

Speaker Adaptation Using an Eigenphone Basis

Patrick Kenny, Gilles Boulianne, Pierre Ouellet and Pierre Dumouchel

Abstract— We describe a new method of estimating speaker-dependent HMM's for speakers in a closed population. Our method differs from previous approaches in that it is based on an explicit model of the correlations between all of the speakers in the population, the idea being that if there is not enough data to estimate a Gaussian mean vector for a given speaker then data from other speakers can be used provided that we know how the speakers are correlated with each other. We explain how to estimate inter-speaker correlations using a Kullback-Leibler divergence minimization technique which can be applied to the problem of estimating the parameters of all of the hyperdistributions that are currently used in Bayesian speaker adaptation.

EDICS Category: 1-RECO

I. INTRODUCTION

SUPPOSE that we are given a speaker-independent HMM containing a total of C mixture components and a speaker population comprising S speakers. For each mixture component $c = 1, \dots, C$, let μ_c denote the speaker-independent mean vector associated with the mixture component. For each speaker $s = 1, \dots, S$, we would like to provide a reasonable estimate of the corresponding speaker-dependent mean vector using statistics collected from all of the speakers in the population. (We will also address the problem of variance adaptation but not that of adapting transition probabilities and mixture weights.)

For a speaker s and mixture component c , let us denote the speaker-dependent mean vector by $\mu_c(s)$ and let $S_{X,c}(s)$ denote the corresponding first order statistics extracted from the training data (we will define these statistics properly later). The simplest assumption we can make is that $\mu_c(s)$ is a function of $S_{X,c}(s)$ alone (so that none of the other first order statistics are deemed to be useful in estimating $\mu_c(s)$). This assumption underlies both maximum likelihood and classical MAP speaker adaptation [1] and it gives good results provided large amounts of training data are available for each speaker.

Extended MAP (or EMAP) adaptation, in the form in which it was introduced in [2], is less restrictive in that, for each speaker s , it predicts $\mu_c(s)$ from the entire collection of first order statistics for the speaker, namely

$$\{S_{X,c'}(s) : c' = 1, \dots, C\}.$$

An explicit formula for the EMAP estimator is given in Corollary 1 to Proposition 1 in [3]. Similar linear predictive estimators have been investigated by numerous authors beginning with Furui [4]. A few examples are [5], [6], [7], [8], [9], [10], [11]; others can be found in the bibliography in [12].

If we define a *speaker supervector* to be a supervector obtained by concatenating the mean vectors $\mu_c(s)$ ($c = 1, \dots, C$) for some speaker s then the basic assumption underlying this type of extended MAP adaptation is that speaker supervectors are statistically independent with a common Gaussian distribution. In [3] we showed how the principal eigenvectors of the covariance matrix of this distribution can be estimated in situations where speaker-dependent training is not feasible. We will refer to this covariance matrix as the intra-speaker correlation matrix and denote it by B .

Our purpose in this article is to develop a new type of extended MAP estimation in which, for each mixture component c , the MAP estimator for $\mu_c(s)$ is a function of the entire collection of first order statistics for the mixture component, namely

$$\{S_{X,c}(s') : s' = 1, \dots, S\}.$$

Similar types of linear predictive estimator have been studied in [13], [14]. Our approach is based on the idea that an explicit model of the correlations between all pairs of speakers in the population can serve as a basis for sharing data among speakers. It can also be regarded as a type of soft speaker clustering (in the sense that hard speaker clustering uses binary valued predictor coefficients).

If we define a *phone supervector* to be a supervector obtained by concatenating the mean vectors $\mu_c(s)$ ($s = 1, \dots, S$) for some mixture component c , then our basic assumption in this article is that phone supervectors are statistically independent with a common Gaussian distribution. This is the dual of the assumption underlying intra-speaker correlation modeling. We will refer to the covariance matrix of this distribution as the inter-speaker correlation matrix and denote it by A . If we define the eigenphones to be the eigenvectors of A corresponding to non-zero eigenvalues then phone supervectors can be decomposed into eigenphones in the same way that speaker supervectors can be decomposed into eigenvoices. (But just as we did not use eigenvoices explicitly in [3] we will not use eigenphones explicitly in this article.) We will use the terms ‘eigenphone modeling’ and ‘inter-speaker correlation modeling’ interchangeably to refer to MAP speaker adaptation using an inter-speaker correlation matrix. The principal theoretical contribution of this article is to explain how to estimate the inter-speaker correlation matrix A using a new divergence minimization technique which can be applied (at least in principle) to the problem of estimating the parameters of all of the hyperdistributions (normal, Dirichlet or Wishart) that are currently used in Bayesian speaker adaptation [12].

Both eigenvoices and eigenphones can claim to be models of inter-speaker variability. In the case of eigenvoice modeling

Manuscript received
The authors are with the Centre de recherche informatique de Montréal (CRIM).

this claim is based on the fact that a relatively small number of eigenvoices generally suffices to capture most of the variation between speakers. But the argument here implicitly assumes that speakers are statistically independent whereas the eigenphone approach is based on an explicit model of inter-speaker dependencies. It will be interesting to see how these two approaches to speaker modeling perform on various tasks. We will report the results of a back to back comparison of eigenvoice and eigenphone modeling on the *AUPELF* French language speech recognition task in this article; see [15] for some comparisons in the context of speaker identification. Our experience generally has been that eigenphone modeling has a slight but consistent edge over eigenvoice modeling.

Despite the fact that speakers are demonstrably not statistically independent, relatively little work has been done on the problem of how to exploit dependencies between speakers. On the other hand an enormous amount of effort has been devoted to modeling intra-speaker correlations. A major reason for this emphasis seems to be that intra-speaker correlation modeling has a strong intuitive appeal because it fits neatly into a Bayesian framework. If the given speaker population consists of designated training and test speakers then intra-speaker correlations can be estimated using *only* the training speakers' data. (This statement obviously does not apply to inter-speaker correlations.) Then, given some adaptation data from a test speaker, these estimates can be used to derive the posterior distribution of the speaker's supervector at recognition time in a straightforward way. Thus intra-speaker correlation modeling is naturally suited to (supervised or unsupervised) on-line speaker adaptation. On the other hand inter-speaker correlation modeling does not allow this type of temporal interpretation of priors and posteriors since the correlations between all pairs of speakers in the population have to be estimated before constructing a speaker-adapted model for a given test speaker. Consequently using an eigenphone model for on-line speaker adaptation is not so easy.

Its intuitive appeal notwithstanding, intra-speaker correlation modeling runs into a formidable obstacle in practice, namely that the intra-speaker correlation matrix \mathbf{B} is very difficult to estimate because its dimensions are huge ($CF \times CF$ where F is the dimension of the acoustic feature vectors). Many different ways of attacking this problem have been suggested such as [16], [17], [18] (see [19] for some comparisons). The eigenvoice approach in [3] is an exact solution to the problem of maximum likelihood estimation of \mathbf{B} but it results in an estimate whose rank is bounded by the number of speakers in the training set. Thus, in practice, there is no way to ensure that eigenvoice MAP speaker adaptation is asymptotically equivalent to speaker-dependent training as the amount of adaptation data increases. As we discussed in [3], constraining the intra-speaker correlation matrix in various ways helps to alleviate this problem but only to a limited extent. So a substantial amount of further research (perhaps involving probabilistic factor analysis rather than probabilistic principal components analysis) seems to be needed in order to estimate intra-speaker correlation matrices properly.

One of the principal motivations for pursuing the eigenphone approach is that it does *not* encounter this type of

difficulty in practice since the dimensions of the inter-speaker correlation matrix \mathbf{A} are much less than those of the intra-speaker correlation matrix \mathbf{B} ($SF \times SF$ versus $CF \times CF$ where S is the number of speakers in the population) and the number of eigenphones that can be estimated is bounded only by the number of mixture components in a speaker HMM which can be made as large as we wish. (We will use the terms 'speaker HMM' and 'speaker-adapted HMM' interchangeably.)

Another difference between eigenvoice and eigenphone MAP becomes apparent when we consider how they are related to classical MAP. A little thought shows that if the intra-speaker correlation matrix \mathbf{B} is constrained to be diagonal (so that speakers and Gaussians are assumed to be statistically independent) then EMAP adaptation reduces to classical MAP adaptation (with normal rather than normal-Wishart priors). However it is difficult to develop a unified approach to classical MAP and eigenvoice modeling along the lines of [3] because eigenvoice modeling cannot be applied in practice unless \mathbf{B} is of low rank. (Again, this difficulty could be overcome by probabilistic factor analysis.) On the other hand, there is no fundamental difficulty in working with inter-speaker correlation matrices of full rank since these matrices are of relatively low dimension. In particular, we can constrain an inter-speaker correlation matrix to be diagonal or block diagonal without any problem. Under this assumption, speakers and Gaussians are statistically independent so the eigenphone approach contains another flavor of classical MAP as a special case. This serves as a natural benchmark for evaluating the effectiveness of inter-speaker correlation modeling and the experimental results we report in this article indicate that eigenphone MAP is indeed more effective than classical MAP on the *AUPELF* task.

The question of how to exploit both types of correlation (inter-speaker and intra-speaker) simultaneously raises interesting and difficult problems. We will conclude the article with a brief discussion of this topic.

II. PRELIMINARIES

The principal problem that we have to address is how to estimate the inter-speaker correlation matrix \mathbf{A} in situations where speaker-dependent training is not feasible or, in other words, where the phone supervectors are unobservable. (If speaker-dependent training were feasible we would have no need for MAP adaptation and hence no need for \mathbf{A} .) Our assumptions are:

- 1) For each mixture component c there is an $F \times 1$ mean vector μ_c and an $F \times F$ covariance matrix Σ_c .
- 2) For each speaker $s = 1, \dots, S$ and each mixture component c , there is an unobservable $F \times 1$ vector $O_c(s)$ (O for offset) such that if X is an observation vector (frame) from mixture component c for speaker s then

$$X = \mu_c + O_c(s) + E$$

where E is normally distributed with mean 0 and covariance matrix Σ_c .

- 3) If, for each mixture component c , \mathbf{O}_c is the $SF \times 1$ vector obtained by concatenating $O_c(1), \dots, O_c(S)$,

then the *a priori* distribution of \mathbf{O}_c is normal with mean $\mathbf{0}$ and covariance matrix \mathbf{A} .

We refer to Σ_c as a residual covariance matrix (or simply a covariance matrix). Let Σ be the $CF \times CF$ block diagonal matrix whose diagonal blocks are $\Sigma_1, \dots, \Sigma_C$.

We will assume that we are given a training set in which each frame has been labeled by a mixture component (as in a Viterbi alignment) and we will show how to formulate the problem of estimating \mathbf{A} and Σ as one of maximizing the likelihood of the training data where the likelihood function is evaluated according to assumptions (1), (2) and (3). (See Proposition 2 below for a computational formula.) If R is the rank of \mathbf{A} we can write

$$\mathbf{A} = \mathbf{U}\mathbf{U}^*$$

where \mathbf{U} is a $SF \times R$ matrix of rank R . This implies that for each mixture component c , there is a unique $R \times 1$ vector \mathbf{x}_c such that

$$\mathbf{O}_c = \mathbf{U}\mathbf{x}_c$$

and the prior distribution of \mathbf{x}_c is normal with mean $\mathbf{0}$ and covariance matrix \mathbf{I} . The algorithm for estimating \mathbf{U} and Σ that we develop in this article is an EM algorithm in which the role of the hidden variables is played by the random vectors \mathbf{x}_c (rather than by the random vectors \mathbf{O}_c as in [20]).

One way of developing such an algorithm is to adapt the argument in [3] so as to interchange the roles of speakers and mixture components. (If the residual covariances in [3] are tied across all mixture components, then the roles of speakers and mixture components are essentially symmetric and so can be interchanged without much difficulty. This gives an algorithm for estimating \mathbf{A} and Σ in the special case where the residual covariances are tied across all mixture components. With a bit more effort, this approach can be made to work without any compromise in generality.) This will be an efficient approach if the inter-speaker correlation matrix can be constrained to be of low rank.

The assumption of low rank is reasonable in situations where the number of Gaussians in a speaker model is relatively small (as in the GMM's used for speaker recognition, for instance) but it is open to question in the general case. So we will develop another algorithm to estimate \mathbf{A} in this article which requires less computation in the full rank case. The approach is based on a simple divergence minimization procedure which we will outline here and develop in detail in the appendix.

For simplicity we will ignore the question of how to estimate the residual covariance matrices and concentrate on the problem of how to estimate the inter-speaker correlation matrix. Let $\nu_{\mathbf{A}}$ denote the Gaussian distribution with mean zero and covariance matrix \mathbf{A} . Given an initial estimate \mathbf{A}_0 of the inter-speaker correlation matrix, for each mixture component c , let $\hat{\pi}_c$ be the posterior distribution of \mathbf{O}_c calculated using the prior $\nu_{\mathbf{A}_0}$ and the observation vectors aligned with c . Set

$$\mathbf{A}_1 = \operatorname{argmin}_{\mathbf{A}} \frac{1}{C} \sum_{c=1}^C D(\hat{\pi}_c \| \nu_{\mathbf{A}})$$

where $D(\cdot \| \cdot)$ denotes the Kullback-Leibler divergence. Iterating this procedure we obtain a sequence of inter-speaker correlation matrices $\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2, \dots$. We will show that this sequence converges (at least to the extent that the likelihood of the training data calculated according to our assumptions (1), (2) and (3) increases on successive iterations).

Just as for normal distributions, closed-form expressions for the Kullback-Leibler divergences of Dirichlet and Wishart distributions can be derived without much difficulty. So this type of argument seems to be quite generally applicable to the problem of hyperparameter estimation for Bayesian speaker adaptation. In principle, it could be applied in any situation where there is some tying between the hyperdistributions so that multiple posteriors can be evaluated for each prior. (The usual assumption is that hyperdistributions are tied across speakers rather than across mixture components as in the case at hand.) Of course the question of whether or not this approach is computationally tractable would have to be settled on a case-by-case basis. For example, it turns out to be computationally intractable when applied to the problem of estimating intra-speaker correlation matrices (which is why we had to develop another EM algorithm for that purpose).

III. ADAPTATION AND ESTIMATION PROCEDURES

We describe the MAP adaptation procedure and the procedures for estimating \mathbf{U} and Σ as a series of propositions. We omit the proofs of Propositions 1 and 2 since they can easily be established by adapting the arguments in [3] and we defer the proof of Proposition 3 to the appendix.

Throughout this section we assume that the training data has been aligned with either the speaker-independent or speaker-adapted HMM's, so that each frame is labeled by a mixture component. For each mixture component c , we denote by \mathcal{X}_c the entire collection of frames (that is, collected over all speakers) which are accounted for by the given mixture component. In order to calculate the posterior distribution of \mathbf{x}_c given \mathcal{X}_c we need to extract the following statistics from \mathcal{X}_c . For each speaker s let $N_c(s)$ be the number of frames in the training data for speaker s which are accounted for by the given mixture component. Set

$$S_{X,c}(s) = \sum_t (X_t - \mu_c)$$

where the sum extends over all frames for speaker s that are aligned with the mixture component c . (If it happens that $N_c(s) = 0$ for a given mixture component c then we set $S_{X,c}(s) = 0$.) Let \mathbf{N}_c be the $SF \times SF$ block diagonal matrix whose diagonal blocks are $N_c(1)\mathbf{I}, \dots, N_c(S)\mathbf{I}$ where \mathbf{I} denotes the $F \times F$ identity matrix. Let $\mathbf{S}_{X,c}$ be the $SF \times 1$ vector obtained by concatenating $S_{X,c}(1), \dots, S_{X,c}(S)$. Let Σ_c be the $SF \times SF$ block diagonal matrix each of whose blocks are identical to Σ_c and set

$$\mathbf{l}_c = \mathbf{I} + \mathbf{U}^* \Sigma_c^{-1} \mathbf{N}_c \mathbf{U}.$$

Proposition 1: For each mixture component c , the posterior distribution of \mathbf{x}_c given \mathcal{X}_c and parameters \mathbf{U} and Σ_c is Gaussian with mean

$$\mathbf{l}_c^{-1} \mathbf{U}^* \Sigma_c^{-1} \mathbf{S}_{X,c}$$

and covariance matrix \mathbf{l}_c^{-1} .

Corollary 1: If, for each mixture component c , we denote the posterior mean and covariance of \mathbf{O}_c by $\hat{\mathbf{O}}_c$ and $\hat{\mathbf{A}}_c$, then

$$\begin{aligned}\hat{\mathbf{O}}_c &= \mathbf{U}\mathbf{l}_c^{-1}\mathbf{U}^*\Sigma_c^{-1}\mathbf{S}_{X,c} \\ \hat{\mathbf{A}}_c &= \mathbf{U}\mathbf{l}_c^{-1}\mathbf{U}^*.\end{aligned}$$

This corollary is the key to MAP speaker adaptation. Invoking the Bayesian predictive classification principle [12] just as in [3], we can adapt both the mean vectors and the variances in the speaker-independent HMM to a given speaker s as follows. With each mixture component c , we associate a mean vector given by the expression

$$\mu_c + \hat{\mathbf{O}}_c(s) \quad (1)$$

and a covariance matrix given by

$$\Sigma_c + \hat{\mathbf{A}}_c(s) \quad (2)$$

where $\hat{\mathbf{A}}_c(s)$ is the s th block of $\hat{\mathbf{A}}_c$ (when $\hat{\mathbf{A}}_c$ is considered as an $S \times S$ block matrix with each block being of dimension $F \times F$). Strictly speaking (2) is not variance adaptation but a method of incorporating the uncertainty in the estimate of \mathbf{O}_c into the HMM parameters for each speaker. Our experience has been that it gives mixed results. We refer the reader to [3] for a discussion of this question.

Corollary 2: If, for each mixture component c , $E[\mathbf{x}_c]$ denotes the posterior expectation of \mathbf{x}_c given \mathcal{X}_c and parameters \mathbf{U} and Σ_c and likewise for $E[\mathbf{x}_c\mathbf{x}_c^*]$ then

$$\begin{aligned}E[\mathbf{x}_c] &= \mathbf{l}_c^{-1}\mathbf{U}^*\Sigma_c^{-1}\mathbf{S}_{X,c} \\ E[\mathbf{x}_c\mathbf{x}_c^*] &= E[\mathbf{x}_c]E[\mathbf{x}_c]^* + \mathbf{l}_c^{-1}.\end{aligned}$$

This also follows immediately from Proposition 1 and it is the key to implementing the E-step of the EM algorithm described in Proposition 3 below.

Although it is not strictly necessary, calculating the likelihood of the training data under the assumptions (1), (2) and (3) provides a useful diagnostic for verifying the implementation of the EM algorithm. To explain how this computation is performed we introduce the following notation. For each mixture component c and speaker s , let

$$S_{XX^*,c}(s) = \sum_t (X_t - \mu_c)(X_t - \mu_c)^*$$

where the sum extends over all frames for speaker s that are aligned with the given mixture component. For each mixture component c , let $G_{\Sigma_c}(c)$ denote the Gaussian log likelihood function given by the expression

$$\sum_{s=1}^S \left(N_c(s) \ln \frac{1}{(2\pi)^{F/2} |\Sigma_c|^{1/2}} - \frac{1}{2} \text{tr}(\Sigma_c^{-1} S_{XX^*,c}(s)) \right). \quad (3)$$

(This is the likelihood of \mathcal{X}_c calculated according to our assumptions (1), (2) and (3) in the case where $\mathbf{A} = \mathbf{0}$.)

Proposition 2: If, for each mixture component c , $P_{\mathbf{U},\Sigma_c}(\mathcal{X}_c)$ denotes the total likelihood of \mathcal{X}_c calculated according to assumptions (1), (2) and (3) then

$$\ln P_{\mathbf{U},\Sigma_c}(\mathcal{X}_c) = G_{\Sigma_c}(c) - \frac{1}{2} \ln |\mathbf{l}_c| + \frac{1}{2} \mathbf{S}_{X,c}^* \Sigma_c^{-1} \hat{\mathbf{O}}_c$$

where $G_{\Sigma}(c)$ is defined by formula (3).

Finally, the EM algorithm:

Proposition 3: Suppose we are given initial parameter estimates (\mathbf{U}_0, Σ_0) . For each mixture component c , let $E[\mathbf{x}_c\mathbf{x}_c^*]$ be calculated using these estimates according to Corollary 2 to Proposition 1 and, for each speaker s , let $\hat{\mathbf{O}}_c(s)$ and $\hat{\mathbf{A}}_c(s)$ be calculated according to Corollary 1 to Proposition 1. Let (\mathbf{U}, Σ) be new estimates of the model parameters defined as follows:

$$\mathbf{U} = \mathbf{U}_0 \mathbf{u}$$

where \mathbf{u} is any $R \times R$ matrix such that

$$\frac{1}{C} \sum_{c=1}^C E[\mathbf{x}_c\mathbf{x}_c^*] = \mathbf{u}\mathbf{u}^* \quad (4)$$

and for each $c = 1, \dots, C$, Σ_c is given by

$$\begin{aligned}\frac{1}{n_c} \sum_s \left(S_{XX^*,c}(s) - \left[\hat{\mathbf{O}}_c(s) \mathbf{S}_{X,c}^*(s) + \mathbf{S}_{X,c}(s) \hat{\mathbf{O}}_c^*(s) \right] \right. \\ \left. + \left[\hat{\mathbf{A}}_c(s) + \hat{\mathbf{O}}_c(s) \hat{\mathbf{O}}_c^*(s) \right] N_c(s) \right)\end{aligned}$$

where $n_c = \sum_s N_c(s)$. Then

$$\sum_{c=1}^C \ln P_{\mathbf{U},\Sigma_c}(\mathcal{X}_c) \geq \sum_{c=1}^C \ln P_{\mathbf{U}_0,\Sigma_{c,0}}(\mathcal{X}_c).$$

To interpret the re-estimation formula for the inter-speaker correlation matrix observe that pre-multiplying equation (4) by \mathbf{U}_0 and post-multiplying by \mathbf{U}_0^* gives

$$\frac{1}{C} \sum_{c=1}^C E[\mathbf{O}_c \mathbf{O}_c^*] = \mathbf{A}.$$

This is reasonable since if the \mathbf{O} 's were observable we would use the expression

$$\frac{1}{C} \sum_{c=1}^C \mathbf{O}_c \mathbf{O}_c^*$$

to estimate \mathbf{A} . Note however that

$$\ker(\mathbf{A}_0) = \ker(\mathbf{U}_0^*) \subseteq \ker(\mathbf{U}^*) = \ker(\mathbf{A}).$$

So, since the range of \mathbf{A} is the orthogonal complement of its kernel and similarly for \mathbf{A}_0 , the range of \mathbf{A} is contained in the range of \mathbf{A}_0 . Thus in order to avoid local optima in estimating the inter-speaker correlation matrix using Proposition 3, it is necessary to start with an initial estimate \mathbf{A}_0 of full rank.

Proposition 3 in conjunction with Corollary 1 to Proposition 1 shows how to estimate speaker HMM's for each speaker in a given speaker population but it does not address the problem of how to perform speaker adaptation for a previously unseen speaker. In principle it would be possible to invoke Proposition 3 to estimate a new inter-speaker correlation matrix whenever a previously unseen speaker is encountered but this is hardly practical. The proposition can be modified to develop an efficient procedure to solve this problem but we will not pursue this question here.

IV. IMPLEMENTATION ISSUES

Before describing our experiments, some remarks concerning implementation issues are in order.

Note first that our model exploits the correlations between speakers in a given speaker population to construct HMM's for all of these speakers. Thus it is a multi-speaker modeling technique in the sense of [3] rather than a speaker adaptation technique in the usual sense. In order to carry out recognition experiments on a designated set of test speakers the model has to be applied to a population consisting of the union of the set of test speakers and a standard training set. (The training databases we used were *BREF-80* and *BREF-TOTAL* [21].) Obviously recognition accuracy will depend on the amount of 'adaptation' data collected for each of the test speakers but it is not easy to measure this dependency because the entire model has to be retrained whenever the quantity of adaptation data is changed.

The EM estimation procedure calls for initial estimates of \mathbf{U} and $\mathbf{\Sigma}$; we used random initializations in our experiments. The EM estimation procedure also calls for the first and second order statistics $N_c(s)$, $S_{X,c}(s)$ and $S_{XX^*,c}(s)$ for each mixture component c and speaker s . These statistics can be extracted by aligning the training data with the speaker-independent HMM using either a Viterbi alignment or the Baum-Welch algorithm; we used the Baum-Welch procedure in our experiments. But note that after invoking Proposition 3, we can use MAP adaptation (Corollary 1 to Proposition 1) to construct speaker-adapted models for each of the training speakers and use these to align the training data instead. Accordingly in training the model, we alternate between alignment and EM iterations until the estimates of $(\mathbf{U}, \mathbf{\Sigma})$ converge. Thus our training procedure produces speaker adapted models for each speaker in the extended training set as a byproduct. These speaker-adapted models can be produced by adapting the mean vectors alone (formula (1)) or by adapting the variances as well (formula (2)). We will report experimental results for both types of adaptation.

Implementing the E-step in Proposition 3 calls for inverting $C \ R \times \ R$ matrices where C is the number of mixture components in the speaker-independent HMM and R is the rank of \mathbf{A} . Since the initial estimate of \mathbf{A} has to be of full rank, this will be computationally expensive if C is large (and it generally is in the large vocabulary case). In order to reduce the computational burden we treat the acoustic features as being statistically independent and apply the EM procedure with $F = 1$ in each feature dimension separately in most of the experiments reported below. Thus we do not attempt to address the problem of whether it is worth taking account of correlations between acoustic features in eigenphone modeling. (Since eigenphone modeling is loosely speaking a regression over all speakers, the evidence from MLLR [22], [23] suggests that this question might be worth pursuing; on the other hand the evidence from eigenvoice modeling is mixed [3].)

There is however one situation of interest where it is not necessary to resort to the statistical independence assumption, namely the case where we impose a diagonal or block diagonal

constraint on \mathbf{A} . This type of constraint is equivalent to assuming that the speakers are statistically independent, just as in classical MAP adaptation so it provides a natural benchmark for evaluating the effectiveness of inter-speaker correlation modeling. (But note that the prior specified by a block diagonal \mathbf{A} is not the same as the normal-Wishart prior used in [1].) For these experiments we divided the acoustic features into streams (cepstral coefficients in one stream, their first derivatives in another). We modeled the correlations between the features within each stream but treated the streams as being statistically independent.

If the number of mixture components in a speaker HMM is large (tens or hundreds of thousands) it seems reasonable to use more than one inter-speaker correlation matrix to model the prior distribution of the phone supervectors. Accordingly we used one correlation matrix per phoneme for most of our experiments. This has the drawback that sufficient data has to be collected from each of the test speakers to ensure that each phoneme is adequately represented. Other solutions are possible but they only occurred to us after reading [24]. For example, just as probabilistic principal components analysis can handle multiple correlation matrices, our model can be extended to handle prior distributions on phone supervectors specified by a Gaussian mixture (rather than by a single inter-speaker correlation matrix).

Finally, we used HMM's with one covariance matrix per phoneme in most of our experiments. This creates a complication in applying the variance adaptation procedure given in Corollary 1 to Proposition 1, since the effect of this procedure is to untie the variances. To get around this difficulty, we re-tied the adapted variances by averaging over mixture components within each phoneme.

V. EXPERIMENTS

We carried out multi-speaker recognition experiments on the French language *AUPELF* dictation task to test our model. In all of our experiments we used a 10ms frame rate and a 26-dimensional acoustic feature vector (13-dimensional mel-frequency cepstral coefficients together with their first derivatives).

A. Small HMM

In our first series of experiments we used a small training set, a small language model and a small HMM. Our purpose was to ensure that the eigenphone model was implemented correctly and to compare its performance with speaker-independent training, speaker-dependent training and classical MAP speaker adaptation.

For training we used the *BREF-80* database which comprises 10.6 hours of data collected from 80 speakers (36 male and 44 female). The language model consisted of 311,000 bigrams and 800,000 trigrams and the recognizer's vocabulary contained 20,000 words. The test set consisted of 576 sentences (8,647 words) from *BREF-TOTAL* representing 20 speakers not seen in the training set (8 male and 12 female). The percentage of out-of-vocabulary words in the test set was 3.8.

The speaker-independent HMM was a tied-state triphone model with 782 output distributions each having four mixture components (3,128 Gaussians in all) and one covariance matrix per phoneme. The speaker-independent recognition accuracy (Line 1 of Table I) was 66.4%, admittedly a low figure. We estimated that about 21% of the errors were due to out-of-vocabulary words and another 21% were due to homophone confusions. (A powerful language model is needed to disambiguate homophones in French.)

Speaker-dependent training of the HMM mean vectors using an average of 20 minutes of speech data per speaker gave an accuracy of 73.9% (Line 4 of Table I). This is an upper bound on what one can reasonably hope to achieve with any type of speaker-adaptation procedure.

For each of the speaker adaptation experiments reported below we used 50 utterance files (~ 5 minutes of speech) as adaptation data for each test speaker.

	Type of Adaptation	Accuracy
1	Speaker-independent ($\mathbf{A} = \mathbf{0}$)	66.4%
2	Classical MAP, 2 streams	71.2%
3	Eigenphone MAP, 26 streams	71.9%
4	Speaker-dependent	73.9%

TABLE I

WORD ACCURACIES (AVERAGED OVER 20 SPEAKERS) FOR THE FIRST SERIES OF EXPERIMENTS.

Classical MAP adaptation using 5 minutes of adaptation data per speaker gave a major improvement in accuracy over the speaker-independent result. (71.2% versus 66.4%). In this experiment we followed the procedure described in Section IV, dividing the acoustic feature vector into 2 streams and using one block diagonal inter-speaker correlation matrix per phoneme.

The full inter-speaker correlation model (without block diagonal constraints) gave a small improvement over classical MAP adaptation (71.9% versus 71.2%). We used 2 streams in the latter case and 26 in the former but these results are comparable in all other respects. (We used the same amount of adaptation data and one inter-speaker correlation matrix per phoneme in both cases.)

Thus in the case of a small HMM and a generous amount of adaptation data, classical MAP is capable of achieving close to speaker-dependent performance with little scope for improvement by modeling inter-speaker correlations.

B. Large HMM

Next we carried out a parallel series of experiments using a larger training set and a HMM containing 10 times as many Gaussians as in the first series (3980 output distributions each with 8 mixture components). Again we used one covariance matrix per phoneme. The new training set was a subset of *BREF-TOTAL* consisting of 52.7 hours of speech (29 K sentences) collected from 100 speakers (46 males and 54 females).

We report results using the same language model as in the first series of experiments and also with a much larger language model containing 2.8 million bigrams and 10.6 million

trigrams extracted from 186 million words of newspaper text (*Le Monde*, 1989–1994). We also report results obtained with different amounts of adaptation data: 5 minutes per speaker (as in the first series of experiments) and 17 minutes per speaker. We used the same 20,000 word dictionary as in the first series of experiments.

The experimental set up is roughly comparable to the QO hub in [25]. (The test set was different from ours but it seems to be comparable.) The LIMSI system obtained the best word accuracy (87.2%) in that evaluation. This system used a dictionary consisting of 64K words (resulting in an out-of-vocabulary rate of 1.4%), a language model consisting of 14 million bigrams and 22 million trigrams (extracted from 270 million words of newspaper text), gender-dependent models trained on *BREF-TOTAL* (67K sentences from 120 speakers) and *BREF2* (19K sentences from 298 speakers) and MLLR speaker adaptation.

The results of our second series of experiments are summarized in Table II. When the speaker-independent HMM was

	Type of Adaptation	AD	LM	Accuracy
1	Speaker-Independent	—	small	73.2%
2	Classical MAP, 2 streams	5	small	73.8%
3	Classical MAP, 2 streams	17	small	76.4%
4	Eigenphone MAP, 26 streams	5	small	76.2%
5	Eigenphone MAP, 26 streams	17	small	78.0%
6	Eigenphone MAP, 26 streams + vars	5	small	75.4%
7	Eigenphone MAP, 26 streams + vars	17	small	77.7%
8	Speaker-Independent	—	large	80.6%
9	Eigenphone MAP, 26 streams	5	large	82.8%
10	Eigenphone MAP, 26 streams	17	large	83.6%
11	Eigenphone MAP, 26 streams + vars	5	large	83.1%
12	Eigenphone MAP, 26 streams + vars	17	large	84.2%

TABLE II

WORD ACCURACIES (AVERAGED OVER 20 SPEAKERS) FOR THE SECOND SERIES OF EXPERIMENTS. LM INDICATES LANGUAGE MODEL. AD INDICATES THE AVERAGE AMOUNT OF ADAPTATION DATA PER SPEAKER IN MINUTES.

trained by the Baum-Welch algorithm in the usual way the recognition accuracy obtained on the test set was 73.2% using the small language model (compared with 66.3% in the case of the small HMM) and 80.6% using the large language model. (Lines 1 and 8 of Table II.)

Line 2 of Table II refers to classical MAP adaptation using 5 minutes of adaptation data per speaker. Although classical MAP adaptation gave a major improvement in the first series of experiments we were unable to obtain an improvement here despite extensive tuning of the recognizer. Since the HMM is 10 times larger much more adaptation data seems to be needed. As the result in line 3 shows, classical MAP adaptation with 17 minutes of adaptation data per speaker does indeed result in a substantial improvement (76.4% versus 73.2%).

Lines 4, 5, 9 and 10 refer to mean vector adaptation with full (that is, not block diagonal) speaker correlation matrices. Note that a comparison of lines 3 and 4 shows that adaptation with full speaker correlation matrices and 5 minutes of adaptation data performs almost as well as classical MAP adaptation with 17 minutes of data (76.2% vs 76.4%). Further improvement

is obtained by increasing the amount of adaptation data from 5 minutes to 17 minutes: 78.0% vs 76.2% in the case of the small language model and 83.6% vs 82.8% in the case of the large language model.

The situation in lines 6 and 7 is the same as in lines 4 and 5 except that in addition to adapting the mean vectors using the speaker correlation matrices we also adapted the covariance matrices. Lines 11 and 12 stand in the same relation to lines 9 and 10. As we mentioned in Section III covariance adaptation gives mixed results: we obtained slight improvements with the large language model (83.1% vs 82.8% and 84.2% vs 83.6%) but, despite careful tuning of the recognizer, slight degradations with the small language model (75.4% vs 76.2% and 77.7% vs 78.0%).

C. Eigenphones, Eigenvoices and MLLR

We used the experimental set up described in Section V B of [3] to compare the effectiveness of eigenphone modeling, eigenvoice modeling and MLLR. This set up is the same as in Section V A above, except that (i) each mixture component had its own diagonal covariance matrix, (ii) we used 100 adaptation sentences for each test speaker, (iii) we used a smaller test set comprising 100 sentences (5 from each test speaker) and (iv) we used a single inter-speaker correlation matrix (rather than one per phoneme).

When the speaker-independent HMM was trained on the extended training set consisting of *BREF-80* and the adaptation data for each of the test speakers, the benchmark word recognition accuracy was 72.06%. MLLR gave a word accuracy of 74.2%. The relative improvement here is rather small (7.0%) because the adaptation data has been included in the extended training set. (If the adaptation data is not included in the training set, the benchmark recognition accuracy is 70.9% and MLLR gives 73.7%, that is, a 9.6% relative improvement which is about what one would expect.)

Mean vector adaptation using eigenvoice MAP [3] gave 76.4% recognition accuracy compared with 77.4% for eigenphones. Mean and variance adaptation using eigenvoices gave the same result as adapting only the mean vectors (namely 76.4%) compared with 77.6% for eigenphones.

Thus eigenphone modeling seems to be more effective than MLLR and slightly more effective than eigenvoice modeling although we observed that the HMM likelihoods of the adaptation data were higher in the case of eigenvoice modeling than in the case of eigenphone modeling. Given the enormous number of free parameters in the eigenvoice model, it seems likely that the performance deficit is attributable to over-fitting.

VI. DISCUSSION

We have developed a new way of estimating speaker HMM's for speakers in a closed population using a type of MAP adaptation based on inter-speaker correlation modeling which is dual to the eigenvoice MAP approach in [3]. We found that MAP adaptation using inter-speaker correlation matrices yielded error reductions of 10–20% on a French language dictation task and gave consistently lower error rates than classical MAP adaptation under comparable conditions.

We also obtained better results with eigenphone MAP in back-to-back comparisons with eigenvoice MAP and MLLR.

Our method is a multi-speaker modeling technique rather than a speaker-adaptation technique. It can be extended to perform speaker adaptation to a previously unseen speaker (like classical MAP, eigenvoice MAP and MLLR) although we have not pursued this possibility. Thus in its present form our work is of limited practical use. We believe that the most natural application for multi-speaker modeling is in closed-set speaker recognition [15] but some speech recognition applications can also be suggested. An obvious example is transcribing the debates of legislative assemblies. Multi-speaker modeling may also play a useful role in providing a medium-term solution to the problem of captioning TV broadcasts. In the absence of very high performance speaker-independent speech recognition systems the most effective solution to this problem may be to replace stenographers by ‘parrot’ speakers whose job is to dictate carefully what they hear to a speech recognizer. A multi-speaker recognition capability would be needed since parroting requires intense concentration so that parrot speakers have to work in short rotations.

Our experimental results show that the assumption of statistical independence among speakers (an assumption which is implicit in almost all current speech and speaker modeling) is suboptimal and raises the question of whether it would be possible to construct a reasonable model of the *joint* probability distribution of speech for a given set of speakers which takes account of both inter-speaker and intra-speaker correlations. Given such a joint distribution it should be a straightforward matter to calculate the marginal distribution for each speaker in the population. Note that since the joint distribution has to be estimated from the data for *all* of the speakers, the same is true for each of the marginals. So this line of research would lead to new ways of viewing the problem of acoustic phonetic modeling. (The ‘posterior pooling’ approach of [12], [26] seems to be motivated by similar considerations.)

To see why a good model for the joint distribution of speech could prove to be useful, it is only necessary to consider how much work is needed to build a speech recognizer for a new task or environment. Ideally, what one would like to have is a *universal* speech recognizer—universal in the sense that it can easily be adapted to any given task. In order to construct such a universal speech recognizer it would first be necessary to collect data from a huge number of speakers under all possible conditions—broadcast news, movie sound tracks, telephone conversations, parliamentary debates (or whatever). If a good model for the joint distribution for this entire population can be constructed then the acoustic phonetic modeling problem for a new task or condition should (hopefully) reduce to calculating the appropriate marginal distributions.

A general approach to this problem would be to assume that the two dimensional lattice of offset vectors $\{O_c(s)\}$ has the structure of a hidden Gaussian random field specified by a simple prior correlation structure. The posterior random field (calculated using the two dimensional lattice of first order statistics $\{S_{X,c}(s)\}$) would in the general case be fully connected so that, for each speaker s and mixture component

c , the MAP estimator of the speaker-dependent mean vector $\mu_c(s)$ would be a function of the entire lattice of first order statistics.

In the present article and in [3] we have worked out two simple cases of this model in which the posterior random field is *not* fully connected. The two prior correlation structures that we have considered are $\mathbf{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \mathbf{B}$ where the symbol \otimes denotes the tensor (or Kroeneker) product of two matrices. Natural generalizations (which do result in fully connected posteriors) are structures of the form $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B}$. Unfortunately, although analytical expressions for the posterior means and covariances (along the lines of Proposition 1 in the current article and Proposition 1 in [3]) can be derived without much difficulty in both of these cases, these posteriors seem to be intractable. (At any rate, it is unclear to us how techniques such as approximate matrix inversions, Gibbs sampling or variational Bayes estimation can be brought to bear on them. See [27] for an unsuccessful attempt to experiment with the structure $\mathbf{A} \otimes \mathbf{B}$.) Thus the problem of how to exploit inter-speaker and intra-speaker correlations simultaneously remains beyond our reach.

APPENDIX

The proof of Proposition 3 makes use of the Kullback-Leibler divergence between Gaussian probability distributions. The covariance matrices of these Gaussian distributions are not necessarily invertible so we first have to give a sufficiently general definition of the divergence.

Two probability measures $\hat{\rho}$ and ρ are said to be *equivalent* if there is a strictly positive function ϕ such that $d\rho = \phi d\hat{\rho}$. The function ϕ is called the Radon-Nikodym derivative of ρ with respect to $\hat{\rho}$ and it is denoted by $\frac{d\rho}{d\hat{\rho}}$. If ρ and $\hat{\rho}$ are equivalent, we define the divergence $D(\hat{\rho}||\rho)$ to be

$$- \int \ln \frac{d\rho}{d\hat{\rho}} d\hat{\rho}.$$

In order to calculate the divergence of two equivalent Gaussian distributions defined by singular covariance matrices we use the following properties of the divergence:

- 1) If T is a 1-1 measurable transformation then

$$D(T^* \hat{\rho} || T^* \rho) = D(\hat{\rho} || \rho)$$

where $T^* \rho$ is the probability measure defined by

$$\int f(x) dT^* \rho(x) = \int f(Ty) d\rho(y)$$

(for all positive measurable functions f) and similarly for $T^* \hat{\rho}$.

- 2) If ρ is the Gaussian distribution with mean μ and covariance matrix Σ and $\hat{\rho}$ is the Gaussian distribution with mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ where Σ and $\hat{\Sigma}$ are invertible $N \times N$ matrices then

$$\begin{aligned} D(\hat{\rho} || \rho) &= -\frac{N}{2} - \frac{1}{2} \ln |\hat{\Sigma} \Sigma^{-1}| \\ &\quad + \frac{1}{2} \text{tr} \left(\left[\hat{\Sigma} + (\hat{\mu} - \mu)(\hat{\mu} - \mu)^* \right] \Sigma^{-1} \right). \end{aligned}$$

Lemma 1: Given a parameter set (\mathbf{U}, Σ) let $\pi_{\mathbf{U}}$ be the Gaussian probability measure with mean $\mathbf{0}$ and covariance

matrix $\mathbf{U}\mathbf{U}^$. For each mixture component c , let $\hat{\pi}_c$ be the posterior distribution of \mathbf{O}_c calculated with \mathbf{U} and Σ_c according to Corollary 1 to Proposition 1. Let $P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})$ be the likelihood of \mathcal{X}_c conditioned on the event that $\mathbf{O}_c = \mathbf{O}$. Then*

$$\frac{d\hat{\pi}_c}{d\pi_{\mathbf{U}}}(\mathbf{O}) = \frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\mathbf{U}, \Sigma_c}(\mathcal{X}_c)}.$$

Proof: Note first that if \mathbf{O} is given then \mathbf{U} is not needed to calculate the likelihood of \mathcal{X}_c so we denote this conditional likelihood by $P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})$ rather than by $P_{\mathbf{U}, \Sigma_c}(\mathcal{X}_c|\mathbf{O})$. For any positive measurable functions $f(\mathcal{X})$ and $g(\mathbf{O})$,

$$E[f(\mathcal{X})E[g(\mathbf{O})|\mathcal{X}]] = E[f(\mathcal{X})g(\mathbf{O})].$$

That is,

$$\begin{aligned} &\int f(\mathcal{X}) \left(\int g(\mathbf{O}) d\hat{\pi}_c(\mathbf{O}) \right) dP_{\mathbf{U}, \Sigma_c}(\mathcal{X}) d\mathcal{X} \\ &= \iint f(\mathcal{X}) g(\mathbf{O}) P_{\Sigma_c}(\mathcal{X}|\mathbf{O}) d\mathcal{X} d\pi_{\mathbf{U}}(\mathbf{O}). \end{aligned}$$

Thus

$$d\hat{\pi}_c(\mathbf{O}) dP_{\mathbf{U}, \Sigma_c}(\mathcal{X}) = P_{\Sigma_c}(\mathcal{X}|\mathbf{O}) d\pi_{\mathbf{U}}(\mathbf{O})$$

as required. \blacksquare

Lemma 2: Let $E[\cdot]$ denote the conditional expectation operator $E[\cdot|\mathcal{X}_c]$ evaluated with the initial parameter estimates $\mathbf{U}_0, \Sigma_{c,0}$. Then

$$\begin{aligned} &\ln \frac{P_{\mathbf{U}, \Sigma_c}(\mathcal{X}_c)}{P_{\mathbf{U}_0, \Sigma_{c,0}}(\mathcal{X}_c)} \\ &\geq E[\ln P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})] - E[\ln P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})] \\ &\quad - D(\hat{\pi}_c || \pi_{\mathbf{U}}) + D(\hat{\pi}_c || \pi_{\mathbf{U}_0}) \end{aligned}$$

Proof: This is a straightforward consequence of Jensen's inequality

$$\ln E[U] \geq E[\ln U]$$

applied to the random variable U defined by

$$U = \frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}).$$

We can simplify the left-hand side of the inequality by invoking Lemma 1 with the parameter set (\mathbf{U}_0, Σ_0) . This gives

$$\begin{aligned} &\ln E \left[\frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) \right] \\ &= \ln \int \frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) d\hat{\pi}_c(\mathbf{O}) \\ &= \ln \int \frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) \frac{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})}{P_{\mathbf{U}, \Sigma_{c,0}}(\mathcal{X}_c)} d\pi_{\mathbf{U}_0}(\mathbf{O}) \\ &= \ln \frac{1}{P_{\mathbf{U}, \Sigma_{c,0}}(\mathcal{X}_c)} \int P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O}) d\pi_{\mathbf{U}}(\mathbf{O}) \\ &= \ln \frac{P_{\mathbf{U}, \Sigma_c}(\mathcal{X}_c)}{P_{\mathbf{U}, \Sigma_{c,0}}(\mathcal{X}_c)}. \end{aligned}$$

Simplifying the right hand side of the inequality using the 'chain rule' for Radon-Nikodym derivatives, namely

$$\frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}} = \frac{d\pi_{\mathbf{U}}}{d\hat{\pi}_c} \frac{d\hat{\pi}_c}{d\pi_{\mathbf{U}_0}},$$

gives

$$\begin{aligned}
& E \left[\ln \frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) \right] \\
&= \int \ln \left(\frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) \right) d\hat{\pi}_c(\mathbf{O}) \\
&= \int \ln \left(\frac{P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})}{P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})} \frac{d\pi_{\mathbf{U}}}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) \frac{d\hat{\pi}_c}{d\pi_{\mathbf{U}_0}}(\mathbf{O}) \right) d\hat{\pi}_c(\mathbf{O}) \\
&= \int \left(\ln P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O}) - \ln P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O}) \right. \\
&\quad \left. + \ln \frac{d\pi_{\mathbf{U}}}{d\hat{\pi}_c}(\mathbf{O}) - \ln \frac{d\pi_{\mathbf{U}_0}}{d\hat{\pi}_c}(\mathbf{O}) \right) d\hat{\pi}_c(\mathbf{O}) \\
&= E[\ln P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})] - E[\ln P_{\Sigma_{c,0}}(\mathcal{X}_c|\mathbf{O})] \\
&\quad - D(\hat{\pi}_c \parallel \pi_{\mathbf{U}}) + D(\hat{\pi}_c \parallel \pi_{\mathbf{U}_0})
\end{aligned}$$

as required. \blacksquare

Lemma 2 implies that we can increase the total likelihood function

$$\sum_{c=1}^C P_{\mathbf{U}, \Sigma_c}(\mathcal{X}_c)$$

by choosing \mathbf{U} so as to minimize

$$\frac{1}{C} \sum_{c=1}^C D(\hat{\pi}_c \parallel \pi_{\mathbf{U}})$$

and by choosing Σ_c so as to maximize $E[\ln P_{\Sigma_c}(\mathcal{X}_c|\mathbf{O})]$ for $c = 1, \dots, C$.

Lemma 3: Suppose \mathbf{u} is an $R \times R$ matrix such that $\mathbf{U} = \mathbf{U}_0 \mathbf{u}$. Then, for each mixture component c ,

$$\begin{aligned}
D(\hat{\pi}_c \parallel \pi_{\mathbf{U}}) &= \frac{R}{2} - \frac{1}{2} \ln |(\mathbf{u}\mathbf{u}^*)^{-1}| \\
&\quad + \frac{1}{2} \text{tr} \left(E[\mathbf{x}_c \mathbf{x}_c^*] (\mathbf{u}\mathbf{u}^*)^{-1} \right).
\end{aligned}$$

Proof: Recall that we are assuming that \mathbf{U}_0 is an $SF \times R$ matrix of rank R . Hence \mathbf{U}_0 is 1-1 when it is regarded as a mapping $\mathbb{R}^R \rightarrow \mathbb{R}^{SF}$. Let $\rho_{\mathbf{u}}$ be the Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{u}\mathbf{u}^*$. Let $\hat{\rho}_c$ be the posterior distribution of \mathbf{x}_c . Since \mathbf{U}_0 maps $\rho_{\mathbf{u}}$ onto $\pi_{\mathbf{U}}$ and $\hat{\rho}_c$ onto $\hat{\pi}_c$, property 1 of the Kullback-Leibler divergence implies that

$$D(\hat{\pi}_c \parallel \pi_{\mathbf{U}}) = D(\hat{\rho}_c \parallel \rho_{\mathbf{u}}).$$

The desired result follows by invoking property 2. \blacksquare

Lemma 4: The minimum value of

$$\frac{1}{C} \sum_{c=1}^C D(\hat{\pi}_c \parallel \pi_{\mathbf{U}})$$

subject to the constraint that $\mathbf{U} = \mathbf{U}_0 \mathbf{u}$ is attained by choosing \mathbf{u} so that

$$\frac{1}{C} \sum_{c=1}^C E[\mathbf{x}_c \mathbf{x}_c^*] = \mathbf{u}\mathbf{u}^*$$

Proof: By Lemma 3, we have to minimize

$$-\frac{1}{2} C \ln |\mathbf{a}^{-1}| + \frac{1}{2} \text{tr} \left(\sum_{c=1}^C E[\mathbf{x}_c \mathbf{x}_c^*] \mathbf{a}^{-1} \right).$$

The result follows by differentiating this with respect to \mathbf{a}^{-1} and setting the gradient to be zero. (See the remarks on differentiating functions of a matrix variable in [3].) \blacksquare

Lemma 4 establishes the re-estimation formula (4) for the inter-speaker correlation matrix. It only remains to derive the re-estimation formulas for the residual covariance matrices.

Lemma 5: For each mixture component c ,

$$\ln P_{\Sigma}(\mathcal{X}_c|\mathbf{O}) = G_{\Sigma}(c) + \mathbf{O}^* \Sigma^{-1} \mathbf{S}_{X,c} - \frac{1}{2} \mathbf{O}^* \mathbf{N}_c \Sigma^{-1} \mathbf{O}.$$

Proof: See the proof of Lemma 1 in [3]. \blacksquare

Lemma 6: For each mixture component c , the maximum value of $E[\ln P_{\Sigma}(\mathcal{X}_c|\mathbf{O})]$ is attained by setting Σ equal to

$$\begin{aligned}
\frac{1}{n_c} \sum_s \left(S_{XX^*,c}(s) - [\hat{\mathbf{O}}_c(s) S_{X,c}^*(s) + S_{X,c}(s) \hat{\mathbf{O}}_c^*(s)] \right. \\
\left. + [\hat{\mathbf{A}}_c(s) + \hat{\mathbf{O}}_c(s) \hat{\mathbf{O}}_c^*(s)] N_c(s) \right)
\end{aligned}$$

where $n_c = \sum_s N_c(s)$.

Proof: Note that we can write the equation in the statement of Lemma 5 in the form

$$\begin{aligned}
\ln P_{\Sigma}(\mathcal{X}_c|\mathbf{O}) \\
= G_{\Sigma}(c) + \text{tr} \left(\Sigma^{-1} \left(\mathbf{S}_{X,c} \mathbf{O}^* - \frac{1}{2} \mathbf{O} \mathbf{O}^* \mathbf{N}_c \right) \right)
\end{aligned}$$

and that, just as in Corollary 2 to Proposition 1,

$$\begin{aligned}
E[\mathbf{O}] &= \hat{\mathbf{O}}_c \\
E[\mathbf{O} \mathbf{O}^*] &= \hat{\mathbf{A}}_c + \hat{\mathbf{O}}_c \hat{\mathbf{O}}_c^*.
\end{aligned}$$

Hence

$$\begin{aligned}
& E[\ln P_{\Sigma}(\mathcal{X}_c|\mathbf{O})] \\
&= G_{\Sigma}(c) + E \left[\text{tr} \left(\Sigma^{-1} \left(\mathbf{S}_{X,c} \mathbf{O}^* - \frac{1}{2} \mathbf{O} \mathbf{O}^* \mathbf{N}_c \right) \right) \right] \\
&= G_{\Sigma}(c) + \text{tr} \left(\Sigma^{-1} \left(\mathbf{S}_X \hat{\mathbf{O}}_c^* - \frac{1}{2} (\hat{\mathbf{A}}_c + \hat{\mathbf{O}}_c \hat{\mathbf{O}}_c^*) \mathbf{N}_c \right) \right) \\
&= G_{\Sigma}(c) + \sum_s \text{tr} \left(\Sigma^{-1} \left(S_{X,c}(s) \hat{\mathbf{O}}_c^*(s) \right. \right. \\
&\quad \left. \left. - \frac{1}{2} [\hat{\mathbf{A}}_c(s) + \hat{\mathbf{O}}_c(s) \hat{\mathbf{O}}_c^*(s)] N_c(s) \right) \right).
\end{aligned}$$

In order to differentiate this expression with respect to Σ^{-1} let us write it in the form

$$\sum_{s=1}^S N_c(s) \ln \frac{1}{(2\pi)^{F/2} |\Sigma|^{1/2}} - \frac{1}{2} \text{tr} (\Sigma^{-1} M)$$

where

$$\begin{aligned}
M &= \sum_{s=1}^S \left(S_{XX^*,c}(s) + 2S_{X,c}(s) \hat{\mathbf{O}}_c^*(s) \right. \\
&\quad \left. - [\hat{\mathbf{A}}_c(s) + \hat{\mathbf{O}}_c(s) \hat{\mathbf{O}}_c^*(s)] N_c(s) \right).
\end{aligned}$$

Let V be a matrix of the same dimensions as Σ^{-1} . Note that the derivative of $\ln |\Sigma^{-1}|$ with respect to Σ^{-1} in the direction of V is $\text{tr}(\Sigma V)$. Hence the directional derivative of $E[\ln P_{\Sigma}(\mathcal{X}_c|\mathbf{O})]$ with respect to Σ^{-1} in the direction of V is

$$\frac{1}{2} \sum_{s=1}^S N_c(s) \text{tr}(V \Sigma) - \frac{1}{2} \text{tr}(VM).$$

This expression evaluates to 0 for all symmetric matrices V iff

$$(n_c \Sigma - M) + (n_c \Sigma - M)^* = 0$$

where $n_c = \sum_s N_c(s)$. The desired result follows immediately. (With a little bit of extra work Σ can be shown to be positive definite.) ■

REFERENCES

- [1] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
- [2] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP*, Detroit, Michigan, May 1995, pp. 676–679.
- [3] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, accepted for publication.
- [4] S. Furui, "A training procedure for isolated word recognition systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, Apr. 1980.
- [5] M. Afi fy, Y. Gong, and J.-P. Haton, "Correlation based predictive adaptation of hidden Markov models," in *Proc. Eurospeech*, Rhodes, Greece, Sept. 1997.
- [6] S. Ahadi and P. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 11, pp. 187–206, 1997.
- [7] L. He, J. Wu, *et al.*, "Speaker adaptation based on combination of MAP estimation and weighted neighbour regression," in *Proc ICASSP*, Istanbul, Turkey, June 2000.
- [8] S. Takahashi and S. Sagayama, "Tied-structure HMM based on parameter correlation for efficient model training," in *Proc. ICASSP*, Atlanta, Georgia, May 1996, pp. 467–470.
- [9] S. Chen and P. V. De Souza, "Speaker adaptation by correlation (ABC)," in *Proc. DARPA SLT Workshop*, 1997.
- [10] M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 417–428, 2000.
- [11] Z. Wang and F. Liu, "Speaker adaptation using maximum likelihood model interpolation," in *Proc. ICASSP*, Phoenix, Arizona, Mar. 1999.
- [12] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, pp. 1241–1268, Aug. 2000.
- [13] T. Hazen, "The use of speaker correlation information for automatic speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1998.
- [14] S. Yoshizawa, A. Baba, *et al.*, "Unsupervised speaker adaptation based on sufficient HMM statistics of selected speakers," in *Proc. ICASSP*, Salt Lake City, Utah, May 2001.
- [15] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003.
- [16] B. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 183–191, Mar. 1997.
- [17] Q. Huo and C.-H. Lee, "Online adaptive learning of the correlated continuous-density hidden Markov models for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 386–397, 1998.
- [18] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 276–287, 2001.
- [19] V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis, *et al.*, "Rapid speech recognizer adaptation to new speakers," in *Proc. ICASSP*, Phoenix, Arizona, Mar. 1999.
- [20] P. Kenny, G. Boulianne, and P. Dumouchel, "Bayesian adaptation revisited," in *Proc. ISCA ITRW*, Paris, France, Sept. 2000.
- [21] L. Lamel, J.-L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for French," in *Proc. Eurospeech*, Genoa, Italy, Sept. 1991.
- [22] H. Jin, S. Matsoukas, *et al.*, "Fast robust inverse transform speaker adapted training using diagonal transformations," in *Proc. ICASSP*, Seattle, Washington, May 1998.
- [23] A. Sankar, L. Neumeyer, and M. Weintraub, "An experimental study of acoustic adaptation algorithms," in *Proc. ICASSP*, Atlanta, Georgia, May 1996, pp. 713–716.
- [24] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, pp. 435–474, 1999.
- [25] J. Dolmazon, F. Bimbot, *et al.*, "Première campagne AUPELF d'évaluation des systèmes de dictée vocale: organisation et résultats," in *Ressources et Evaluation en Ingénierie des Langues*, K. Chibout, J. Mariani, N. Masson, and F. Néel, Eds. Louvain-la-Neuve: Champs Linguistiques, De Boeck université, 2000.
- [26] Q. Huo and B. Ma, "Online adaptive learning of continuous-density hidden Markov models based on multiple-stream prior evolution and posterior pooling," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 388–398, 2001.
- [27] P. Kenny, G. Boulianne, and P. Dumouchel, "What is the best type of prior distribution for EMAP speaker adaptation?" in *Proc. Eurospeech 2001*, Aalborg, Denmark, Sept. 2001.

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, RIP) 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 BSc degree in Unified Engineering from the Université du Québec (Chicoutimi) and the MSc degree in Telecommunications from INRS-Télécommunications. He worked on speech analysis and articulatory speech modeling at the Linguistics Department in the Université du Québec (Montreal) until 1990 and then on large vocabulary speech recognition at INRS and Spoken Word Technologies until 1998 when he joined CRIM. His research interests include finite state transducer approaches and practical applications of large vocabulary speech recognition.

Pierre Ouellet obtained the BSc degree in Computer Science from McGill University in 1994. He joined the Ecole 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 received the Ph D degree in Telecommunications from INRS-Télécommunications in 1995. He is currently vice-president R&D of CRIM, a professor at the Ecole de Technologie Supérieure and a board member of the Canadian Language Industry Association (AILIA). His research interests include broadcast news speech recognition, speaker recognition and audio-visual content extraction.

Table I Word accuracies (averaged over 20 speakers) for the first series of experiments.

Table II Word accuracies (averaged over 20 speakers) for the second series of experiments. LM indicates language model. AD indicates the average amount of adaptation data per speaker in minutes.