Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Joint Factor Analysis for Text-Dependent Speaker Verification

P. Kenny[1]*, T. Stafylakis*[1]*, J. Alam*[1]*, P. Ouellet*[1] *and M. Kockmann*[2]

[1]Centre de Recherche Informatique de Montreal (CRIM), Canada
[2]VoiceTrust, Canada
Patrick.Kenny@crim.ca

Abstract

We tackle the problem of text-dependent speaker verification using a version of Joint Factor Analysis (JFA) in which speaker-phrase variability is modeled with a factorial prior and channel variability with a subspace prior. We implemented this using Zhao and Dong's variational Bayes algorithm, an extension of Vogt's Gauss-Seidel method that supports UBM adaptation to the speaker and channel effects in enrollment and test utterances. We report results on the RSR2015 dataset obtained with two types of likelihood ratio and several strategies for UBM adaptation. We found that using a large UBM and decomposing JFA into a feature extractor and a simple back end classifier (in a way broadly analogous to the i-vector/PLDA cascade) gives better results than using likelihood ratios of either type to make verification decisions. This method involves no UBM adaptation other than to the lexical content of utterances and it is based on Vogt's algorithm rather than Zhao and Dong's. It results in an equal error rate of 0.5% on the RSR2015 evaluation set.

1. Introduction

In text-dependent speaker recognition, the classes to be recognized are speaker-phrase combinations rather than speakers as such and it is not generally possible to collect sufficient training data to model speaker-phrase variability using the subspace methods that have proved to be so successful in text-independent speaker recognition. For instance, several authors have found that i-vector based methods are not notably more successful than less sophisticated approaches [1, 2, 3, 4] (and papers cited there). Channel variability, on the other hand, ought to be amenable to subspace modeling in text-dependent speaker recognition just as in text-independent speaker recognition. This suggests that a JFA model based on a factorial prior for speaker-phrase variability and a subspace prior for channel variability may be suitable for text-dependent speaker recognition. We obtained encouraging results using this type of model in [5] but our investigation there was superficial as that paper was primarily concerned with another question, namely using JFA models to provide fixed-dimensional feature representations of utterances and speakers which could serve as alternatives to i-vectors. We did not attempt to explore the traditional role of JFA as a monolithic classifier (as distinct from a feature extractor) in text-dependent speaker recognition and we left unexplained an anomalous result, namely a degradation in performance that occurs when feature vectors are extracted by adapting the Universal Background Model (UBM) to the data. (The result was anomalous because we did not observe this type of degradation for JFA models based on subspace priors.) Our purpose in this paper is to resolve these issues.

In JFA and i-vector modeling it is usual to collect Baum-Welch statistics with a fixed universal background model (UBM). Since JFA was originally conceived as a method to characterize utterances by GMMs adapted from the UBM, it seems unnatural not to use utterance-dependent GMMs to collect Baum-Welch statistics instead. But experience has shown that this practice is warranted as long as there are no gross mismatches between the UBM and individual utterances. (An outstanding exception is the use of vector Taylor series methods to adapt the UBM to additive noise effects [6].) In the early literature on subspace methods (eigenvoices and eigenchannels), this type of adaptation was generally performed [7, 8, 9] and it is important in subspace GMM modeling for speech recognition [10], but experiments with fully-fledged JFA models containing both speaker and channel subspaces suggested that collecting Baum-Welch statistics with a UBM was the most effective procedure [11].

Zhao and Dong [12] developed a variational Bayes treatment of JFA which combines UBM adaptation with Vogt's Gauss-Seidel method [13] in a principled way but they failed to achieve an improvement in performance on a text-independent speaker recognition task (Table V in [12]). This failure may have been due to the fact that, in their implementation, Zhao and Dong performed UBM adaptation at run time but not in training their JFA models (they used the training algorithms in [11] instead). In this paper we will remedy this defect by developing a variational Bayes EM training algorithm for JFA which optimizes the same criterion as the run-time variational Bayes computation in [12].

We used the RSR2015 dataset as a test bed. For our first experiments we used a PLDA-like speaker verification likelihood ratio which Zhao and Dong refer to as "batch variational Bayes". This can be evaluated with or without UBM adaptation. If the UBM is adapted, then it is adapted to both the speaker and channel effects in all enrollment and test utterances and, in our implementation, this adaptation is performed in training the JFA model as well as at run time. UBM adaptation turned out to give mixed results depending on the number of mixture components in the UBM (it is effective in the case of small UBMs but not large ones). On the other hand, we obtained consistent and substantial improvements from a partial UBM adaptation by calculating verification likelihood ratios as in Section III-G of [11]. Adaptation here is partial because the UBM is adapted to the speaker effects (but not to the channel effects) in the enrollment data, and there is *no adaptation to the test data*. Thus the question of how best to do UBM adaptation in calculating likelihood ratios for speaker verification turns out to be quite subtle.

For most of our experiments we used a UBM with 64 Gaussians. We found that the most effective type of UBM adaptation is to the lexical content of individual phrases (using relevance MAP). If speaker verification is performed by calculating likelihood ratios then adapting to the speaker effects in the enrollment data as in [11] is important. Adapting to the channel effects in the test utterance appears to make little difference one way or the other. On the other hand, it seems that adapting to the speaker effects in a single test utterance needs to be avoided for reasons which can be traced to the weakness of factorial priors and the relative scarcity of test data compared to enrollment data. In the case of a UBM with 512 Gaussians, only adaptation to the lexical content of utterances proved to be helpful. The weakness of the factorial prior is exacerbated by increasing the number of mixture components in this way but it turns out that, in the case of very short utterances such as those encountered in text-dependent speaker verification, extracting Baum-Welch statistics using a large, unadapted UBM provides a very good representation of such utterances. This appears to explain why UBM adaptation did not prove to be beneficial in this case.

Our main conclusion is that deploying a JFA model as a feature extractor in conjunction with a simple cosine distance classifier leads to better results than basing speaker verification decisions on likelihood ratio calculations. In order to be comparable, features need to be extracted in the same way from enrollment and test utterances and since UBM adaptation to a single test utterance needs to be avoided, UBM adaptation to enrollment utterances has to be avoided as well. On the other hand, adapting the UBM to the lexical content of individual phrases is very effective just as in the likelihood ratio calculations. Thus our recipe for extracting a feature from a collection of utterances turns out to be very simple: use a phrase-dependent background model to collect Baum-Welch statistics and use Vogt's algorithm rather than Zhao and Dong's to calculate the feature vector.

As for the back end, a straightforward cosine distance classifier with s-norm score normalization works very well. We did not gain any measurable performance improvement by attempting to model session effects in the back end using nuisance attribute projection or PLDA. We conclude that, at least on the RSR2015 test set, the JFA feature extractor is very effective at suppressing session effects.

2. JFA with UBM adaptation

Recall that the general JFA model assumes that, given a UBM with mean supervector $\boldsymbol{m}$ and multiple recordings of a speaker indexed by r, each recording can be modeled by a GMM whose mean supervector has the form

$$\boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}^r + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z} \qquad (1)$$

where the hidden variables $\boldsymbol{x}^r, \boldsymbol{y}$ and $\boldsymbol{z}$ are assumed to have standard normal priors. The hidden variable $\boldsymbol{x}^r$ varies from one recording to another and is intended to model channel effects. In text-independent speaker recognition, the term $\boldsymbol{D}\boldsymbol{z}$ is usually dropped and speakers are characterized by the low-dimensional vector $\boldsymbol{y}$. For text-dependent speaker recognition, we drop the term $\boldsymbol{V}\boldsymbol{y}$ and we use the variables $\boldsymbol{z}$ to characterize speaker-phrase combinations. The prior on $\boldsymbol{z}$ is factorial in the sense that $P(\boldsymbol{z}) = \prod_c P(\boldsymbol{z}_c)$ where c ranges over mixture components and $\boldsymbol{z}_c$ is the part of $\boldsymbol{z}$ that corresponds to mixture component c. In the case of relevance MAP with relevance factor f, the corresponding submatrix $\boldsymbol{D}_c$ of $\boldsymbol{D}$ is defined by the condition that $f\boldsymbol{D}_c^*\boldsymbol{\Lambda}_c\boldsymbol{D}_c$ is the identity matrix where $\boldsymbol{\Lambda}_c$ is the precision matrix of the mixture component [13]. For the expository portion of this paper we will retain both the subspace term $\boldsymbol{V}\boldsymbol{y}$ and the factorial term $\boldsymbol{D}\boldsymbol{z}$ but we will not use the subspace term in our experiments.

If the alignment of acoustic observations with mixture components is given, posterior distributions of the hidden variables can be calculated exactly using the methods in [17], but a much more efficient iterative approach to calculating posterior expectations was developed by Vogt [13] (the Gauss-Seidel method). These posterior calculations are usually implemented by collecting Baum-Welch statistics with the UBM (rather than with utterance dependent GMMs), so that the effect of the hidden variables $\boldsymbol{x}^r, \boldsymbol{y}$ and $\boldsymbol{z}$ in (1) on frame alignments is ignored.

Zhao and Dong [12] showed how this effect could be handled in a variational Bayes calculation which includes Vogt's Gauss-Seidel algorithm as a special case by introducing additional hidden variables to account for the alignment between frames and mixture components. This variational Bayes calculation enables a coherent development of JFA in which probabilities are evaluated by integrating out hidden variables (rather than plugging in point estimates) and hyperparameters can be estimated using a variational Bayes EM algorithm that implements the maximum likelihood II principle.

In this section we will explain Zhao and Dong's variational Bayes calculation and, in the next section, how it can be used to form likelihood ratios for speaker verification. In the appendix, we explain how to develop a variational Bayes EM training algorithm for JFA models that takes account of the extra hidden variables. (Zhao and Dong did not address this question. They trained JFA models using the heuristics in [11] which are plausible only in the case where all Baum-Welch statistics are collected with the UBM.)

2.1. Notation

Firstly, it is convenient to re-write (1) in terms of mean vectors rather than supervectors:

$$\boldsymbol{m}_c + \boldsymbol{U}_c\boldsymbol{x}^r + \boldsymbol{V}_c\boldsymbol{y} + \boldsymbol{D}_c\boldsymbol{z}_c \qquad (2)$$

(c for mixture component) and to set

$$\boldsymbol{W}_c = \begin{pmatrix} \boldsymbol{U}_c & \boldsymbol{V}_c & \boldsymbol{D}_c \end{pmatrix}.$$

We denote the concatenation of $\boldsymbol{x}^r, \boldsymbol{y}$ and $\boldsymbol{z}$ by $\boldsymbol{X}^r$ and the concatenation of $\boldsymbol{x}^r, \boldsymbol{y}$ and $\boldsymbol{z}_c$ by $\boldsymbol{X}_c^r$ so that (2) can be written as $\boldsymbol{m}_c + \boldsymbol{W}_c\boldsymbol{X}_c^r$. Thus if $\boldsymbol{O}_t^r$ denotes the observation at time t in recording r and the hidden variables $\boldsymbol{x}^r, \boldsymbol{y}$ and $\boldsymbol{z}$ are given, we can calculate the conditional probability $P(\boldsymbol{O}_t^r|\boldsymbol{X}_c^r)$ by plugging the mean vector $\boldsymbol{m}_c + \boldsymbol{W}_c\boldsymbol{X}_c^r$ into the Gaussian kernel to obtain

$$\ln P(\boldsymbol{O}_t^r|\boldsymbol{X}_c^r) = \frac{1}{2}\ln\frac{|\boldsymbol{\Lambda}_c|}{(2\pi)^F} - \frac{1}{2}\boldsymbol{\epsilon}_{tc}^{r*}\boldsymbol{\Lambda}_c\boldsymbol{\epsilon}_{tc}^r \qquad (3)$$

where F is the dimension of the acoustic observations, $\boldsymbol{\Lambda}_c$ is the precision matrix for the mixture compnent c in the UBM, and

$$\boldsymbol{\epsilon}_{tc}^r = \boldsymbol{O}_t^r - \boldsymbol{m}_c - \boldsymbol{W}_c\boldsymbol{X}_c^r. \qquad (4)$$

Secondly, for each $t = 1, \ldots, T^r$, we denote the mixture component which accounts for the acoustic observation $\boldsymbol{O}_t^r$ by c_t^r and we set $\boldsymbol{c}^r = c_{1:T^r}^r$. Here T^r is the duration of the rth recording. We use underlining to indicate aggregation over $r = 1, \ldots, R$ so that $\underline{\boldsymbol{c}}$ denotes the set $\{\boldsymbol{c}^1, \ldots, \boldsymbol{c}^R\}$ (and similarly for $\underline{\boldsymbol{X}}$, $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{O}}$).

For each mixture component c, Baum-Welch statistics are defined by

$$\begin{aligned} N_c^r &= \sum_{t=1}^{T^r} Q(c_t^r = c) \\ \boldsymbol{F}_c^r &= \sum_{t=1}^{T^r} Q(c_t^r = c)(\boldsymbol{O}_t^r - \boldsymbol{m}_c) \end{aligned}$$

Here $Q(c_t^r = c)$ is the posterior probability of the event that the observation $\boldsymbol{O}_t^r$ is accounted for by the mixture component c. We define 'whitened' Baum-Welch statistics $\tilde{\boldsymbol{F}}_c^r$ by setting $\tilde{\boldsymbol{F}}_c^r = \boldsymbol{L}_c^{-1}\boldsymbol{F}_c^r$ where $\boldsymbol{L}_c$ is the lower triangular matrix such that $\boldsymbol{L}_c\boldsymbol{L}_c^*$ is the Cholesky decomposition of $\boldsymbol{\Lambda}_c^{-1}$. Similarly we set $\tilde{\boldsymbol{U}}_c = \boldsymbol{L}_c^{-1}\boldsymbol{U}_c$ and likewise for $\tilde{\boldsymbol{V}}_c, \tilde{\boldsymbol{D}}_c$ and $\tilde{\boldsymbol{W}}_c$.

2.2. Variational posterior calculations

To calculate the variational posteriors for $(\underline{\boldsymbol{X}}, \underline{\boldsymbol{c}})$ we assume a factorization

$$Q(\underline{\boldsymbol{X}})Q(\underline{\boldsymbol{c}})$$

which induces a factorization

$$Q(\underline{\boldsymbol{c}}) = \prod_r Q(\boldsymbol{c}^r).$$

It turns out that $Q(\underline{\boldsymbol{X}})$ is Gaussian but intractable so we impose a factorization

$$Q(\underline{\boldsymbol{X}}) = Q(\underline{\boldsymbol{x}})Q(\boldsymbol{y})Q(\boldsymbol{z})$$

which induces a factorization

$$Q(\underline{\boldsymbol{x}}) = \prod_r Q(\boldsymbol{x}^r).$$

Consider first $Q(\underline{\boldsymbol{X}})$. For each recording r,

$$\ln Q(\boldsymbol{x}^r) \equiv E_{\boldsymbol{X}^r \backslash \boldsymbol{x}^r, \boldsymbol{c}^r}\left[\ln P(\boldsymbol{X}^r, \boldsymbol{c}^r, \boldsymbol{O}^r)\right]$$

where $\boldsymbol{X}^r \backslash \boldsymbol{x}^r$ indicates the complement of $\boldsymbol{x}^r$ in $\boldsymbol{X}^r$, that is $\{\boldsymbol{y}, \boldsymbol{z}\}$, and we use $\equiv$ to indicate equality up to an additive constant. This expression is quadratic in $\boldsymbol{x}^r$ so the variational posterior distribution of $\boldsymbol{x}^r$ is Gaussian with precision matrix $\boldsymbol{P}$ and expectation $\langle \boldsymbol{x}^r \rangle$ given by

$$\begin{aligned} &\boldsymbol{I} + \sum_{c=1}^{C} N_c^r \tilde{\boldsymbol{U}}_c^* \tilde{\boldsymbol{U}}_c \\ &\boldsymbol{P}^{-1} \sum_{c=1}^{C} \tilde{\boldsymbol{U}}_c^* (\tilde{\boldsymbol{F}}_c^r - N_c^r \tilde{\boldsymbol{V}}_c \langle \boldsymbol{y} \rangle - N_c^r \tilde{\boldsymbol{D}}_c \langle \boldsymbol{z}_c \rangle) \end{aligned}$$

Similarly for the variational posterior of $\boldsymbol{y}$, the precision matrix $\boldsymbol{P}$ and expectation $\langle \boldsymbol{y} \rangle$ are given by

$$\begin{aligned} &\boldsymbol{I} + \sum_{c=1}^{C} N_c \tilde{\boldsymbol{V}}_c^* \tilde{\boldsymbol{V}}_c \\ &\boldsymbol{P}^{-1} \sum_{r=1}^{R} \sum_{c=1}^{C} \tilde{\boldsymbol{V}}_c^* (\tilde{\boldsymbol{F}}_c^r - N_c^r \tilde{\boldsymbol{U}}_c \langle \boldsymbol{x}^r \rangle - N_c^r \tilde{\boldsymbol{D}}_c \langle \boldsymbol{z}_c \rangle) \end{aligned}$$

and, for each mixture component c, the corresponding expressions for $\boldsymbol{z}_c$ are

$$\begin{aligned} &\boldsymbol{I} + N_c \tilde{\boldsymbol{D}}_c^* \tilde{\boldsymbol{D}}_c \\ &\boldsymbol{P}^{-1} \sum_{r=1}^{R} \tilde{\boldsymbol{D}}_c^* (\tilde{\boldsymbol{F}}_c^r - N_c^r \tilde{\boldsymbol{U}}_c \langle \boldsymbol{x}^r \rangle - N_c^r \tilde{\boldsymbol{V}}_c \langle \boldsymbol{y} \rangle). \end{aligned}$$

Turning now to $Q(\underline{\boldsymbol{c}})$, by the variational update formula,

$$\begin{aligned} Q(\boldsymbol{c}^r) &\equiv E_{\boldsymbol{X}^r}\left[\ln P(\boldsymbol{c}^r, \boldsymbol{O}^r | \boldsymbol{X}^r)\right] \\ &\equiv \sum_{t=1}^{T^r} \langle \ln P(c_t^r, \boldsymbol{O}_t^r | \boldsymbol{X}^r) \rangle \end{aligned}$$

so that

$$\begin{aligned} Q(c_t^r = c) &\equiv \langle \ln P(c_t^r = c, \boldsymbol{O}_t^r | \boldsymbol{X}^r) \rangle \\ &\equiv \ln \pi_c + \langle \ln P(\boldsymbol{O}_t^r | \boldsymbol{X}_c^r) \rangle \end{aligned}$$

where π_c is the mixture weight for component c. Denoting this quantity by $\tilde{\gamma}_{tc}^r$, we have

$$Q(c_t^r = c) = \frac{\tilde{\gamma}_{tc}^r}{z_t^r} \qquad (5)$$

where z_t^r is determined by the condition that probabilities sum to 1. To evaluate $\tilde{\gamma}_{tc}^r$, we can write it in the form

$$\ln \pi_c + \frac{1}{2} \ln \frac{|\boldsymbol{\Lambda}_c|}{(2\pi)^F} - \frac{1}{2} \langle \boldsymbol{\epsilon}_{tc}^{r*} \boldsymbol{\Lambda}_c \boldsymbol{\epsilon}_{tc}^r \rangle .$$

To evaluate $\langle \boldsymbol{\epsilon}_{tc}^{r*} \boldsymbol{\Lambda}_c \boldsymbol{\epsilon}_{tc}^r \rangle$ we can write it as

$$\langle \boldsymbol{\epsilon}_{tc}^{r*} \rangle \, \boldsymbol{\Lambda}_c \, \langle \boldsymbol{\epsilon}_{tc}^r \rangle + \operatorname{tr}\left(\boldsymbol{\Lambda}_c \operatorname{Cov}\left(\boldsymbol{\epsilon}_{tc}^r, \boldsymbol{\epsilon}_{tc}^r\right)\right)$$

and write the second term as $\operatorname{tr}\left(\tilde{\boldsymbol{W}}_c^* \tilde{\boldsymbol{W}}_c \operatorname{Cov}\left(\boldsymbol{X}_c^r, \boldsymbol{X}_c^r\right)\right)$ which simplifies to

$$\begin{aligned} &\operatorname{tr}\left(\tilde{\boldsymbol{U}}_c^* \tilde{\boldsymbol{U}}_c \operatorname{Cov}\left(\boldsymbol{x}^r, \boldsymbol{x}^r\right)\right) \\ &\quad + \operatorname{tr}\left(\tilde{\boldsymbol{V}}_c^* \tilde{\boldsymbol{V}}_c \operatorname{Cov}\left(\boldsymbol{y}, \boldsymbol{y}\right)\right) + \operatorname{tr}\left(\tilde{\boldsymbol{D}}_c^* \tilde{\boldsymbol{D}}_c \operatorname{Cov}\left(\boldsymbol{z}_c, \boldsymbol{z}_c\right)\right). \end{aligned}$$

2.3. Variational Lower Bound

The variational lower bound $\mathcal{L}$ which serves as a proxy for $\ln P(\underline{\boldsymbol{O}})$ is given by

$$\mathcal{L} = E\left[\ln \frac{P(\underline{\boldsymbol{c}}, \underline{\boldsymbol{X}}, \underline{\boldsymbol{O}})}{Q(\underline{\boldsymbol{c}})Q(\underline{\boldsymbol{X}})}\right]$$

where the expectation is taken with respect to the variational posterior of the hidden variables. It is generally convenient to write this sort of expression in the form

$$\langle \ln P(\underline{\boldsymbol{O}} | \underline{\boldsymbol{X}}, \underline{\boldsymbol{c}}) \rangle + \text{negative divergences} \qquad (6)$$

but, in this situation, it is most easily evaluated by writing it in the form

$$E\left[\ln \frac{P(\underline{\boldsymbol{c}}, \underline{\boldsymbol{O}} | \underline{\boldsymbol{X}})}{Q(\underline{\boldsymbol{c}})}\right] - D\left(Q(\underline{\boldsymbol{X}}) || P(\underline{\boldsymbol{X}})\right). \qquad (7)$$

The divergence term can be evaluated using the formula for the divergence of two normal distributions. A simple calculation using (5) shows that the contribution of the first term reduces to

$$\sum_{r=1}^{R} \sum_{t=1}^{T^r} \ln z_t^r .$$

Alternatively, the lower bound can be expressed in terms of Baum-Welch statistics by evaluating the first term in (7) by writing

$$E\left[\ln \frac{P(\boldsymbol{c}^r, \boldsymbol{O}^r | \boldsymbol{X}^r)}{Q(\boldsymbol{c}^r)}\right] = \langle \ln P(\boldsymbol{c}^r, \boldsymbol{O}^r | \boldsymbol{X}^r) \rangle + H(Q(\boldsymbol{c}^r))$$

and writing the first term here as

$$\begin{aligned}
&\sum_{c}\sum_{t=1}^{T^r} Q(c_t^r = c)\ln\tilde{\gamma}_{rc}^{t} \\
&= \quad \sum_{c}\sum_{t=1}^{T^r} Q(c_t^r = c)\left(\ln\pi_c + \frac{1}{2}\ln\frac{|\boldsymbol{\Lambda}_c|}{(2\pi)^F} - \frac{1}{2}\left\langle \boldsymbol{\epsilon}_{tc}^{r*}\boldsymbol{\Lambda}_c\boldsymbol{\epsilon}_{tc}^{r}\right\rangle\right)
\end{aligned}$$

then substituting the expression

$$\begin{aligned}
&\operatorname{tr}\left(\boldsymbol{\Lambda}_c\boldsymbol{S}_c^r\right) - 2\tilde{\boldsymbol{F}}_c^{r*}\tilde{\boldsymbol{W}}_c\langle\boldsymbol{X}_c^r\rangle + N_c^r\langle\boldsymbol{X}_c^r\rangle^*\tilde{\boldsymbol{W}}_c^*\tilde{\boldsymbol{W}}_c\langle\boldsymbol{X}_c^r\rangle \\
&+ N_c^r\operatorname{tr}\left(\tilde{\boldsymbol{W}}_c^*\tilde{\boldsymbol{W}}_c\operatorname{Cov}\left(\boldsymbol{X}_c^r,\boldsymbol{X}_c^r\right)\right)
\end{aligned}$$

for

$$\sum_{t=1}^{T^r} Q(c_t^r = c)\left\langle \boldsymbol{\epsilon}_{tc}^{r*}\boldsymbol{\Lambda}_c\boldsymbol{\epsilon}_{tc}^{r}\right\rangle .$$

Here $\boldsymbol{S}_c^r$ denotes the second order Baum-Welch statistics. (In some situations the contributions of the second order statistics and the entropy $H(Q(\boldsymbol{c}^r))$ can be ignored and need not be calculated.)

3. Three Approaches to Speaker Verification

3.1. Bayesian model selection

Given a speaker verification trial consisting of a collection of enrollment utterances E and a test utterance T, the most straightforward way to make a verification decision is to calculate a likelihood ratio of the form

$$\frac{P(E,T)}{P(E)P(T)} \tag{8}$$

using variational lower bounds as proxies for each of the terms in this expression. Zhao and Dong [12] obtained their best results using this type of likelihood ratio (Table IV).

In experiments in text-dependent speaker recognition on the RSR2015 dataset which we will describe in detail below, we found that doing UBM adaptation (both in JFA training and at verification) gave mixed results when likelihood ratios are evaluated in this way (performance degraded in the case of large UBMs but improved in the case of small UBMs).

To gain some insight into the reasons for this misbehavior, it can be noted that (8) can be interpreted as a Bayesian model selection criterion. The question is whether the union of the enrollment data E and the test data T can be better accounted for by positing a two speaker model or a single speaker model. Referring to the expression for the variational lower bound (6), the first term measures how well a model fits the data and the second term penalizes model complexity. Modeling the data with two $\boldsymbol{z}$ vectors rather than one increases the value of the first term but turns out to have a relatively minor effect on the second term. Rather than having a single divergence of the form $D\left(Q(\boldsymbol{z})||P(\boldsymbol{z})\right)$, two divergences are introduced, one calculated with the enrollment data and the other calculated with the test data. Irrespective of whether or not UBM adaptation is performed at run time, this type of divergence (which is primarily determined by the value of the posterior covariance matrices $\operatorname{Cov}\left(\boldsymbol{z},\boldsymbol{z}\right)$) tends to be small, particularly if it is evaluated with a single, short test utterance because $\boldsymbol{z}$ is of very high dimension and the prior $P(\boldsymbol{z})$ is factorial. (Subspace priors would result in larger divergences.) On the other hand the effect of UBM adaptation is to greatly increase the value of the first term in (6). Thus if UBM adaptation is performed the model selection criterion is apt to break down because the prior on $\boldsymbol{z}$ is too weak.

Henceforth we will refer to (8) as the model selection likelihood ratio.

3.2. JFA as a feature extractor

In [5] we introduced the idea of decomposing a JFA model into a front end and a back end in a manner which is broadly analogous to the i-vector/PLDA cascade. Given a collection of (enrollment or test) utterances by a given speaker, we characterize the speaker by a point estimate of the $\boldsymbol{z}$ vector calculated using Vogt's Gauss-Seidel algorithm. Even though multiple utterances are typically available for enrollment, a single feature vector is extracted at enrollment time and similarly at test time.

Our experience in [5] was that performance degraded when we tried to extract $\boldsymbol{z}$-vectors with UBM adaptation (by training JFA models with UBM adaptation and using Zhao and Dong's algorithm rather than Vogt's at run time). If a $\boldsymbol{z}$-vector is extracted from multiple recordings (as at enrollment time) and UBM adaptation is performed then it is to be expected that the JFA model will succeed in finding a good alignment of the acoustic observations with the mixture components in the adapted UBMs. On the other hand in processing a *single* test utterance with a factorial prior (rather than a subspace prior), the UBM mean vectors are allowed to adapt to the data *independently* of each other, so that the constraints on the way acoustic observations align with mixture components in the adapted UBM are extremely weak. We believe that this asymmetry between enrollment and test data accounts for our lack of success with UBM adaptation in [5].

This suggests that in exploring the question of UBM adaptation with JFA models based on factorial priors, particular attention needs to be paid to the way test utterances are handled.

3.3. Alternative likelihood ratios

In evaluating the numerator of (8) using the variational posterior calculation to iteratively improve the variational lower bound, one way to proceed is to alternate between updating the variational posteriors of (i) $\boldsymbol{y}$ and $\boldsymbol{z}$ (which are tied across all recordings, be they enrollment or test) and (ii) $\underline{\boldsymbol{c}}$ and $\underline{\boldsymbol{x}}$ (which are untied). This is referred to as batch variational Bayes in [12]. It involves processing the enrollment data E from scratch in each verification trial, something that is impractical and intuitively unappealing. An alternative approach (called sequential variational Bayes in [12]) is to split the variational posterior calculation into two stages, one of which is carried out at enrollment time and the other at test time. That is, in stage one alternate between (i) and (ii) using only the enrollment data and in stage two alternate between (i) and (ii) using only the test data. When stage one has been completed, posteriors for the tied hidden variables $\boldsymbol{y}$ and $\boldsymbol{z}$ have been calculated and these posteriors are further updated in stage two. This two stage calculation gives similar results to the straightforward method which makes no distinction between enrollment and test recordings. As such, it cannot be expected to improve on the anomalous result that we obtained with UBM adaptation using the likelihood ratio (8). (The results of the two approaches are similar but not identical because the order in which the variational updates are performed in the two stage approach breaks the symmetry between the enrollment and test data.)

The two stage approach can be thought of as evaluating a likelihood ratio of the form

$$\frac{P(T|E)}{P(T)} \tag{9}$$

which highlights the test utterance T and which we will refer to as a predictive likelihood ratio. Several possibilities can be explored to evaluate the numerator and denominator here. In a two stage calculation such as the one we have described, it is usual to simplify the second stage by setting to 0 the posterior covariance matrices of $\boldsymbol{y}$ and $\boldsymbol{z}$ calculated in stage one. This is equivalent to using a point estimate of a speaker-dependent GMM in stage two which is *not further adapted to the speaker effects in the test utterance* and it is well motivated if there is adequate enrollment data. Depending on the implementation, this speaker-dependent GMM may be adapted to the channel effects in the test utterance (as in eigenchannel modeling where several alignment iterations are performed and speaker verification decisions are made with conventional GMM scoring [8, 9]) or the speaker-dependent GMM may merely be used to collect Baum-Welch statistics (in which case scores for verification decisions are evaluated by integrating over channel factors as in equation (19) in [14]).

A strong argument in favour of adapting the UBM to the speaker effects in the enrollment utterances is that in the case of text-independent speaker recognition with a JFA model based on a factorial prior, it turns out to be better to collect Baum-Welch statistics with speaker-dependent GMMs than with the UBM (whereas the opposite is true in the case of JFA models based on subspace priors). See Section III-G in [11] entitled “Note on collecting Baum-Welch statistics”. For that experiment, we used a JFA model of the form $\boldsymbol{m}+\boldsymbol{D}\boldsymbol{z}+\boldsymbol{U}\boldsymbol{x}^r$ to create speaker-dependent GMMs for target speakers and we used the channel component of this model, namely $\boldsymbol{U}\boldsymbol{x}^r$, to evaluate both the numerator and denominator of (9) by centering the Baum-Welch statistics with the target speaker’s supervector in one case and with the UBM supervector in the other. This ignores the uncertainty in the point estimate of the target speaker’s supervector that arises from the fact that the factorial prior is relatively weak (compared with a subspace prior) and the amount of data available to enroll the target speaker is limited. This consideration led us to adopt a slightly different approach for our experiments here.

We used two factor analysis models, one at enrollment time and the other at verification time, which we trained independently. For enrollment, we trained JFA models of the form $\boldsymbol{m}+\boldsymbol{D}\boldsymbol{z}+\boldsymbol{U}\boldsymbol{x}^r$ (using the VBEM algorithms summarized in the appendix) with and without UBM adaptation. To enroll a target speaker, we created a speaker-dependent GMM by estimating the speaker’s $\boldsymbol{z}$ vector from the enrollment data (using Vogt’s or Zhao and Dong’s algorithm) taking the speaker-dependent supervector to be $\boldsymbol{m}+\boldsymbol{D}\langle\boldsymbol{z}\rangle$.

To train the factor analysis model used at verification time, we first created speaker-dependent GMMs for each of the training speakers using the enrollment JFA model. We then trained a model of the form $\boldsymbol{m}(s)+\boldsymbol{U}\boldsymbol{x}^r$ to account for channel effects at verification time. The notation $\boldsymbol{m}(s)$ here indicates the mean supervector in the model varies from one target speaker to another; formally, this model is just an i-vector extractor with Baum-Welch statistics being centered in different ways for different speakers. Again, we trained this type of model with and without further adapting target speakers’ GMMs.

If a target speaker’s GMM is not adapted at verification time, then evaluating the numerator $P(T|E)$ of (9) using the variational lower bound described in Section 2.3 implements exact integration over the channel factors and so is equivalent to equation (19) in [14]. (Integration is exact here because the JFA model used at verification time is formally just an i-vector extractor so posteriors of the hidden variables are evaluated exactly by the variational Bayes algorithm.) On the other hand, if adaptation is performed at verification time, then acoustic observation vectors in a test utterance are aligned with mixture components in the target speaker’s GMM in such a way as to take account of channel effects. In this case the lower bound calculation can be viewed as combining eigenchannel modeling as implemented in [8, 9] with channel factor integration.

Generally speaking the question of UBM adaptation needs to be investigated because mismatches between the UBM and the enrollment and test data may be sufficiently serious that some sort of compensation is required. In the context of text-dependent speaker recognition, in each verification trial the same phrase is repeated at enrollment and test time but different trials may involve different phrases. For example, there are 30 different phrases in Part I of RSR2015 and lexical variability is the principal source of UBM mismatch in this dataset. Adapting a UBM to compensate for lexical mismatch is straightforward (relevance MAP works well) and we included this type of adaptation in our experiments by modifying the JFA model used at enrollment time to have the form $\boldsymbol{m}(p)+\boldsymbol{D}\boldsymbol{z}+\boldsymbol{U}\boldsymbol{x}^r$ where the notation $\boldsymbol{m}(p)$ indicates that the mean supervector in the JFA model varies from one phrase p to another. Similarly, to evaluate the denominator of (9) we centered the Baum-Welch statistics with the phrase-dependent supervector $\boldsymbol{m}(p)$ rather than the UBM supervector.

Thus we experimented with three distinct types of UBM adaptation in evaluating predictive likelihood ratios: adapting the UBM to produce phrase-dependent background models, adapting phrase-dependent background models to produce speaker-phrase-dependent GMMs (in such a way as to take account of the speaker effects in the enrollment data but not the channel effects) and adapting speaker-phrase models to the channel effects in test utterances. We introduce the notation $a-b-c$ to keep track of the number of alignment iterations performed at each stage. In our experiments we took a to be 0, 1 or 5 (0 indicates that we used the UBM rather than a phrase background model); we took b to be 1 or 5 (1 indicates that a single alignment iteration was performed in enrolling a speaker) and similarly for c.

4. Experiments

4.1. Data

We used the RSR2015 data set for our experiments (using the background set for UBM and JFA training and the Part I evaluation set for testing) [2]. For algorithmic development we used a restricted test set consisting of all of the female trials obtained by selecting all of the target trials and 50 000 high scoring non-target trials. (Low scoring non-target trials are too easy to be interesting. Working with the restricted test set inflates the error rates by a factor of two.)

4.2. Model configurations

We used a standard 60 dimensional front end (MFCCs with short term Gaussianization) and UBMs having 64 and 512 diagonal Gaussians. In all of our experiments (except where otherwise indicated), we took the rank of $\boldsymbol{U}$ to be 50, we used a relevance factor of 2 and we used s-norm for score normalization.

4.3. Benchmarks

We used the Bayesian model selection likelihood ratio (8) for benchmarking. (Results obtained with a standard GMM/UBM approach will be presented later.) We did not use phrase-dependent background models at this stage. The results are summarized in Table 1. In lines 1 and 2, speaker verification likelihood ratios were evaluated using (8). UBM adaptation was performed for line 2 but not for line 1. UBM adaptation is seen to be helpful in the case of 64 Gaussians but not in the case of 512 Gaussians. (Reducing the number of components in the UBM increases the mismatch between the UBM and the data, so this can be expected to favor UBM adaptation.) For line 3 we used a predictive likelihood ratio (9) calculated using two JFA models, one of which is used to enroll speakers and the other to evaluate test utterances, as explained in Section 3.3. A single alignment iteration was used to create a speaker-dependent GMM from the UBM at enrollment time and a single alignment iteration was performed at test time. (Thus the only type of UBM adaptation performed here is adaptation to the speaker effects in the enrollment data and $a = 0$, $b = 1$ and $c = 1$.) Substantial improvements in performance are observed for both UBM configurations.

Table 1: *Restricted test set, 512 component UBM (columns 1 and 2) and 64 component UBM (columns 3 and 4).*

	EER	2008 NDCF	EER	2008 NDCF
1	2.2%	0.085	3.6%	0.145
2	2.7%	0.096	3.4%	0.133
3	**1.7%**	**0.065**	**2.7%**	**0.110**

4.4. Predictive likelihood ratios

In line 3 of Table 1, the denominator of the likelihood ratio is evaluated using the UBM supervector. In effect, we are trying to determine whether the test speaker is closer to the target speaker or the UBM “speaker”. This sort of calculation is typically used in text-independent speaker recognition because, in practice, it works as well as the more principled sequential variational Bayes calculation described in Section 3.3. But it fails to take account of the fact that in text-dependent speaker recognition the lexical content of the test utterance (the “phrase”) is given and this has a greater effect on the acoustics of the test utterance than the identity of the test speaker. Thus we retrained the JFA model used to enroll target speakers by centering the Baum-Welch statistics with phrase-dependent supervectors and we treated the Baum-Welch statistics in the same way in evaluating the denominator of the likelihood ratio. Line 1 of Table 2 shows that adapting the UBM to the lexical content of phrases leads to a substantial gain in performance (compare with line 3 of Table 1). Further minor gains can be obtained by performing multiple alignment iterations on test utterances (line 2) and in enrolling speakers (line 3).

Table 2: *Restricted test set, 64 component UBM adapted to each phrase*

	a-b-c	EER	2008 NDCF
1	1-1-1	2.1%	0.092
2	1-1-5	2.0%	0.086
3	1-5-1	2.0%	0.080

Creating phrase-dependent background models by adapting the UBM with 5 iterations of relevance MAP rather than one, turned out to give further major improvements, as shown in Table 3 (compare with Table 2). The EER of 1.6% (line 3) is the lowest that we achieved in any of our experiments with likelihood ratios on the restricted test set.

Table 3: *Restricted test set, 64 component UBM, phrase adaptation with 5 iterations of relevance MAP*

	a-b-c	EER	2008 NDCF
1	5-1-1	1.7%	0.076
2	5-5-1	1.7%	0.070
3	5-5-5	1.6%	0.066

Another question that we sought to resolve in this paper is whether it is possible to improve on relevance MAP by estimating the matrices $\boldsymbol{D}_c$ using the maximum likelihood II principle rather than relevance MAP. (That is, with the same criterion used to estimate $\boldsymbol{U}$. See the appendix.) It turns out that minor improvements can be achieved in this way. Line 1 of Table 4 is copied from line 2 of Table 3 which was obtained with relevance MAP (relevance factor 2); for line 2, the matrices $\tilde{\boldsymbol{D}}_c$ were constrained to be diagonal and estimated using maximum likelihood II and, for line 3, these matrices were taken to be full. This yielded the best result that we obtained with likelihood ratios on the restricted test set as measured by the 2008 detection cost function. (And likewise for the 2010 cost function, whose minimum value was a respectable 0.211.) An aspect of this experiment which took us by surprise is that when the matrices $\boldsymbol{D}_c$ are not constrained to be diagonal, they turn out to be of low rank (so that, for each mixture component, speaker variation is confined to a low dimensional subspace of the acoustic feature space). This provides intuitive support for the acoustic factor analysis in [15].

Table 4: *Restricted test set, 64 component UBM, relevance MAP vs maximum likelihood II*

	a-b-c	EER	2008 NDCF
1	5-5-1	1.7%	0.070
2	5-5-1	1.7%	0.069
3	5-5-1	1.7%	0.065

We replicated the $5-1-1$ and $5-5-5$ experiments from Table 3 with 512 Gaussians rather than 64 and found that $5-1-1$ works best in this case. Results are presented in Table 5. Comparing line 1 of this table with line 3 of Table 1, we note a modest improvement (attributable to the phrase-dependent background models). But comparing with lines 1 and 2 of Table 1 shows that multiple alignment iterations at enrollment and test time (the 5-5-5 configuration in Table 5) leads to degradations in performance in the case of 512 Gaussians. This is consistent with our initial observation that UBM adaptation was not helpful in evaluating likelihood ratios of the form (8) in the case of 512 Gaussians, although it did work in the case of 64 Gaussians.

The results in line 3 and 4 of Table 5 were obtained by using a JFA model with 512 component phrase-dependent background models as a feature extractor and a cosine distance back end as in [5] which we will explain in the next section. These results turn out to be better than any of the results that we obtained on the restricted test set by using likelihood ratios of either form to make verification decisions (although the performance gap is not very wide). The notation $5-1-$ in lines 3 and 4 indicates that 5 alignment iterations were used to create phrase-dependent background models and one alignment iteration was used in extracting the feature vector at enrollment and test time, as in [5] (so that we used Vogt's algorithm rather than Zhao and Dong's to obtain point estimates of the $\boldsymbol{z}$-vector features).

Table 5: *Restricted test set, 512 component UBM, phrase-dependent background model*

	a-b-c	EER	2008 NDCF
1	5-1-1	1.5%	0.062
2	5-5-5	2.0%	0.083
3	5-1-	1.3%	0.056
4	5-1-	1.4%	0.055

4.5. Results on the full evaluation set

We now report results on the full RSR2015 evaluation set using phrase-dependent background modeling and JFA as a feature extractor (Zhao and Dong's algorithm as well as Vogt's), cosine distance scoring, two versions of PLDA and a GMM/UBM benchmark. In using the $\boldsymbol{z}$-vectors as features we did not find that any type of pre-processing was helpful. The results in lines 4 and 3 of Table 5 were obtained with and without 50 dimensional nuisance attribute projection (NAP); they led us to conclude that NAP is not needed apparently because JFA succeeds well in removing session effects from the $\boldsymbol{z}$-vectors.

As an alternative to cosine distance scoring, we implemented a PLDA classifier with diagonal matrices as in [5]. Because the $\boldsymbol{z}$ features are supervector-sized, estimating a full rather than a covariance residual covariance matrix is not feasible but we attempted to compensate for this by incorporating channel factors (50 in our implementation) as in [16].

In Table 6, all of the results except those in line 3 were obtained by extracting $\boldsymbol{z}$ vectors with Vogt's algorithm rather than Zhao and Dong's. Line 1 refers to cosine distance scoring without score normalization, line 2 (the best result on the female trials) to cosine distance scoring with s-norm and similarly for line 3, line 4 refers to a PLDA classifier without channel factors or score normalization, line 5 to a PLDA classifier with 50 channel factors but with score normalization, line 6 to the PLDA classifier without channel factors but with s-norm (the best result on the male trials, but not significantly better than line 2), and line 7 to a standard GMM/UBM implementation with t-norm.

As we found with NAP, neither version of PLDA performs appreciably better than simple cosine distance scoring. Thus there appears to be no benefit from modeling session effects in the back end.

Table 6: *Full RSR2015 evaluation set, female trials (columns 1 and 2), male trials (columns 3 and 4)*

	EER	2008 NDCF	EER	2008 NDCF
1	0.92%	0.045	0.61%	0.038
2	**0.61%**	**0.027**	0.44%	0.028
3	0.93%	0.042	0.79%	0.042
4	1.12%	0.050	0.54%	0.033
5	1.11%	0.050	0.54%	0.033
6	0.87%	0.041	**0.37%**	**0.024**
7	1.06%	0.045	0.60%	0.034

5. Conclusion

We have conducted a comprehensive investigation of the question of UBM adaptation in JFA-based text-dependent speaker verification with Bayesian model selection and predictive likelihood ratios using the RSR2015 dataset as a testbed. We found that, with 64 Gaussians and the predictive likelihood ratio, using phrase-dependent background models leads to major improvements and best results are obtained with multiple iterations of relevance MAP; further minor improvements can be obtained from doing more than one alignment iteration at enrollment time or test time; and small improvements can also be obtained by using maximum likelihood II rather than relevance MAP to estimate the matrices $\boldsymbol{D}_c$. Although the Bayesian model selection likelihood ratio with UBM adaptation has a better theoretical motivation, the predictive likelihood ratio works better in practice because it avoids adapting to the speaker effects in test utterances.

This type of adaptation is treacherous in the case of a *single* test utterance because the factorial prior on $\boldsymbol{z}$ is so weak: contrary to a subspace prior, the mean vectors for the UBM mixture components adapt to the data in *statistically independent* ways so, if multiple alignment iterations are performed, the constraints on the way the data aligns with the adapted mixture components are extremely weak.

In the case of a large UBM with 512 Gaussians, even the restricted types of adaptation used in evaluating predictive likelihood ratios are unhelpful. It seems likely that the reason for this is that the weakness of the factorial prior is exacerbated by an eightfold increase in the number of mixture components. In the case of a 2 second utterance, the occupation count for a mixture component is less than 1 on average. Fortunately, it turns out that Baum-Welch statistics collected without UBM adaptation provide a very good representation of such utterances so that UBM adaptation (other than to the lexical content of utterances) is not needed.

We obtained our best results by using a JFA model built with the 512 component UBM as a feature extractor together with a simple cosine distance based back end. We used phrase-dependent mean supervectors in the JFA model and extracted the feature vectors with Vogt's algorithm rather than Zhao and Dong's. We obtained no benefit from modeling session effects in the back end, leading us to conclude that the JFA feature extractor does a very good job of suppressing such effects (at least on the RSR2015 testbed). An appealing aspect of this result is that labeled training is not required to train the back end classifier.

6. References

[1] H. Aronowitz and O. Barkan, “On leveraging conversational data for building a text dependent speaker verification system,” *Interspeech* 2013.

[2] A. Larcher, K.-A. Lee, B. Ma, and H. Li, “Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances,” *ICASSP* 2013.

[3] T. Stafylakis, P. Kenny, *et al.*, “I-Vector/PLDA

Variants for Text-Dependent Speaker Recognition," http://www.crim.ca/perso/patrick.kenny

[4] T. Stafylakis, P. Kenny, *et al.*, "Text-dependent speaker recognition using PLDA with uncertainty propagation," *Interspeech* 2013.

[5] P. Kenny, T. Stafylakis, P. Ouellet, and M. J. Alam, "JFA-based front ends for speaker recognition," *ICASSP* 2014.

[6] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using Vector Taylor Series for speaker recognition," in *ICASSP* 2013.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. SAP*, May 2005.

[8] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," *Eurospeech* 2003.

[9] L. Burget, P. Matejka, *et al.*, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Trans. ASLP* , Sept. 2007.

[10] D. Povey, L. Burget, M. Agarwal, *et al.*, "The subspace Gaussian mixture model – a structured model for speech recognition," *Computer Speech and Language*, 2011.

[11] P. Kenny, P. Ouellet, *et al.* "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, July 2008.

[12] X. Zhao and Y. Dong, "Variational Bayesian Joint Factor Analysis Models for Speaker Verification," *IEEE Trans. ASLP*, Mar. 2012.

[13] R. J. Vogt and S. Sridharan, "Explicit modeling of session variability for speaker verification," *Computer Speech and Language*, 2008.

[14] P. Kenny, G. Boulianne, *et al.* "Joint Factor Analysis versus eigenchannels in speaker recognition," *IEEE Trans. ASLP*, May 2007.

[15] T. Hasan and J. H. L. Hansen, "Acoustic factor analysis for robust speaker verification." *IEEE Trans. ASLP*, 2013.

[16] Y. Jiang, K. A. Lee, *et al.* "PLDA modeling in i-vector and supervector space for speaker verification," *Interspeech* 2012.

[17] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms, Tech. Report CRIM-06/08-13," 2005. http://www.crim.ca/perso/patrick.kenny

Appendix: Training a JFA model with UBM adaptation

Contrary to simpler models such as an i-vector extractor or Gaussian PLDA, posterior distributions of the hidden variables in JFA with UBM adaptation cannot be calculated exactly so we use Zhao and Dong's variational Bayes algorithm to calculate approximate posteriors. We have already seen how to use variational lower bounds as a proxy for the model evidence; we now explain how to use them as a criterion for hyperparameter training with a variational Bayes EM algorithm in which VB updates alternate with updates of the hyperparameters.

We assume that the UBM means and precision matrices are given. We have to estimate the matrices $\boldsymbol{U}_c$, $\boldsymbol{V}_c$ and $\boldsymbol{D}_c$, or equivalently, $\tilde{\boldsymbol{U}}_c$, $\tilde{\boldsymbol{V}}_c$ and $\tilde{\boldsymbol{D}}_c$, where $\tilde{\boldsymbol{D}}_c$ may be subject to diagonal constraints.

We assume that we have at our disposal recordings of multiple training speakers indexed by s. For each speaker s, the recordings are indexed by $r = 1, \ldots, R(s)$, the hidden variables by $\boldsymbol{X}^r(s)$ and $\boldsymbol{c}^r(s)$ and so forth. The evidence criterion is $\sum_s \mathcal{L}(s)$. Writing this in the form (6), only the first term depends on the hyperparameters $\tilde{\boldsymbol{W}}_c$. Ignoring additive constants, we can write it as

$$-\frac{1}{2}\sum_s \sum_{r=1}^{R(s)} \sum_{t=1}^{T^r(s)} \sum_{c=1} Q(c_t^r(s) = c)\langle \boldsymbol{\epsilon}_{tc}^{r*}(s)\boldsymbol{\Lambda}_c\boldsymbol{\epsilon}_{tc}^{r}(s)\rangle$$

which is the auxiliary function for maximum likelihood estimation.

Setting to zero the derivative of the auxiliary function with respect to $\tilde{\boldsymbol{W}}_c$ gives

$$\begin{aligned} \tilde{\boldsymbol{W}}_c \sum_s \sum_{r=1}^{R(s)} N_c^r(s) \, \langle \boldsymbol{X}_c^r(s)\boldsymbol{X}_c^{r*}(s)\rangle & \\ = \sum_s \sum_{r=1}^{R(s)} \tilde{\boldsymbol{F}}_c^r(s) \, \langle \boldsymbol{X}_c^{r*}(s)\rangle \,. & \qquad (10) \end{aligned}$$

This is the update formula for $\tilde{\boldsymbol{W}}_c$; it is formally identical to the update formula used in training an i-vector extractor [7]. (Slight modifications are needed if the matrices $\tilde{\boldsymbol{D}}_c$ are constrained to be diagonal rather than full or if they are fixed by relevance MAP. See for example equation (25) in [17].) In order to ensure that the evidence criterion increases monotonically from one training iteration to the next, variational posterior calculations performed after updating the model parameters have to be initialized with the variational posterior distributions used to evaluate the expressions in (10). (This complication does not arise in training a standard i-vector extractor where exact posteriors can be calculated on each training iteration and no initialization is required.)

The contribution of the negative divergences to the evidence criterion can be minimized by replacing the standard normal priors by non-standard normal priors of the form

$$\begin{aligned} P'(\boldsymbol{x}^r) &= N(\boldsymbol{x}^r|\boldsymbol{0}, \boldsymbol{A}) \\ P'(\boldsymbol{y}) &= N(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{B}) \\ P'(\boldsymbol{z}_c) &= N(\boldsymbol{z}_c|\boldsymbol{0}, \boldsymbol{C}_c) \qquad (11) \end{aligned}$$

where $N(\cdot|\cdot)$ is the Gaussian kernel and

$$\begin{aligned} \boldsymbol{A} &= \frac{1}{R}\sum_s \sum_{r=1}^{R(s)} \langle \boldsymbol{x}^r(s)\boldsymbol{x}^{r*}(s)\rangle \\ \boldsymbol{B} &= \frac{1}{S}\sum_s \langle \boldsymbol{y}(s)\boldsymbol{y}^*(s)\rangle \\ \boldsymbol{C}_c &= \frac{1}{S}\sum_s \langle \boldsymbol{z}_c(s)\boldsymbol{z}_c^*(s)\rangle. \end{aligned}$$

Here S is the number of training speakers and $R = \sum_s R(s)$. The priors can be brought to standard form in the usual way by modifying the matrix hyperparameters in such a way as to preserve the value of the evidence criterion. Let $\boldsymbol{T}_c$ be the upper triangular matrix such that $\boldsymbol{T}_c^*\boldsymbol{T}_c$ is the Cholesky decomposition of the block diagonal matrix whose diagonal blocks are $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}_c$. Then the model with non-standard priors can be brought to standard form by making the substitutions $\tilde{\boldsymbol{W}}_c \leftarrow \tilde{\boldsymbol{W}}_c\boldsymbol{T}_c^*$ and replacing the posterior expectation and covariance of $\boldsymbol{X}_c^r(s)$ by

$$\begin{aligned} & \boldsymbol{T}_c^{-*}\langle \boldsymbol{X}_c^r(s)\rangle \\ \text{and} \quad & \boldsymbol{T}_c^{-*}\,\mathrm{Cov}\left(\boldsymbol{X}_c^r(s), \boldsymbol{X}_c^r(s)\right)\boldsymbol{T}_c^{-1}. \qquad (12) \end{aligned}$$