well matched training corpus and evaluation set. We attempted to mitigate the mismatch problem by using the adaptation techniques described in Section II-D but our efforts in this direction were largely unsuccessful. Thus, the factor analysis model seems to require a training set which is well matched with the application domain in which it is expected to function and there does not seem to be any easy way of getting around this obstacle.

As we explained in Section IV, we have only been in a position to experiment with the factor analysis model using properly matched training corpora and evaluation sets since 2005. The results reported in the companion paper [7] were obtained on the NIST 2005 evaluation set (strictly in accordance with the NIST protocol). These results include back-to-back comparisons of a simplified version of the likelihood II approach developed here with eigenchannel MAP and they clearly show that the likelihood II approach is far superior. We found that the likelihood II approach had to be simplified because the computational demands of strictly adhering to the likelihood II principle made it difficult for us to turn around large numbers of experiments. After this work was completed, we showed how training factor analysis models could be speeded up by decoupling the speaker and channel components of the model [6], and how the CPU time needed for verification trials could be greatly reduced by using an approximation to the likelihood II statistic which can be evaluated without having to perform a Cholesky decomposition [26], [7].

The most interesting result to emerge from the experiments in this article is the very strong synergy between the joint factor analysis model and feature warping. This method of feature normalization consists in sliding an analysis window

(typically of length 3 sec) over the signal and mapping the distribution of each cepstral coefficient in the window onto a standard normal distribution. The technique is currently used in most, but by no means all, state of the art text-independent speaker recognition systems, although the reasons for its effectiveness do not seem to be fully understood. The most comprehensive reference on the subject is Pelecanos' unpublished thesis [16]. The arguments and empirical evidence presented there in favor of normalizing the first and second order moments of the cepstral distributions in this manner seem to be incontrovertible but it is less clear why normalizing *all* of the moments in this way should be an effective strategy for dealing with channel effects in speaker recognition. The theoretical arguments advanced by Pelecanos only pertain to the case of additive white noise, but the results using this technique on real data, namely the NIST 1999 evaluation set, were very impressive.

The surprising thing about our results on the same evaluation set is that feature warping in conjunction with factor analysis actually gave us a much larger gain in performance (30% reductions in error rates as measured both by DCF and EER) than Pelecanos and others have obtained by using feature warping in conjunction with the GMM/UBM approach. We suspect that the reason for this is that the joint factor analysis model depends on Gaussian assumptions at the level of supervectors, and the short term Gaussianization performed in feature warping reinforces these assumptions. However, in order to test this hypothesis, it would be necessary to relax the Gaussian assumptions that we used to model supervector distributions (by using independent components analysis rather than principal components analysis or factor analysis, for example). This seems to be a difficult problem which would only be worth tackling if additional evidence could be found to suggest that it might be of practical importance.

REFERENCES

[1] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 6, pp. 780–788, June 2002.

[2] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[3] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13," 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

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

[5] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[7] ——, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[8] ——, "Improvements in factor analysis based speaker verification," in *Proc. ICASSP 2006*, Toulouse, France, May 2006.

[9] ——, "The geometry of the channel space in GMM-based speaker recognition," in *Proc. IEEE Odyssey 2006*, San Juan, Puerto Rico, June 2006.

[10] S.-C. Yin, P. Kenny, and R. Rose, "Experiments in speaker adaptation for factor analysis based speaker verification," in *Proc. IEEE Odyssey 2006*, San Juan, Puerto Rico, June 2006.

[11] H. Botterweck, "Anisotropic MAP defined by eigenvoices for large vocabulary continuous speech recognition," in *Proc. ICASSP*, Salt Lake City, Utah, May 2001.

[12] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[13] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, Hong Kong, China, Apr. 2003.

[14] J. Campbell *et al.*, "The MMSR bilingual and crosschannel corpora for speaker recognition research and evaluation," in *Proc. Odyssey 2004*, Toledo, Spain, June 2004, pp. 29–32.

[15] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, June 2001, pp. 213–218.

[16] J. Pelecanos, "Robust automatic speaker recognition," Ph.D. dissertation, Queensland University of Technology, Brisbane, Australia, 2003.

[17] D. Rubin and D. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, vol. 47, pp. 69–76, 1982.

[18] D. Kim and N. Kim, "Online adaptation of continuous density hidden Markov models based on speaker space model evolution," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 1393–1396.

[19] S. Lucey and T. Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 2021–2024.

[20] R. Vogt and S. Sridharan, "Bayes factor scoring of GMMs for speaker verification," in *Proc. Odyssey 2004*, Toledo, Spain, June 2004, pp. 173–178.

[21] ——, "Frame-weighted Bayes factor scoring for speaker verification," in *Proc. 10th Australian International Conference on Speech Science and Technology*, Sydney, Australia, 2004, pp. 404–409.

[22] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker adaptation using an eigenphone basis," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 6, Nov. 2004. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[23] B. Mak, T.-C. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," in *Proc. ICASSP*, Toulouse, France, May 2006.

[24] H. Jiang and L. Deng, "A Bayesian approach to the verification problem: Applications to speaker verification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 874–884, 2001.

[25] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–52, 2000.

[26] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. ICASSP*, Montreal, Canada, May 2004. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[27] (2001) The NIST year 2001 speaker recognition evaluation plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf

[28] S. Kajarekar, N. Malayath, and H. Hermansky, "Analysis of speaker and channel variability in speech," in *Workshop on Automatic Speech Recognition and Understanding*, Keystone, CO, Dec. 1999, pp. 59–62.

Patrick Kenny received the BA degree in Mathematics from Trinity College, Dublin and the MSc and PhD degrees, also in Mathematics, from McGill University. He was a professor of Electrical Engineering at INRS-Télécommunications in Montreal from 1990 to 1995 when he started up a company (Spoken Word Technologies) to spin off INRS's speech recognition technology. He joined CRIM in 1998 where he now holds the position of principal research scientist. His current research interests are concentrated on Bayesian speaker and channel adaptation for speech and speaker recognition.

Gilles Boulianne received the B.Sc. degree in Unified Engineering from Université du Québec à Chicoutimi and the M.Sc. degree in Telecommunications from INRS-Telecommunications, Montreal. He worked on speech analysis and articulatory speech modeling at the UQAM Linguistics Department until 1990, then on large vocabulary speech recognition at INRS and Spoken Word Technologies until 1998. He has been since with the Computer Research Institute of Montreal Speech Recognition Team. His research interests include finite state transducer approaches and practical applications of large vocabulary speech recognition such as live closed-captioning and content indexation.

Pierre Ouellet obtained the BSc degree in Computer Science from McGill University in 1994. He joined the École de Technologie Supérieure in Montreal in 1997 to work on speaker identification in the context of dialogs in noisy environments. Since 1998, he has been working in the CRIM Speech Recognition team, where he contributes to ASR software development. His interests are software implementation issues and the application of adaptation techniques.

Pierre Dumouchel , B.Eng. (McGill University), M.Sc., Ph.D. (INRS-Télécommunications). Pierre is currently Scientific Vice-President at CRIM and full professor at the École de technologie supérieure (ETS) of the Université du Qubec. Pierre was the vice-president of Research and Development at CRIM from 1999 to 2004. Before that he assumed the role of Principal Researcher of the CRIMs Automatic Speech Recognition team and was a scientific columnist at Radio-Canada, the French Canadian National Radio. He has more than 20 years of expertise in Speech Recognition Research, eight years in managing a research team and three years in managing the Research and Development unit of CRIM. His research has resulted in many technology transfers to such companies as Nortel, Locus Dialog, Canadian National Defence, Le Groupe TVA, as well as many small and medium size businesses, as such as Ryshco Media. His research interests are in search by transduction and automatic adaptation to new environments. He has favoured applications of speech recognition for the hard-of-hearing and audio-visual film indexation.