MAXIMUM LIKELIHOOD ESTIMATION OF EIGENVOICES AND RESIDUAL VARIANCES FOR LARGE VOCABULARY SPEECH RECOGNITION TASKS

P. Kenny, G. Boulianne and P. Dumouchel

Centre de recherche informatique de Montréal (CRIM)
{pkenny, gboulian, pdumouch}@crim.ca

ABSTRACT
We describe a new algorithm for estimating eigenvoices (or, equivalently, EMAP correlation matrices) for large vocabulary speech recognition tasks. The algorithm is an EM procedure based on a novel maximum likelihood formulation of the estimation problem which is similar to the mathematical model underlying probabilistic principal components analysis. It enables us to extend eigenvoice/EMAP adaptation in a natural way to adapt variances as well as mean vectors. It differs from other approaches in that it does not require that speaker dependent or speaker adapted models for the training speakers be given in advance (these are derived as a byproduct of the estimation procedure). Accordingly our algorithm can be applied directly to large vocabulary tasks even if the training data is sparse in the sense that only a small fraction of the total number of Gaussians is observed for each training speaker.

1. INTRODUCTION
Let $C$ be the number of mixture components in a given speaker-independent HMM and let $F$ be the dimensionality of the acoustic feature vectors. We consider the problem of maximum likelihood estimation of the $CF \times CF$ covariance matrix $\mathbf{B}$ which is used in EMAP (extended MAP) speaker adaptation [8]. The assumptions underlying EMAP adaptation can be formulated as follows:

1. For each mixture component $c$ there is an $F \times 1$ mean vector $\mu_c$ and an $F \times F$ covariance matrix $\Sigma_c$.

2. For each (training or test) speaker $s$ and each mixture component $c$, there is an unobservable $F \times 1$ vector $O_c(s)$ ($O$ for offset) such that if $Y$ is an observation from mixture component $c$ for speaker $s$ then

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

where $E$ is normally distributed with mean 0 and covariance matrix $\Sigma_c$.

3. If, for each (training or test) speaker $s$, $O(s)$ is the $CF \times 1$ vector obtained by concatenating the $F \times 1$ vectors $O_1(s), \ldots, O_C(s)$, then the a priori distribution of $O(s)$ is normal with mean $0$ and covariance matrix $\mathbf{B}$.

Once the matrix $\mathbf{B}$ has been estimated, the EMAP adaptation procedure can be used to adapt the speaker independent HMM to a given training or test speaker. In this paper we are concerned with how to estimate the matrix $\mathbf{B}$ from a given training set in the large vocabulary case.

This problem is equivalent to the problem of estimating the eigenvoices of the training set as we now explain briefly. Suppose that $S$ is the number of speakers in the training set and $\hat{\mathbf{B}}$ is the matrix defined by

$$\hat{\mathbf{B}} = \frac{1}{S} \sum_s O(s)O(s)^*$$

where the sum extends over all speakers in the training set. Departing slightly from the usage in [5], we define the eigenvoices of the training set to be the eigenvectors of $\hat{\mathbf{B}}$. If assumption 3 is given, then $\hat{\mathbf{B}}$ is an unbiased estimate of $\mathbf{B}$ so the problem of estimating the eigenvoices is equivalent to that of estimating $\mathbf{B}$. In the large vocabulary case the training data is generally so sparse that speaker dependent training for the training speakers is not feasible. Thus the vectors $O(s)$ and hence the matrix $\hat{\mathbf{B}}$ are unobservable making it difficult to see how the problem of estimating the eigenvoices could be tackled directly. Our purpose in this paper is to show that the equivalent problem of estimating $\mathbf{B}$ is tractable even under these conditions.

More specifically, we will show how to estimate both $\mathbf{B}$ and the residual covariance matrices $\Sigma_c$ in such a way as to optimize the likelihood of a given labeled training set, that is, a training set in which each frame is labeled by a mixture component (as in a Viterbi alignment). The likelihood function referred to here is the likelihood calculated according to assumptions 1, 2 and 3 above rather than the usual HMM likelihood function (see Proposition 2 for a computational formula). By providing estimates of the variability in the training data which is not accounted for by the eigenvoices (i.e. the residual covariance matrices $\Sigma_c$) our approach enables us to extend eigenvoice/EMAP adaptation in a natural way to handle variance as well as mean vector adaptation.
Our starting point for deriving the estimation procedure is the observation that it is not necessary that the matrix \( B \) be of full rank in order for EMAP adaptation to be applicable. This can be seen by inspecting the formulas for speaker adaptation given in Corollary 1 to Proposition 1 below: there is no requirement that \( B \) be invertible. (In fact working with correlation matrices of low rank has enormous computational advantages. This is implicit in Botterweck [1] but it does not seem to have been noticed by other authors.)

Indeed, in large vocabulary applications it is difficult to see how a matrix \( B \) of dimension \( CF \times CF \) could be estimated if it is required to be of full rank. This is because the number of eigenvoices that can be estimated from a training set comprising \( S \) speakers is at most \( S \) and in large vocabulary tasks it is generally the case that \( S \ll CF \). In view of this we reformulate our assumptions 1, 2 and 3 as follows: If \( R \) is the rank of \( B \) then we write

\[
B = VV^* \\
O(s) = Vx(s)
\]

where \( V \) is a \( CF \times R \) matrix and, for each speaker \( s \), \( x(s) \) is an \( R \times 1 \) random vector whose prior distribution is normal with mean \( 0 \) and covariance matrix \( I \).

Of course we are not in a position to know the rank of \( B \) ahead of time so \( R \) has to be fixed arbitrarily (subject to the condition that \( R \leq S \) since there is nothing to be gained by assuming a larger value for \( R \)). The fact that the number of eigenvoices that can be estimated is at most \( S \) implies that the set of eigenvoices may be incomplete in the sense that (contrary to assumption 3) the space spanned by the eigenvoices fails to contain \( O(s) \) for all test speakers \( s \). In this case the effectiveness of EMAP adaptation could be compromised. Botterweck [1] suggests that this problem should be dealt with by forcing \( B \) to be of full rank and he shows how this can be done without making EMAP adaptation computationally intractable. The strategy we adopt is to make a statistical independence assumption about the acoustic features and estimate \( S \) eigenvoices in each feature dimension separately. This boosts the number of eigenvoices in the joint feature space from \( S \) to \( SF \). (A little thought shows that this is equivalent to imposing a sparsity condition on \( B \).)

2. ADAPTATION AND ESTIMATION PROCEDURES

Let \( \Sigma \) be the \( CF \times CF \) block diagonal matrix whose diagonal blocks are \( \Sigma_1, \ldots, \Sigma_C \). The algorithm for estimating \( V \) and \( \Sigma \) is an EM algorithm in which the the role of the hidden variables is played by the random vectors \( x(s) \) (rather than by the random vectors \( O(s) \) as in [4]). We state the estimation formulas without proof in Proposition 3. Readers familiar with probabilistic principal components analysis [7] will be able (with effort) to derive these formulas by adapting the arguments in [7]. (Probabilistic principal components analysis has previously been used as for speaker adaptation in a digit recognition task [3].)

We note in passing that the assumption of a single correlation matrix \( B \) common to all speakers (assumption 3 above) could be improved upon by mixture modeling as in [7]. However we have not explored this possibility since the training set we are working with contains only 80 speakers and more speakers may be needed to estimate even one complete set of eigenvoices.

**Notation** Fix a training or test speaker \( s \). Assuming that the data for the speaker has been aligned with either the speaker-independent or a speaker-adapted HMM, so that each frame is labeled by a mixture component, we denote by \( X(s) \) the entire collection of labeled frames for the speaker.

We extract the following statistics from \( X(s) \). For each mixture component \( c \), let \( N_c(s) \) be the number of observation vectors in the training data for speaker \( s \) which are accounted for by the given mixture component. Let \( S_{XX}^c(s) \) be the \( F \times 1 \) vector of first order moments of these observations about the speaker independent mean vector \( \mu_c \) and let \( S_{XX^*}^c(s) \) be the \( F \times F \) matrix of second order moments. (If it happens that \( N_c(s) = 0 \) for a given mixture component \( c \) then we set \( S_{X}^c(s) = 0 \) and \( S_{XX^*}^c(s) = 0 \).)

Let \( N(s) \) be the \( CF \times CF \) block diagonal matrix whose diagonal blocks are \( N_1(s)I \ldots N_C(s)I \) where \( I \) denotes the \( F \times F \) identity matrix. Let \( S_X(s) \) be the \( CF \times 1 \) vector obtained by concatenating \( S_{X,1}(s), \ldots, S_{X,C}(s) \). Let \( S_{XX}^*(s) \) be the \( CF \times CF \) block diagonal matrix whose diagonal blocks are \( S_{XX,1}(s), \ldots, S_{XX,C}(s) \).

Let \( I(s) \) be the \( R \times R \) matrix defined by

\[
I(s) = I + V^* \Sigma^{-1} N(s)V.
\]

Note that \( I(s) \) is positive definite and hence invertible (and since it is of low dimension its inverse can be inverted without difficulty).

Let \( G(s) \) denote the Gaussian log likelihood function given by the expression

\[
\sum_{c=1}^{C} \left( N_c(s) \ln \left( \frac{1}{(2\pi)^{F/2} |\Sigma_c|^{1/2}} \right) - \frac{1}{2} \text{tr} \left( \Sigma_c^{-1} S_{XX^*}^c(s) \right) \right).
\]

(This is the likelihood of \( X(s) \) calculated according to assumptions 1, 2 and 3 in the case where \( B = 0 \).)

**Proposition 1** For each training or test speaker \( s \), the posterior distribution of \( x(s) \) given \( X(s) \) and a parameter set \( (V, \Sigma) \) is Gaussian with mean

\[
I^{-1}(s)V^* \Sigma^{-1} S_X(s)
\]

and covariance matrix \( I^{-1}(s) \).


Corollary 1 If, for each training or test speaker \( s \), we denote the posterior mean and covariance of \( \mathbf{O}(s) \) by \( \hat{\mathbf{O}}(s) \) and \( \hat{\mathbf{B}}(s) \), then

\[
\hat{\mathbf{O}}(s) = \mathbf{V}^{-1}(s)\mathbf{V}^\top \Sigma^{-1} \mathbf{S}_X(s),
\]
\[
\hat{\mathbf{B}}(s) = \mathbf{V}^{-1}(s)\mathbf{V}^\top.
\]

This corollary is the key to EMAP speaker adaptation. Invoking the Bayesian predictive classification principle [6], 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

\[
\mu_c + \hat{\mathbf{O}}_c(s)
\]

and a covariance matrix given by

\[
\Sigma_c + \hat{\mathbf{B}}_{cc}(s)
\]

where \( \hat{\mathbf{B}}_{cc}(s) \) is the \( cc \)th entry of \( \hat{\mathbf{B}}(s) \) when \( \hat{\mathbf{B}}(s) \) is considered as a \( C \times C \) block matrix (each block being of dimension \( F \times F \)).

Corollary 2 If, for each training speaker \( s \), \( E[\mathbf{x}(s)] \) denotes the posterior expectation of \( \mathbf{x}(s) \) given \( \mathbf{X}(s) \) and a parameter set \( (\mathbf{V}, \Sigma) \) and likewise for \( E[\mathbf{x}(s)\mathbf{x}^*(s)] \) then

\[
E[\mathbf{x}(s)] = \mathbf{1}^{-1}(s)\mathbf{V}^\top \Sigma^{-1} \mathbf{S}_X
\]
\[
E[\mathbf{x}(s)\mathbf{x}^*(s)] = E[\mathbf{x}(s)] E[\mathbf{x}(s)^*] + \mathbf{1}^{-1}(s).
\]

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.

Proposition 2 If \( P(\mathbf{X}(s)|\mathbf{V}, \Sigma) \) denotes the total likelihood of \( \mathbf{X}(s) \) calculated according to assumptions 1, 2 and 3 then \( \ln P(\mathbf{X}(s)|\mathbf{V}, \Sigma) \) is given by the expression

\[
G(s) - \frac{1}{2} \ln |\mathbf{I}(s)| + \frac{1}{2} \hat{\mathbf{O}}^*(s)\Sigma^{-1} \mathbf{S}_X(s).
\]

Summing this expression over all speakers \( s \) in the training set gives the likelihood function whose values are guaranteed to increase on successive iterations of the EM estimation procedure described in Proposition 3.

Proposition 3 Suppose we are given initial parameter estimates \( (\mathbf{V}_0, \Sigma_0) \). For each training speaker \( s \), let \( E[\mathbf{x}(s)] \) and \( E[\mathbf{x}(s)\mathbf{x}^*(s)] \) be the first and second moments of \( \mathbf{x}(s) \) calculated with these estimates according to Corollary 2. Let \( (\mathbf{V}, \Sigma) \) be new estimates of the model parameters defined as follows: \( \mathbf{V} \) is the solution of the equation

\[
\sum_s \mathbf{N}(s)\mathbf{V} E[\mathbf{x}(s)\mathbf{x}^*(s)] = \sum_s \mathbf{S}_X(s) E[\mathbf{x}(s)]^*
\]

and for each \( c = 1, \ldots, C \),

\[
\Sigma_c = \frac{1}{n_c} \sum_s (S_{XX^*c}(s) - \alpha_c(s))
\]

where \( n_c = \sum_s N_c(s) \) and \( \alpha_c(s) \) denote the \( cc \)th diagonal block of the \( CF \times CF \) matrix \( \mathbf{S}_X(s) E[\mathbf{x}(s)]^* \mathbf{V}^\top \). Then

\[
\sum_s \ln P(\mathbf{X}(s)|\mathbf{V}, \Sigma) \geq \sum_s \ln P(\mathbf{X}(s)|\mathbf{V}_0, \Sigma_0)
\]

where the sums extend over all speakers in the training set.

Implementation issues In order to apply the EM estimation procedure we need the first and second order statistics \( \mathbf{N}(s), \mathbf{S}_X(s) \) and \( \mathbf{S}_{XX^*}(s) \) for each training speaker \( s \). These statistics can be extracted by aligning the training data with the speaker-independent HMM. However after estimates of \( \mathbf{V} \) and \( \Sigma \) have been obtained we can apply Corollary 1 to construct speaker adapted models for each training speaker and use these to align the training data instead. Accordingly, in the training phase of our experiments, we alternate between alignment and EM iterations until the estimates of \( (\mathbf{V}, \Sigma) \) converge. (A similar issue arises when Corollary 1 is invoked to construct speaker-adapted models for a given test speaker. We deal with it in the same way by performing several alignment iterations on the speaker’s adaptation data.)

Note that unlike Botterweck’s [1] our estimation procedure does not require that speaker dependent or speaker adapted models for the training speakers be available at the outset. (Speaker adapted HMMs for the training speakers are derived as a byproduct.) It follows that our algorithm can be applied directly to large vocabulary tasks even if the training data is sparse in the sense that only a small fraction of the total number of Gaussians is observed for each training speaker.

In order to palliate the incompleteness problem in the large vocabulary case, we implemented the EM algorithm (with \( F = 1 \)) in each feature dimension separately. (As we mentioned in the introduction, this effectively increases the number of eigenvoices that can be estimated from \( S \) to \( SF \)).

3. EXPERIMENTS

We carried out some pilot experiments on the French language AUPELF task [2] using the BREF-80 training set which consists of 10.6 hours of data collected from 80 speakers. The speaker-independent HMM had 3128 Gaussians, which consists of 10.6 hours of data collected from 80 speakers. However, in the training phase of our experiments, we alternate between alignment and EM iterations until the estimates of \( (\mathbf{V}, \Sigma) \) converge. (A similar issue arises when Corollary 1 is invoked to construct speaker-adapted models for a given test speaker. We deal with it in the same way by performing several alignment iterations on the speaker’s adaptation data.)

Note that unlike Botterweck’s [1] our estimation procedure does not require that speaker dependent or speaker adapted models for the training speakers be available at the outset. (Speaker adapted HMMs for the training speakers are derived as a byproduct.) It follows that our algorithm can be applied directly to large vocabulary tasks even if the training data is sparse in the sense that only a small fraction of the total number of Gaussians is observed for each training speaker.

In order to palliate the incompleteness problem in the large vocabulary case, we implemented the EM algorithm (with \( F = 1 \)) in each feature dimension separately. (As we mentioned in the introduction, this effectively increases the number of eigenvoices that can be estimated from \( S \) to \( SF \)).
and 80,000 trigrams) and a dictionary containing 20,000 words.

For the test set we chose 20 speakers (not represented in the training set) and five sentences per speaker for a total of 1,435 words of which 3.0% were out of vocabulary.

3.1. Speaker Adaptation

We estimated 80 eigenvoices and performed supervised adaptation with various adaptation sets comprising 1, 2, 5, 10, 15, 20 and 100 sentences per speaker; the average length of a sentence was 6 seconds. Recognition results averaged over the 20 test speakers are reported in Table 1. The table shows that adapting the mean vectors is effective with small amounts of adaptation data and that performance saturates after 20 sentences (2 minutes of speech). It is also apparent that adapting the variances as well as the means gives less satisfactory results. We observed that adapting the variances produced generally lower likelihoods for the adaptation sets than adapting the means alone; this suggests that the problem may be due to incompleteness of the set of eigenvoices.

![Table 1](image)

**Table 1.** Recognition accuracies averaged over 20 speakers. S indicates the number of adaptation sentences, M indicates mean adaptation, MV indicates mean and variance adaptation.

3.2. Multi-Speaker Modeling

Another way of producing speaker adapted models for the test speakers is to add the adaptation data to the training data and estimate the model parameters \((\mathbf{V}, \Sigma)\) using the extended training set. This increases the number of eigenvoices that can be estimated (from 80 to 100 in the case at hand) and so enables us to sidestep the incompleteness problem if sufficient adaptation data is available. We refer to this type of training as multi-speaker modeling since it only produces speaker adapted models for the speakers in the extended training set.

In order to see if EMAP adaptation was adversely affected by the incompleteness problem we performed a multi-speaker recognition experiment using 100 adaptation sentences for each of the test speakers. Performing Baum-Welch estimation of the speaker independent HMM with the extended training set and running recognition on the test speakers gave a new benchmark recognition accuracy of 72.06%. Performing mean adaptation on all of the speakers in the extended training set in the course of estimating \((\mathbf{V}, \Sigma)\) and running recognition on the test speakers gave a recognition accuracy of 76.38%; adapting both the means and the variances also led to a recognition accuracy of 76.38%. Comparing these figures with Table 1, we conclude that incompleteness is indeed a problem for variance adaptation but not for mean adaptation.

4. REFERENCES


