A Study of Interspeaker Variability in Speaker Verification

Patrick Kenny, *Member, IEEE*, Pierre Ouellet, Najim Dehak, Vishwa Gupta, *Senior Member, IEEE*, and Pierre Dumouchel, *Member, IEEE*

***Abstract*—We propose a new approach to the problem of estimating the hyperparameters which define the interspeaker variability model in joint factor analysis. We tested the proposed estimation technique on the NIST 2006 speaker recognition evaluation data and obtained 10%–15% reductions in error rates on the core condition and the extended data condition (as measured both by equal error rates and the NIST detection cost function). We show that when a large joint factor analysis model is trained in this way and tested on the core condition, the extended data condition and the cross-channel condition, it is capable of performing at least as well as fusions of multiple systems of other types. (The comparisons are based on the best results on these tasks that have been reported in the literature.) In the case of the cross-channel condition, a factor analysis model with 300 speaker factors and 200 channel factors can achieve equal error rates of less than 3.0%. This is a substantial improvement over the best results that have previously been reported on this task.**

***Index Terms*—Channel factors, Gaussian mixture model (GMM), speaker factors, speaker verification.**

I. Introduction

FACTOR analysis is a model of speaker and session variability in Gaussian mixture models (GMMs). This paper is concerned with the speaker variability component of our version of factor analysis. In our approach to speaker recognition, the role of this component is to provide a prior distribution for target speaker models (we use the term prior distribution in the sense in which it is used in Bayesian statistics [1]). As such, it plays a key role in estimating target speaker models at enrollment time.

In order to formulate precisely the problem that we address, we begin by recapitulating the basic assumptions in factor analysis. Let C be the number of components in a universal background model (UBM) and F the dimension of the acoustic feature vectors. We use the term supervector to refer to the CF-dimensional vector obtained by concatenating the F-dimensional mean vectors in the GMM corresponding to a given utterance.

Manuscript received January 10, 2008; revised April 4, 2008. 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 associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

The authors are with the Centre de Recherche Informatique de Montréal (CRIM), Montreal, QC H3A 1B9, Canada (e-mail: patrick.kenny@crim.ca; pierre.ouellet@crim.ca; najim.dehak@crim.ca; vishwa.gupta@crim.ca; pierre.dumouchel@crim.ca).

Digital Object Identifier 10.1109/TASL.2008.925147

Our assumptions are as follows. First, we assume that a speaker- and channel-dependent supervector $\boldsymbol{M}$ can be decomposed into a sum of two supervectors, a speaker supervector $\boldsymbol{s}$, and a channel supervector $\boldsymbol{c}$

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c} \tag{1}$$

where $\boldsymbol{s}$ and $\boldsymbol{c}$ are statistically independent and normally distributed.

Second, we assume that the distribution of $\boldsymbol{s}$ has a hidden variable description of the form

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{2}$$

where $\boldsymbol{m}$ is a $CF \times 1$ supervector, $\boldsymbol{v}$ is a rectangular matrix of low rank and $\boldsymbol{y}$ is a normally distributed random vector, $\boldsymbol{d}$ is a $CF \times CF$ diagonal matrix, and $\boldsymbol{z}$ is a normally distributed CF-dimensional random vector. We will refer to the columns of $\boldsymbol{v}$ as eigenvoices, and we will refer to the components of $\boldsymbol{y}$ as speaker factors.[1]

Third, we assume that the distribution of $\boldsymbol{c}$ has a hidden variable description of the form

$$\boldsymbol{c} = \boldsymbol{u}x \tag{3}$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank, and $\boldsymbol{x}$ is a normally distributed random vector. We refer to the components of $\boldsymbol{x}$ as channel factors, and we use the term eigenchannels to refer to the columns of $\boldsymbol{u}$.

Finally, we associate a diagonal covariance matrix Σ_c with each mixture component c whose role is to model the variability in the acoustic observation vectors which is not captured by either the speaker model (2) or the channel model (3). We denote by $\boldsymbol{\Sigma}$ the $CF \times CF$ supercovariance matrix whose diagonal is the concatenation of these covariance matrices. Although most authors (e.g., [4] and [5]) use the term *factor analysis* to refer to the channel model (3) alone, we have always used this term in a broader sense which includes the speaker model (2) as well. (Where it is necessary to make this distinction explicitly we speak of *joint factor analysis*.) Our concern in this paper is with the way the hyperparameters $\boldsymbol{v}$ and $\boldsymbol{d}$ in (2) are estimated. These hyperparameters provide a prior distribution for *maximum a priori* (MAP) estimation of speaker-dependent GMMs at enrollment time, and they are critically important to the success of our approach to speaker recognition. (The MAP

[1]As in our previous work, we are following the usage of [2]. A different usage prevails in the general statistical literature: the columns of $\boldsymbol{v}$ would be referred to as speaker factors and the entries of $\boldsymbol{y}$ as factor loadings. The terminology is used in this way in [3] where factor analysis methods are applied to the face recognition problem.

calculation is explained in Section III of [6].) Since the assumption in (2) is equivalent to saying that $\boldsymbol{s}$ is normally distributed with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{d}^2 + \boldsymbol{v}v^*$, (2) is a model of interspeaker variability. Our goal in this paper is to show how improved interspeaker variability modeling can lead to substantial gains in speaker recognition performance. If $\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{u} = \boldsymbol{0}$, then the assumption in (2) is the same as in classical MAP [7]; on the other hand, if $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{u} = \boldsymbol{0}$, the assumption is the same as in eigenvoice MAP [8].

Classical MAP adaptation (including relevance MAP [9]) is by far the most popular type of speaker modeling in text-independent speaker recognition, but our experience has been that MAP adaptation using eigenvoices is generally much more effective, at least in situations where limited amounts of enrollment data are available. Classical MAP adaptation can only adapt those Gaussians which are seen in the enrollment data but, if large amounts of enrollment data are available, it is arguably the best way of estimating speaker supervectors since it is asymptotically equivalent to maximum-likelihood estimation. On the other hand, eigenvoice MAP is helpful if only small amounts of enrollment data are available, since only a small number of free parameters need to be estimated at enrollment time. The fact that the supervector covariance matrix is full rather than diagonal in this case ensures that MAP adaptation takes account of the correlations between the different Gaussians in a speaker supervector so that all of the Gaussians are updated at enrollment time even if only a small fraction of them are observed. An extreme example of the effectiveness of eigenvoices can be found in [10], which is concerned with the use of factor analysis to model syllable-level prosodic features. The number of feature vectors per conversation side is only about 400; it is unrealistic to expect classical MAP adaptation to be very effective in this situation.

It should be possible to capitalize on the advantages of both classical MAP and eigenvoice MAP by including both terms $\boldsymbol{vy}$ and $\boldsymbol{dz}$ in (2) (this was first suggested in [11]). However, extensive experimentation in [12] showed that the term $\boldsymbol{dz}$ was only helpful on an extended data task where 15–20 min of enrollment data are available for each target speaker. (The term $\boldsymbol{vy}$ is helpful in all circumstances, even in the extended data task.) Since including the term $\boldsymbol{dz}$ is the source of most of the mathematical complication in [6], and the extended data task is of secondary interest to most researchers, this led us to wonder if we would not be better off suppressing the term $\boldsymbol{dz}$ altogether.

The reason why the term $\boldsymbol{dz}$ was not helpful in [12] is that in a typical factor analysis training scenario with, say, 1000 training speakers and 300 speaker factors, almost all of the speaker variability in the training set can be well accounted for by $\boldsymbol{v}$ alone ($\boldsymbol{v}$ has 300 times as many free parameters as $\boldsymbol{d}$). Thus, if the maximum likelihood criterion is used to estimate $\boldsymbol{d}$ and $\boldsymbol{v}$, what tends to happen is that $\boldsymbol{d}$ ends up playing no useful role unless very large amounts of enrollment data are available (as in the extended data task). However, there is reason to doubt that the maximum-likelihood criterion is appropriate for this type of estimation problem. Even if the linear/Gaussian assumptions in (2) are granted, there is no reason to believe that (2) is a correct model of interspeaker variability—it is just a compromise that is forced on us by the fact that a supervector covariance matrix of sufficiently high rank to be realistic would probably be impossible to estimate or to calculate with. (Impossible to calculate with because the rank of the covariance matrix would be too high; impossible to estimate because many more training speakers would be needed than are currently available.)

This led us to explore another way of estimating $\boldsymbol{v}$ and $\boldsymbol{d}$ which we will explain in Section II, and which we refer to as *decoupled estimation*. In Section III, we show how decoupled estimation leads to 10%–15% reductions in error rates (as measured both by equal error rates and the NIST detection cost function) on both the core condition and the extended data condition of the NIST 2006 speaker recognition evaluation data. In order to be able to turn around these experiments in a reasonable time, we used factor analysis models of relatively modest dimensions. Our final results, using a much larger factor analysis model, are presented in Section IV. These results show that a stand-alone joint factor analysis model is capable of performing at least as well as fusions of large numbers of systems of other types (based on comparisons with the best results that have been reported in the literature). The results on the NIST 2006 cross channel condition are particularly impressive: equal error rates of less than 3% can be achieved without any special-purpose signal processing. Results of other tests are presented in [13]; these include cross-channel tests in which microphone speech is used for enrollment as well as for verification and tests involving very short utterances at verification time.

II. Estimating the Hyperparameters

The supervector defined by a UBM can serve as an estimate of $\boldsymbol{m}$, and the UBM covariance matrices are good first approximations to the residual covariance matrices Σ_c $(c = 1, \ldots, C)$. The problem of estimating $\boldsymbol{v}$ in the case where $\boldsymbol{d} = \boldsymbol{0}$ was addressed in [8] and a very similar approach can be adopted for estimating $\boldsymbol{d}$ in the case where $\boldsymbol{v} = \boldsymbol{0}$. We first summarize the estimation procedures for these two special cases and then explain how they can be combined to tackle the general case.

A. Baum–Welch Statistics

Given a speaker s and acoustic feature vectors $Y_1, Y_2, \ldots$, for each mixture component c we define the Baum–Welch statistics in the usual way

$$N_c(s) = \sum_t \gamma_t(c)$$

$$F_c(s) = \sum_t \gamma_t(c) Y_t$$

$$S_c(s) = \operatorname{diag}\left(\sum_t \gamma_t(c) Y_t Y_t^*\right)$$

where, for each time t, $\gamma_t(c)$ is the posterior probability of the event that the feature vector Y_t is accounted for by the mixture component c. We calculate these posteriors using the UBM.

We denote the centralized first- and second order Baum–Welch statistics by $\tilde{F}_c(s)$ and $\tilde{S}_c(s)$

$$\tilde{F}_c(s) = \sum_t \gamma_t(c)(Y_t - m_c)$$

$$\tilde{S}_c(s) = \operatorname{diag}\left(\sum_t \gamma_t(c)(Y_t - m_c)(Y_t - m_c)^*\right)$$

where m_c is the subvector of $\boldsymbol{m}$ corresponding to the mixture component c. In other words

$$\tilde{F}_c(s) = F_c(s) - N_c(s)m_c$$
$$\tilde{S}_c(s) = S_c(s) - \mathrm{diag}(F_c(s)m_c^* + m_cF_c(s)^* - N_c(s)m_cm_c^*).$$

Let $\boldsymbol{N}(s)$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c(s)I$ $(c = 1, \ldots, C)$. Let $\tilde{\boldsymbol{F}}(s)$ be the $CF \times 1$ supervector obtained by concatenating $\tilde{F}_c(s)$ $(c = 1, \ldots, C)$. Let $\tilde{\boldsymbol{S}}(s)$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $\tilde{S}_c(s)$ $(c = 1, \ldots, C)$.

B. Training an Eigenvoice Model

In this section, we consider the problem of estimating $\boldsymbol{m}, \boldsymbol{v}$ and $\boldsymbol{\Sigma}$ under the assumption that $\boldsymbol{d} = \boldsymbol{0}$. We assume that initial estimates of the hyperparameters are given. (Random initialization of $\boldsymbol{v}$ works fine in practice.)

The approach that we adopt is similar to Gales's cluster adaptive training [14] in that it does not rely on techniques such as MAP or MLLR to produce GMMs for the training speakers (all of the computation is done with Baum–Welch statistics rather than GMMs). It differs from the approach in [14] in that the hyperparameter estimation problem is formulated in terms of maximum likelihood II [15]. As such, our approach is very similar to the probabilistic principal components analysis (PPCA) of [16] (which is formally a special case of our procedure). Note that, as in PPCA, the terms eigenvector and eigenvalue do not appear in our formulation but it is known that, unless the optimization gets stuck locally, PPCA does succeed in finding principal eigenvectors. Thus, it is appropriate for us to speak of eigenvoices even though our estimation procedure is not formulated as an eigenvalue problem.

1) Posterior Distribution of the Hidden Variables: For each speaker s, set $\boldsymbol{l}(s) = \boldsymbol{I} + \boldsymbol{v}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(s)\boldsymbol{v}$. Then the posterior distribution of $\boldsymbol{y}(s)$ conditioned on the acoustic observations of the speaker is Gaussian with mean $\boldsymbol{l}^{-1}(s)\boldsymbol{v}^*\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}(s)$ and covariance matrix $\boldsymbol{l}^{-1}(s)$. (See [8, Prop. 1].) We will use the notation $E[\cdot]$ to indicate posterior expectations; thus, $E[\boldsymbol{y}(s)]$ denotes the posterior mean of $\boldsymbol{y}(s)$ and $E[\boldsymbol{y}(s)\boldsymbol{y}^*(s)]$ the posterior correlation matrix.

2) Maximum-Likelihood Re-Estimation: This entails accumulating the following statistics over the training set, where the posterior expectations are calculated using initial estimates of $\boldsymbol{m}, \boldsymbol{v}, \boldsymbol{\Sigma}$ and s ranges over the training speakers

$$N_c = \sum_s N_c(s) \; (c = 1, \ldots, C)$$
$$\boldsymbol{\mathfrak{A}}_c = \sum_s N_c(s)E[\boldsymbol{y}(s)\boldsymbol{y}^*(s)] \; (c = 1, \ldots, C)$$
$$\boldsymbol{\mathfrak{C}} = \sum_s (\tilde{\boldsymbol{F}}(s)E[\boldsymbol{y}^*(s)])$$
$$\boldsymbol{N} = \sum_s \boldsymbol{N}(s).$$

For each mixture component $c = 1, \ldots, C$ and for each $f = 1, \ldots, F$, set $i = (c-1)F + f$; let v_i denote the ith row of $\boldsymbol{v}$ and $\boldsymbol{\mathfrak{C}}_i$ the ith row of $\boldsymbol{\mathfrak{C}}$. Then $\boldsymbol{v}$ is updated by solving the equations

$$v_i\boldsymbol{\mathfrak{A}}_c = \boldsymbol{\mathfrak{C}}_i \; (i = 1, \ldots, CF).$$

The update formula for $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}(\sum_s \tilde{\boldsymbol{S}}(s) - \mathrm{diag}(\boldsymbol{\mathfrak{C}}\boldsymbol{v}^*)).$$

(See [8, Prop. 3].)

3) Minimum-Divergence Re-Estimation: Given initial estimates $\boldsymbol{m}_0$ and $\boldsymbol{v}_0$, the update formulas for $\boldsymbol{m}$ and $\boldsymbol{v}$ are

$$\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{v}_0\boldsymbol{\mu}_{\boldsymbol{y}}$$
$$\boldsymbol{v} = \boldsymbol{v}_0\boldsymbol{T}_{\boldsymbol{yy}}^*.$$

Here

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \frac{1}{S}\sum_s E[\boldsymbol{y}(s)].$$

$\boldsymbol{T}_{\boldsymbol{yy}}$ is an upper triangular matrix such that

$$\boldsymbol{T}_{\boldsymbol{yy}}^*\boldsymbol{T}_{\boldsymbol{yy}} = \frac{1}{S}\sum_s E[\boldsymbol{y}(s)\boldsymbol{y}^*(s)] - \boldsymbol{\mu}_{\boldsymbol{y}}\boldsymbol{\mu}_{\boldsymbol{y}}^*$$

(i.e., Cholesky decomposition), S is the number of training speakers, and the sums extend over all speakers in the training set. (See [6, Theorem 7].) This update formula leaves the range of the covariance matrix $\boldsymbol{v}\boldsymbol{v}^*$ unchanged. The only freedom it has is to rotate the eigenvoices and scale the corresponding eigenvalues. This type of hyperparameter estimation was introduced in [17]; its role is to get good estimates of the eigenvalues corresponding to the eigenvoices ([18], Section II-C). Thus, it is useful for diagnostic purposes; for example, in comparing the eigenvalues of $\boldsymbol{u}\boldsymbol{u}^*$ with those of $\boldsymbol{v}\boldsymbol{v}^*$ as in Table VI. Maximum-likelihood estimation on its own produces eigenvalues which are difficult to interpret [6].

C. Training a Diagonal Model

An analogous development can be used to estimate $\boldsymbol{m}, \boldsymbol{d}$ and $\boldsymbol{\Sigma}$ if $\boldsymbol{v}$ is constrained to be $\boldsymbol{0}$.

1) Posterior Distribution of the Hidden Variables: For each speaker s, set $\boldsymbol{l}(s) = \boldsymbol{I} + \boldsymbol{d}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(s)$. Then the posterior distribution of $\boldsymbol{z}(s)$ conditioned on the acoustic observations of the speaker is Gaussian with mean $\boldsymbol{l}^{-1}(s)\boldsymbol{d}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}(s)$ and covariance matrix $\boldsymbol{l}^{-1}(s)$. (The derivation here is essentially the same as in Section II-B1.)

Again, we will use the notation $E[\cdot]$ to indicate posterior expectations; thus, $E[\boldsymbol{z}(s)]$ denotes the posterior mean of $\boldsymbol{z}(s)$ and $E[\boldsymbol{z}(s)\boldsymbol{z}^*(s)]$ the posterior correlation matrix.

It is straightforward to verify that, in the special case where $\boldsymbol{d}$ is assumed to satisfy

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

this posterior calculation leads to the standard relevance MAP estimation formulas for speaker supervectors [9] (r is the relevance factor). The following two sections summarize data-driven procedures for estimating $\boldsymbol{m}, \boldsymbol{d}$ and $\boldsymbol{\Sigma}$ which do not

depend on the relevance MAP assumption. It can be shown that when these update formulas are applied iteratively, the values of a likelihood function analogous to that given in Proposition 2 of [8] increase on successive iterations.

2) Maximum-Likelihood Re-Estimation: This entails accumulating the following statistics over the training set where the posterior expectations are calculated using initial estimates of $\boldsymbol{m}, \boldsymbol{d}, \boldsymbol{\Sigma}$, and s ranges over the training speakers

$$N_c = \sum_s N_c(s) \; (c = 1, \ldots, C)$$
$$\mathfrak{a} = \sum_s \operatorname{diag}\left(\boldsymbol{N}(s) E[\boldsymbol{z}(s)\boldsymbol{z}^*(s)]\right)$$
$$\mathfrak{b} = \sum_s \operatorname{diag} \tilde{\boldsymbol{F}}(s) E[\boldsymbol{z}^*(s)]$$
$$\boldsymbol{N} = \sum_s \boldsymbol{N}(s).$$

For $i = 1, \ldots, CF$ let d_i be the ith entry of $\boldsymbol{d}$ and similarly for $\mathfrak{a}_i$ and $\mathfrak{b}_i$. Then $\boldsymbol{d}$ is updated by solving the equation

$$d_i \mathfrak{a}_i = \mathfrak{b}_i$$

for each i. The update formula for $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \boldsymbol{N}^{-1} \left(\sum_s \tilde{\boldsymbol{S}}(s) - \operatorname{diag}\left(\mathfrak{b}\boldsymbol{d}\right) \right).$$

3) Minimum-Divergence Re-Estimation: Given initial estimates $\boldsymbol{m}_0$ and $\boldsymbol{d}_0$, the update formulas for $\boldsymbol{m}$ and $\boldsymbol{d}$ are

$$\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{d}_0 \boldsymbol{\mu}_{\boldsymbol{z}}$$
$$\boldsymbol{d} = \boldsymbol{d}_0 \boldsymbol{T}_{\boldsymbol{zz}}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \frac{1}{S} \sum_s E[\boldsymbol{z}(s)].$$

$\boldsymbol{T}_{\boldsymbol{zz}}$ is a diagonal matrix such that

$$\boldsymbol{T}_{\boldsymbol{zz}}^2 = \operatorname{diag}\left(\frac{1}{S} \sum_s E[\boldsymbol{z}(s)\boldsymbol{z}^*(s)] - \boldsymbol{\mu}_{\boldsymbol{z}} \boldsymbol{\mu}_{\boldsymbol{z}}^* \right).$$

S is the number of training speakers, and the sums extend over all speakers in the training set.

We will need a variant of this update procedure which applies to the case where $\boldsymbol{m}$ is forced to be $\boldsymbol{0}$. In this case, $\boldsymbol{d}$ is estimated from $\boldsymbol{d}_0$ by taking $\boldsymbol{T}_{\boldsymbol{zz}}$ to be such that

$$\boldsymbol{T}_{\boldsymbol{zz}}^2 = \operatorname{diag}\left(\frac{1}{S} \sum_s E[\boldsymbol{z}(s)\boldsymbol{z}^*(s)] \right).$$

D. Joint Estimation of $\boldsymbol{v}$ and $\boldsymbol{d}$

There is no difficulty in principle in extending the maximum-likelihood and minimum-divergence training procedures to handle a general factor analysis model in which both $\boldsymbol{v}$ and $\boldsymbol{d}$ are nonzero [6, Theorems 4, 7]. We used this type of joint estimation in all of our previous work in factor analysis and to produce benchmarks for the experiments that we will report in this article.

However, joint estimation of $\boldsymbol{v}$ and $\boldsymbol{d}$ is computationally demanding because, in a general factor analysis model, all of the hidden variables become correlated with each other in the posterior distributions. Our experience has been that, given the Baum–Welch statistics, training a diagonal model runs very quickly, and training a pure eigenvoice model can be made to run quickly (at the cost of some memory overhead) by suitably organizing the computation of the matrices $\boldsymbol{l}(s)$ in Section II-B1. Unfortunately, no such computational shortcuts seem to be possible in the general case. Furthermore, even if the eigenvoice component $\boldsymbol{v}$ is carefully initialized, many iterations of joint estimation seem to be needed to estimate $\boldsymbol{d}$ properly and, because the contribution of $\boldsymbol{d}$ to the likelihood of the training data is minor compared with the contribution of $\boldsymbol{v}$, it is difficult to judge when the training algorithm has effectively converged.

E. Decoupled Estimation of $\boldsymbol{v}$ and $\boldsymbol{d}$

A much more serious problem with joint estimation is that it tends to produce estimates of $\boldsymbol{d}$ which are too small, so that almost all of the speaker variability in a factor analysis training set is accounted for by the term $\boldsymbol{vy}$ in (2) and very little of the variability is accounted for by the term $\boldsymbol{dz}$. Thus, in practice, the term $\boldsymbol{dz}$ is of little use except in situations where large amounts of enrollment data are available (as we observed in [12]).

It is probable that the reason why joint estimation behaves in this way is that it is a maximum-likelihood estimation procedure, and $\boldsymbol{v}$ has many more free parameters than $\boldsymbol{d}$. However, as we mentioned in Section I, there is reason to doubt the appropriateness of maximum-likelihood estimation in this situation, and there is a good argument which suggests that the term $\boldsymbol{dz}$ ought to be helpful in distinguishing between speakers. The term $\boldsymbol{vy}$ can only capture interspeaker variability which is confined to a low-dimensional affine subspace of supervector space (namely the subspace containing $\boldsymbol{m}$ which is spanned by the eigenvoices). It is reasonable to believe that the orientation of this subspace reflects attributes which are common to all speakers. No such constraint is imposed on the term $\boldsymbol{dz}$, so it ought to be capable of capturing attributes which are unique to individual speakers. Similar considerations led the authors in [19] and [20] to construct speaker recognition systems which operate in the orthogonal complement of the principal components of a large training set. (This orthogonal complement is referred to as the "speaker unique subspace" in [20].)

This raises the question of how to produce a reasonable estimate of $\boldsymbol{d}$ which is not "too small." Since the term $\boldsymbol{dz}$ models residual interspeaker variability which is not captured by a large set of eigenvoices, this can be achieved by withholding a subset of the training speaker population; the speakers that are withheld serve to estimate $\boldsymbol{d}$ but they play no role in estimating $\boldsymbol{v}$.

Thus, we split the factor analysis training set in two and use the larger of the two sets to estimate $\boldsymbol{m}$ and $\boldsymbol{v}$ and the smaller to estimate $\boldsymbol{d}$ and $\boldsymbol{\Sigma}$. We first fit a pure eigenvoice model to the larger training set using the procedures described in Sections II-B2 and II-B3. Then, for each speaker s in the residual training set, we calculate the MAP estimate of $\boldsymbol{y}(s)$,

namely $E[\boldsymbol{y}(s)]$, as in Section II-B1. This gives us a preliminary estimate of the speaker's supervector $\boldsymbol{s}$, namely

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}E[\boldsymbol{y}(s)]. \tag{4}$$

We centralize the speaker's Baum–Welch statistics by subtracting the speaker's supervector (that is, we apply the formulas in Section II-A with $\boldsymbol{m}$ replaced by $\boldsymbol{s}$). Finally, we use these centralized statistics together with the procedures described in Sections II-C2 and II-C3 to estimate a pure diagonal model with $\boldsymbol{m} = \boldsymbol{0}$. This gives us estimates of $\boldsymbol{d}$ and $\boldsymbol{\Sigma}$.

Since this training algorithm uses only the diagonal and eigenvoice estimation procedures, it converges rapidly.

III. Experiments

A. Enrollment and Test Data

We used the core condition and the extended data condition (in which eight-conversation sides are available for enrolling each target speaker) of the NIST 2006 speaker recognition evaluation (SRE) for testing [21]. Although we will report results on male speakers as well as female, we used mostly the female trials in the 2006 SRE for our experiments.

B. Feature Extraction

We extracted 19 cepstral coefficients together with a log energy feature using a 25-ms Hamming window and a 10-ms frame advance. These were subjected to feature warping using a 3-s sliding window [22]. Δ coefficients were then calculated using a five-frame window, giving a total of 40 features.

C. Factor Analysis Training Data

We trained two gender-dependent UBMs having 1024 Gaussians and gender-dependent factor analysis models having 0, 100, and 300 speaker factors. Except where otherwise indicated, the number of channel factors was fixed at 50.

For training UBMs, we used Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Parts 1 and 2; the NIST 2003 Language Recognition Evaluation data set; and the NIST 2004 SRE enrollment and test data.

For training factor analysis models, we used the LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 SRE data. We used only those speakers for which five or more recordings were available. For decoupled estimation of $\boldsymbol{v}$ and $\boldsymbol{d}$, we estimated $\boldsymbol{v}$ on the Switchboard data and $\boldsymbol{d}$ on the 2004 SRE data.

In order to ensure strict disjointness between the factor analysis training data and the NIST 2006 SRE data which we used for testing, we made no use of the 2005 SRE data. (For the extended data condition, some of the 2005 data was recycled in 2006. In [23], we reported how failing to keep the training and test sets disjoint could produce extremely misleading results.) Note also that, since the Switchboard corpora consist of English only data and English is predominant in the NIST 2004 SRE data, the factor analysis training set is biased towards English speakers.

D. Implementation Details

The first step in building factor analysis models is to train gender-dependent UBMs in the usual way. Baum–Welch statistics extracted with these UBMs are sufficient statistics for all subsequent processing: hyperparameter estimation, target speaker enrollment, and likelihood calculations at verification time.

To estimate $\boldsymbol{v}$ and $\boldsymbol{d}$, we pooled all of the recordings of each speaker in the factor analysis training set and ignored channel effects as in [24]. (The rationale here is that channel effects can be averaged out if sufficiently many recordings are available for each speaker.) In implementing decoupled estimation of $\boldsymbol{v}$ and $\boldsymbol{d}$, we ran the algorithms in Section II-B to convergence (seven iterations of maximum-likelihood estimation and one of minimum-divergence estimation) before calculating the speaker supervectors for each training speaker according to (4).

We decoupled the estimation of $\boldsymbol{u}$ from that of $\boldsymbol{v}$ and $\boldsymbol{d}$ as in [25] and [24] (rather than using the maximum-likelihood procedures in [6] and [18]).

Recall that (2) can be interpreted as saying that speaker supervectors are normally distributed with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{d}^2 + \boldsymbol{v}\boldsymbol{v}^*$. For the purposes of enrolling target speakers, we interpret this normal distribution as a prior distribution in the sense in which this term is used in Bayesian statistics. Given an enrollment utterance and the hyperparameters $\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{d}$, we enroll a target speaker by calculating the posterior distribution of the hidden variables $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$, using the maximum *a posteriori* estimate of $\boldsymbol{m}+\boldsymbol{v}\boldsymbol{y}+\boldsymbol{d}\boldsymbol{z}$ as a point estimate of the speaker's supervector. (We do not use the point estimate of $\boldsymbol{x}$ since the channel effects in the enrollment data are irrelevant.)

As is generally the case with a Gaussian prior, the posterior is also Gaussian and can be calculated in closed form [1]. The case where $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{u} = \boldsymbol{0}$ is treated in Section II-B1; the case where $\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{v} = \boldsymbol{0}$ is treated in Section II-C1; the case where $\boldsymbol{d} \neq \boldsymbol{0}, \boldsymbol{v} = \boldsymbol{0}, \boldsymbol{u} = \boldsymbol{0}$, and there is a single enrollment utterance is treated in the Appendix to [26]; the case where $\boldsymbol{d} \neq \boldsymbol{0}, \boldsymbol{v} \neq \boldsymbol{0}, \boldsymbol{u} = \boldsymbol{0}$, and there is a single enrollment utterance is formally equivalent to this—one only has to replace $\boldsymbol{u}$ by the matrix

$$(\boldsymbol{u} \quad \boldsymbol{v}).$$

Finally, the general case in which there are multiple enrollment utterances is treated in [6, Sec. III], but we have found that pooling the Baum–Welch statistics from the various utterances together (as if we had a single utterance for enrollment) works just as well as the complicated calculation described there.

Note that our enrollment procedure results in a point of estimate of each target speaker's supervector, rather than a posterior distribution as in the Bayesian approach that we originally attempted [18].

At verification time, likelihoods were evaluated according to (19) in [24]. (We did not use the correction (20) in [24]. This is a minor technical issue which is discussed at length in [12].) Thus, we account for channel effects in test utterances by integrating over the channel factors $\boldsymbol{x}$ in (3) rather than by using a point estimate of the channel factors for each test utterance as other authors do.

If a test utterance is sufficiently long, the posterior distribution of the channel factors will be sharply peaked and using a

TABLE I
RESULTS OBTAINED ON THE CORE CONDITION OF THE NIST 2006 SRE (FEMALE SPEAKERS, ENGLISH LANGUAGE TRIALS)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	4.4%	0.027	3.9%	0.022
300 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	4.1%	0.024	**3.6%**	**0.021**
300 speaker factors, $\boldsymbol{d} = \mathbf{0}$	3.9%	0.024	–	–
0 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	5.2%	0.027	–	–

TABLE II
RESULTS OBTAINED ON THE CORE CONDITION OF THE NIST 2006 SRE (FEMALE SPEAKERS, ALL TRIALS)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	5.9%	0.032	4.9%	0.027
300 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	5.6%	0.030	**4.6%**	**0.025**
300 speaker factors, $\boldsymbol{d} = \mathbf{0}$	5.2%	0.028	–	–
0 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	7.2%	0.034	–	–

point estimate of the channel factors (either a MAP estimate or a maximum-likelihood estimate) will give essentially the same result as integrating over the channel factors. However, in the case of short test utterances (say 10 s of speech), integrating over channel factors seems to be the right thing to do. (Since the integral in question is Gaussian, there is no difficulty in evaluating it in closed form.) It was reported in [27]–[29] that channel factors are unhelpful for tasks involving short test utterances, but this does not agree with our experience. In [13] we present some good results on 10-s test conditions; we believe that our success can be traced to *not* attempting to obtain point estimates of channel factors under these conditions.

Finally, the denominator of the log likelihood ratio statistic used for verification was calculated in exactly the same way as the numerator with the UBM supervector used in place of the hypothesized speaker's supervector.

E. Imposters

The verification decision scores obtained with the factor analysis models were normalized using zt-norm. As in [12], we used 283 t-norm speakers in the female case and 227 in the male case. We used 1000 z-norm utterances for each gender. The imposters were chosen at random from the factor analysis training data. The reasons for using such a large number of z-norm utterances are explained in [12].

F. Results

The results of our experiments on the female portion of the common subset of the core condition of the NIST 2006 SRE are summarized in Table I. (EER refers to the equal error rate, and DCF to the minimum value of the NIST detection cost function. The common subset consists of English language trials only. All results in this paper were obtained using version 5 of the 2006 SRE answer key.) There are some blank entries in the table because decoupled estimation applies only in the case where both $\boldsymbol{v}$ and $\boldsymbol{d}$ are nonzero. The best result is obtained with 300 speaker factors and decoupled estimation. It is apparent that, contrary to our conclusion in [12], the term $\boldsymbol{dz}$ in (2) can play a useful role in restricted data tasks after all. There is an anomaly in the joint estimation column: In the 300-speaker factor case, we obtained a better EER by setting $\boldsymbol{d} = \mathbf{0}$ than by joint estimation of $\boldsymbol{v}$ and $\boldsymbol{d}$. We attribute this to the convergence issue mentioned in Section II-D.

Table II gives the corresponding results on all trials of the female portion of the core condition. Again, the best results are obtained with 300-speaker and decoupled estimation. Comparing the second row of Table II with that of Table I we see that, if $\boldsymbol{d}$ is estimated with decoupled estimation, then the term $\boldsymbol{dz}$ in (2) is particularly effective in modeling non-English speakers. This is to be expected since we used a large amount of English-only data (namely, the Switchboard corpora) to estimate $\boldsymbol{v}$.

We replicated these experiments on the female trials of the extended data condition. The results are summarized in Tables III and IV. Patterns similar to those in Tables I and II are evident.

We report results on male speakers in Table V. Note that these results are much better than the results we obtained for female speakers.

TABLE III
RESULTS OBTAINED ON THE EXTENDED DATA CONDITION OF THE NIST 2006 SRE (FEMALE SPEAKERS, ENGLISH LANGUAGE TRIALS)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	2.2%	0.012	2.1%	0.011
300 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	2.1%	0.014	**1.9%**	**0.011**
300 speaker factors, $\boldsymbol{d} = \mathbf{0}$	2.1%	0.014	–	–
0 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	3.1%	0.017	–	–

TABLE IV
RESULTS OBTAINED ON THE EXTENDED DATA CONDITION OF THE NIST 2006 SRE (FEMALE SPEAKERS, ALL TRIALS)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	2.5%	0.012	2.3%	0.014
300 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	2.7%	0.014	**2.3%**	**0.012**
300 speaker factors, $\boldsymbol{d} = \mathbf{0}$	2.7%	0.015	–	–
0 speaker factors, $\boldsymbol{d} \neq \mathbf{0}$	3.6%	0.016	–	–

TABLE V
RESULTS OBTAINED ON THE CORE CONDITION AND THE EXTENDED DATA CONDITION OF THE NIST 2006 SRE FOR MALE SPEAKERS (50 CHANNEL FACTORS, 300 SPEAKER FACTORS, $\boldsymbol{d} \neq \mathbf{0}$, DECOUPLED ESTIMATION)

	EER	DCF
Core condition, English	2.1%	0.013
Core condition, all trials	4.2%	0.020
Extended data, English	1.4%	0.006
Extended data, all trials	1.7%	0.008

G. Note on Baum–Welch Statistics

The results we have obtained using speaker factors are clearly much better than those obtained using $\boldsymbol{d}$ alone, but the reader may have noticed that the figures presented in the fourth rows of Tables I and II are not quite as good as the best results that have been reported with comparable stand-alone GMM/UBM systems as in [30], [5], and [31]. These systems are comparable because they use relevance MAP for speaker enrollment and channel factors to compensate for intersession variability. As we mentioned in Section II-C1, relevance MAP is essentially a special type of diagonal factor analysis model.

The reason for the discrepancy in performance is that we use the UBM to extract Baum–Welch statistics in our system rather than speaker-dependent GMMs. It turns out that, in the case of a diagonal factor analysis model, using speaker-dependent GMMs does indeed produce better results. For example, on the English language trials in the core condition, a diagonal model with 100 channel factors produces an EER of 2.8% for male speakers, which is similar to the results presented in [30], [5], and [31] (but not as good as the result in the first line of Table V).

However, for a factor analysis model with 300 speaker factors, using speaker-dependent GMMs (estimated with speaker factors) to extract Baum–Welch statistics turns out to be harmful. For example, on the English language trials in the core condition, a factor analysis model with 300 speaker factors and 100 channel factors produces an EER of 4.2% for male speakers if the Baum–Welch statistics are extracted with speaker-dependent GMMs; on the other hand, an EER of 1.4% is obtained if the UBM is used for this purpose. This is the reason why we have always used the UBM to extract Baum–Welch statistics in our work on factor analysis. (The extraordinarily low error rate of 1.4% is attributable to using 100 channel factors rather than 50 as in Table V.)

IV. Results Obtained With a Large Factor Analysis Model

In this section, we report results obtained on the NIST 2006 SRE test set by increasing the dimensions of the male and female factor analysis models. We increased the number of Gaussians from 1024 to 2048, we increased the dimension of the acoustic feature vectors from 40 to 60 by appending $\Delta\Delta$ coefficients, and we increased the number of channel factors from 50 to 100. We kept the number of speaker factors at 300 because previous experience has shown that using larger numbers of speaker factors is not helpful [12].

In presenting the results that we obtained with the large factor analysis models, we will break them out by gender because (as we saw in the previous section) there are large differences in performance between males and females. Some insight into this phenomenon can be gained by inspecting the way the male and female factor analysis models fit the training data. Table VI gives the traces of the matrices $\boldsymbol{v}\boldsymbol{v}^*$, $\boldsymbol{d}^2$ and $\boldsymbol{u}\boldsymbol{u}^*$ for the two gender-dependent factor analysis models. It is clear that, as measured both by the trace of $\boldsymbol{v}\boldsymbol{v}^*$ and by the trace of $\boldsymbol{d}^2$, there is substantially greater variability among male speakers than among female speakers. Thus, distinguishing between male speakers seems to be intrinsically easier than distinguishing between female speakers, at least if the feature set consists of cepstral coefficients.

TABLE VI
Speaker and Channel Variability in Male and Female Factor Analysis Models

	Male speakers	Female speakers
$\mathrm{tr}\,(\boldsymbol{v}\boldsymbol{v}^*)$	1075	975
$\mathrm{tr}\,(\boldsymbol{d}^2)$	334	306
$\mathrm{tr}\,(\boldsymbol{u}\boldsymbol{u}^*)$	501	433
$\mathrm{tr}\,(\boldsymbol{\Sigma})$	25,840	23,535

TABLE VII
Results Obtained With a Large Factor Analysis Model on the Core Condition of the NIST 2006 SRE

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	1.5%	0.011	2.7%	0.017
All trials	3.0%	0.017	3.3%	0.020

(In our earlier work we used 12 cepstral coefficients. We increased the number of cepstral coefficients to 19 in the present work in the hope that this would narrow the gender gap.) Perhaps even more surprisingly, we observe that the trace of $\boldsymbol{u}\boldsymbol{u}^*$ is substantially larger for the male model than for the female model, which seems to indicate that the channel model (3) gives a better fit in the case of male speech. However, although the figures in the fourth row of Table VI are not really comparable with the others, they suggest that most of the variability in the data is not captured by either the speaker model (2) or the channel model (3), and this residual variability is larger in the case of males than in the case of females.

A. Core Condition

The results we obtained on the core condition of the NIST 2006 SRE with the large factor analysis model just described are summarized in Table VII. Even though they were obtained with a stand-alone system, these results compare very favorably with the best results on the 2006 core condition that have been reported in the literature, namely those obtained by STBU [32], SRI [33], and MIT/IBM [4]. The STBU system achieved an EER of 2.3% (English language trials only, results pooled over male and female speakers) by fusing ten subsystems (cepstral and MLLR); SRI achieved an EER of 2.6% by fusing eight subsystems (cepstral, MLLR, and higher level); and MIT/IBM achieved an EER of 2.7% by fusing nine subsystems (cepstral, MLLR, and higher level).

In comparing our results with those of STBU, it should be noted that the individual subsystems of the STBU system were trained on pre-2005 data, but the fusion parameters were estimated using the data made available for the 2005 NIST SRE. (Robust fusion was a key ingredient in the success of the STBU system in the 2006 SRE.) On the other hand, our reason for excluding the 2005 data from the factor analysis training set was simply to enable us to experiment properly with the extended data condition in the 2006 evaluation set (as we explained in Section III-C). Had we included the 2005 data in factor analysis training, the proportion of Mixer data in the factor analysis training set would have increased from 20% to 50%, and this would presumably have resulted in even better performance on the 2006 evaluation set.

TABLE VIII
RESULTS OBTAINED WITH A LARGE FACTOR ANALYSIS MODEL ON THE EXTENDED DATA CONDITION OF THE NIST 2006 SRE

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	0.8%	0.004	1.6%	0.009
All trials	1.1%	0.007	1.8%	0.009

TABLE IX
RESULTS OBTAINED WITH A LARGE FACTOR ANALYSIS MODEL ON THE CROSS-CHANNEL CONDITION OF THE NIST 2006 SRE

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	2.6%	0.011	2.9%	0.014
All trials	2.5%	0.011	3.3%	0.015

B. Extended Data Condition

Table VIII summarizes the results obtained with the large factor analysis model on the extended data condition of the 2006 NIST SRE. The best results on this task in the literature are those reported by MIT/IBM [4] where EERs of 1.5% (English language trials, male and female results pooled) and 2.6% (all trials) were obtained by fusing nine subsystems (cepstral, MLLR, and higher level). It is interesting to note that, although the extended data task was intended to encourage research into higher level systems, and higher level systems (including an MLLR system) play an important role in reducing the error rates in [4], we were able to obtain better results using cepstral features alone.

C. Cross-Channel Condition

In the cross-channel condition, the enrollment data for each target speaker consists of a conversation side extracted from a recording of a telephone conversation but the test data consists of recordings made using one of eight different microphones. (The identity of the microphone is not given. The cross-channel task is described in detail in [34].)

We used the development data provided by NIST to estimate 100 eigenchannels to model the effects of the various microphones, and we appended these eigenchannels to the 100 eigenchannels that we had previously estimated on telephone speech. Thus, the factor analysis model that we used at recognition time had 200 channel factors rather than 100. Since the enrollment data in this task consists of telephone speech, we did not have to make any change to the factor analysis model used at enrollment time. The only other modification that we made was to choose the z-norm utterances from the cross-channel development data rather than from the factor analysis training data described in Section III-C.

The results we obtained on the cross-channel test data are summarized in Table IX. These results are a good deal better than the best results that have been reported on this task, namely an EER of 4.0% obtained by MIT [34]. The MIT results were obtained by fusing two cepstral systems (a support vector machine with nuisance attribute projection and a GMM/UBM system with channel factors) and speech enhancement played an important role in reducing error rates. Our system makes use of no special-purpose signal processing; it relies solely on a large number of channel factors to compensate for transducer effects. The results of other auxiliary microphone tests (where microphone speech is used at enrollment time as well as at verification time) can be found in [13].

V. CONCLUSION

We have shown how careful modeling of interspeaker variability enables a stand-alone joint factor analysis system to perform as well as fusions of large numbers of systems of other types (which typically include models of intersession variability but not of interspeaker variability). Of course, this achievement comes at a cost—the implementation is more complicated and painstaking experimentation is needed—but our approach beats the state-of-the-art on the NIST 2006 extended data task (without using higher level features) and on the cross-channel task (without using speech enhancement).

The principal departure from our earlier work is that we abandoned (at least partially) the maximum-likelihood principle for estimating the hyperparameters which define a joint factor analysis model. This decision was driven by the results in [12], which led us to conclude that the maximum-likelihood principle could not produce useful estimates of the hyperparameter $\boldsymbol{d}$ in (2), apparently because (2) is not a realistic model of interspeaker variability. There is an obvious parallel here with speech recognition: Hidden Markov models are not realistic models of acoustic–phonetic phenomena, and maximum-likelihood estimation does not perform as well as other estimation criteria (such as maximum mutual information or minimum phone error). This raises the question of whether similar discriminative training criteria can be used to estimate factor analysis hyperparameters. First steps in this direction have been taken in [35] and [36], but the results suggest that getting this type of approach to work in speaker recognition may require a large effort, just as it has in speech recognition.

REFERENCES

[1] A. O'Hagan and J. Forster, *Kendall's Advanced Theory of Statistics*. London, U.K.: Arnold, 2004, vol. 2B.
[2] J. A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson, S. P. Khudanpur, M. Ostendorf, and R. Rosenfeld, Eds. New York: Springer-Verlag, 2004, pp. 191–246.
[3] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV'07*, Rio de Janeiro, Brazil, Oct. 2007, CD-ROM.
[4] W. M. Campbell, D. E. Sturim, W. Shen, D. A. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition," in *Proc. ICASSP'07*, Honolulu, HI, Apr. 2007, pp. IV-217–IV-220.
[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1969–1978, Sep. 2007.
[6] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep. CRIM-06/08-13, 2005 [Online]. Available: http://www.crim.ca/perso/patrick.kenny
[7] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[8] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–359, May 2005.
[9] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[10] N. Dehak, P. Kenny, and P. Dumouchel, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2095–2103, Sep. 2007.
[11] S. Lucey and T. Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2021–2024.
[12] P. Kenny, N. Dehak, R. Dehak, V. Gupta, and P. Dumouchel, "The role of speaker factors in the NIST extended data task," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008, CD-ROM.
[13] P. Kenny, N. Dehak, P. Ouellet, V. Gupta, and P. Dumouchel, "Development of the primary CRIM system for the NIST 2008 speaker recognition evaluation," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, CD-ROM.
[14] M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 417–428, Jul. 2000.
[15] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Comput.*, vol. 11, no. 5, pp. 1035–1068, 1999.
[16] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Comput.*, vol. 11, pp. 435–474, 1999.
[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker adaptation using an eigenphone basis," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 579–589, Nov. 2004.
[18] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.
[19] S. Kajarekar, "Four weightings and a fusion," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, 2005, pp. 17–22.
[20] H. Aronowitz, "Speaker recognition using kernel-PCA and inter-session variability modeling," in *Proc. Interspeech'07*, Antwerp, Belgium, Aug. 2007, pp. 298–301.
[21] "The NIST year 2006 Speaker Recognition Evaluation Plan." 2006 [Online]. Available: http://www.nist.gov/speech/tests/spk/2006/index.htm
[22] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, Jun. 2001, pp. 213–218.
[23] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. ICASSP*, Montreal, QC, Canada, May 2004, pp. I-37–I-40 [Online]. Available: http://www.crim.ca/perso/patrick.kenny
[24] 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.
[25] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP'05*, Philadelphia, PA, Mar. 2005, pp. 637–640 [Online]. Available: http://www.crim.ca/perso/patrick.kenny
[26] S.-C. Yin, R. Rose, and P. Kenny, "A joint factor analysis approach to progressive model adaptation in text independent speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1999–2010, Sep. 2007.
[27] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre, and J. Mason, "Influence of task duration in text-independent speaker verification," in *Proc. Interspeech'07*, Antwerp, Belgium, Aug. 2007, pp. 794–797.
[28] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short-duration SVM- and GMM-based speaker verification," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008, CD-ROM.
[29] R. Vogt, C. Lustri, and S. Sridharan, "Factor analysis modeling for speaker verification with short utterances," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008, CD-ROM.
[30] L. Burget, P. Matejka, O. Glembek, P. Schwarz, and J. Cernocky, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1979–1986, Sep. 2007.
[31] B. G. B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1960–1968, 2007.
[32] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. 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.
[33] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez, and G. Tur, "Speech recognition as feature extraction for speaker recognition," in *Proc. SAFE 2007: Workshop Signal Process. Applicat. Public Security and Forensics*, Washington, DC, 2007, pp. 39–43.
[34] D. E. Sturim, W. M. Campbell, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Robust speaker recognition with cross-channel data: MIT-LL results on the 2006 NIST SRE auxiliary microphone task," in *Proc. ICASSP'07*, Honolulu, HI, Apr. 2007, pp. IV-49–IV-52.
[35] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP'05*, Philadelphia, PA, Mar. 2005.
[36] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant NAP for SVM speaker recognition," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008, pp. 629–632, CD-ROM.

Patrick Kenny (M'05) received the B.A. degree in mathematics from Trinity College Dublin, U.K., and the M.Sc. and Ph.D. degrees, also in mathematics, from McGill University, Montreal, QC, Canada.

He was a Professor of electrical engineering at INRS-Télécommunications, 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.

Najim Dehak received the M.S. degree in pattern recognition and artificial intelligence from the Université de Pierre et Marie Curie Paris VI, Paris, France, in 2004 and the B.Eng. degree in artificial intelligence from the Université des Sciences et de la Technologie d'Oran, Oran, Algeria, in 2003. He is currently pursuing the Ph.D. degree at the École de Technologie Supérieure, Montreal, QC, Canada,

He is also with the Centre de Recherche informatique de Montréal. His research interests are speaker modeling and recognition.

Vishwa Gupta (SM'91) received the B.Tech. degree from the Indian Institute of Technology (IIT), Kharagpur, India, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from Clemson University, Clemson, SC.

At Nortel, he worked on many speech recognition applications including ADAS Plus for directory assistance automation. He published many papers and obtained several patents while at Nortel. At SpeechWorks, he was involved in research and development of applications as a member of the product team. At IBM, he worked on their speech recognition engine to incorporate state-of-the-art algorithms and features. At the Centre de Recherche informatique de Montréal (CRIM), he has been active in speech recognition, speaker diarization, keyword spotting, and automated advertisement detection.

Pierre Ouellet received the B.Sc. degree in computer science from McGill University, Montreal, QC, Canada, in 1994.

He joined the École de Technologie Supérieure, Montreal, in 1997 to work on speaker identification in the context of dialogs in noisy environments. Since 1998, he has been working in the Centre de Recherche informatique de Montréal (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 (M'97) received the B.Eng. degree from McGill University, Montreal, QC, Canada, and the M.Sc. and Ph.D. degrees from INRS-Télécommunications, Montreal.

He is Scientific Vice-President at CRIM and Full Professor at the École de Technologie Supérieure, Université du Québec. He was the Vice-President Research and Development of CRIM from 1999 to 2004. Before he assumed the role of Principal Researcher of the CRIM's Automatic Speech Recognition team, he 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-sized enterprises, such as Ryshco Media. His research interests are in search by transduction and automatic adaptation to new environments. He favored applications of speech recognition for the hard-of-hearing and audio–visual film indexation.