

# Bayesian Analysis of Speaker Diarization with Eigenvoice Priors

Patrick Kenny

Centre de recherche informatique de Montréal

Patrick.Kenny@crim.ca

*A year in the lab can save you a day in the library.*

— Panu Somervuo

## I. INTRODUCTION

The speaker diarization problem consists in determining how many speakers there are in a given speech file and in partitioning the speech file into intervals each of which is assigned to one of the speakers. The collection of all intervals assigned to a given speaker is known as a *cluster*. We assume that the given speech file has already been partitioned into *segments*, that is, intervals each containing the speech of a single speaker. These segments may be of very short duration and the possibility that the same speaker is talking in two successive segments is not excluded. The problem then is how to cluster the segments so that there is a 1–1 correspondence between speakers and clusters.

In his thesis, Valente showed how the speaker clustering problem could be formulated in a principled way in terms of Bayesian model selection [1]. The primary problem, namely determining the number of speakers in the given speech file, can be viewed as one of determining the number of components in a mixture distribution where each mixture component is a speaker model; a Bayesian approach formulates this problem more generally, as one of calculating a posterior probability distribution over the number of mixture components. Similarly a Bayesian approach to the question of which segments should be assigned to which speakers results in a posterior probability distribution on all possible assignments.<sup>1</sup> Given these posterior distributions, it is a straightforward matter to make hard decisions as to the actual number of speakers and as to which of these speakers is talking in each segment.

Valente showed that putting the speaker diarization problem on a firm mathematical foundation requires just two ingredients:

<sup>1</sup>More precisely, a joint posterior distribution over both the number of speakers and all possible assignments is calculated.

- 1) Prior distributions on the number of speakers and on the mixing coefficients, where for each speaker, the corresponding mixing coefficient is the *a priori* probability that the speaker is talking in a given segment.
- 2) A prior distribution on the parameters which specify a speaker model.

In principle, a consistent application of the rules of probability (marginalization and conditioning) using these prior distributions is all that is required to calculate the posterior distributions referred to above and hence to produce a solution to the speaker diarization problem. In particular, no tunable fudge factors ought to be needed to determine the number of speakers in the given speech file. The reason for this is that, if the prior distributions are well chosen, then the Bayesian approach is automatically regularized. That is, it is (or it ought to be) immune to the overfitting tendency which maximum likelihood methods are prone to. The tendency of the Bayesian approach to prefer simple models to complex ones is sometimes referred to as a quantitative version of Occam's razor. See [2] for an excellent discussion of this issue.

Thus the key question is how to choose the prior distributions. On the one hand, Occam's razor won't work properly unless the priors are sufficiently realistic; on the other, the Bayesian integrals that need to be evaluated for the posterior calculations have to be approximated by variational methods in practice, and this restricts the choice of prior distributions to a small number of conjugate families.

There is general agreement in both the speaker diarization and text-independent speaker recognition communities that the most effective type of generative model for distinguishing between speakers is a Gaussian mixture model (GMM) derived from a universal background model (UBM) by adapting the Gaussian mean vectors but not the covariance matrices or mixture weights. We adopt the same premise in our approach to speaker diarization, so that the principal question that we need to address is how to specify an appropriate prior on the mean vectors in a speaker GMM. (Unlike Valente we do not propose to make the GMM covariance matrices speaker dependent but our speaker models will have much larger numbers of Gaussians.)

It is convenient to use the supervector formulation: a speaker *supervector* is the high dimensional vector obtained by concatenating all of the mean vectors in a speaker GMM. By far the most popular prior on speaker supervectors is the one which is used in classical *maximum a posteriori* (MAP) estimation (relevance MAP is a special case [3].) This prior has a hidden variable description of the form

$$s = m + Dz.$$

Here  $s$  is a randomly chosen speaker supervector,  $m$  is a speaker-independent supervector,  $D$  is a supervector-sized diagonal matrix and  $z$  is a random vector having a standard normal distribution. Like

the priors used by Valente, this is an example of a *factorial* prior: taking  $\mathbf{D}$  to be diagonal implies that all of the components of  $\mathbf{s}$  are statistically independent so the prior admits a component-wise factorization. Factorial priors are relatively non-informative since they fail to capture any correlations which may exist between the different components.

If they are implemented on a sufficiently large scale, then eigenvoice priors are much more effective in text-independent speaker recognition than the factorial prior that we have just described [4], [5]. The assumption here is that a randomly chosen speaker supervector  $\mathbf{s}$  is distributed according to

$$\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} \quad (1)$$

where  $\mathbf{m}$  is a speaker-independent supervector,  $\mathbf{V}$  is a rectangular matrix of low rank and  $\mathbf{y}$  is a normally distributed random vector. The columns of  $\mathbf{V}$  are eigenvoices and the components of  $\mathbf{y}$  are speaker factors. Note that there is no loss in generality in (1) in assuming that  $\mathbf{y}$  has a standard normal distribution. (Any non-standard normal distribution could be accommodated by modifying  $\mathbf{m}$  and  $\mathbf{V}$ .) In situations where  $\mathbf{y}$  has a standard normal distribution we will say that the distribution (1) is in *canonical form*.

Equation (1) imposes severe constraints on speaker supervectors. Although supervectors typically have tens of thousands of dimensions, equation (1) constrains all supervectors to lie in an affine subspace of the supervector space whose dimension is typically at most a few hundred. (The subspace in question is the affine subspace containing  $\mathbf{m}$  which is spanned by the columns of  $\mathbf{V}$ .) In Bayesian terms, (1) would thus be referred to as a highly informative prior distribution. Our main contribution in this paper is to use this type of prior together with Bayesian methods to do speaker diarization. Earlier (non-Bayesian) work in this direction can be found in [6].

The second contribution that we propose to make is to simplify the determination of the number of speakers in a given speech file by using the maximum likelihood II principle to estimate the mixing coefficients as in [7] rather than by using fully Bayesian methods. (The maximum likelihood II approach to hyperparameter estimation is discussed in a general setting in [8].) This is a reasonable procedure for the problem at hand since, no matter what (reasonable) prior is chosen for the mixing coefficients in a fully Bayesian approach, the posterior distribution of the number of speakers ought to be very sharply peaked — there should be little doubt as to how many speakers there actually are in a given speech file. Thus there should be no need for a prior on the mixing coefficients (such as a Dirichlet prior) to steer the posterior towards a reasonable number of speakers. On the other hand, a fully Bayesian treatment may be needed to discover the number of mixture components in other mixture distributions. For example, there does not seem to be a clear cut answer to the question of how many mixture components should

be used in constructing a universal background model [9].

The maximum likelihood II approach is easier to implement than a fully Bayesian treatment and it is less computationally complex. The fully Bayesian approach requires one training run for each number of speakers hypothesized, something that is likely to be quite impractical in real-world speaker diarization. A single training run suffices for the maximum likelihood II approach and, since the mixing coefficients are treated as constants rather than hidden variables, the variational posterior calculations are simpler than in the fully Bayesian approach.

## II. INTRODUCTION TO VARIATIONAL BAYES

Suppose we are given data  $\mathbf{X}$  generated by a model comprising two hidden variables  $\mathbf{Y}$  and  $\mathbf{I}$ . Set  $\theta = (\mathbf{Y}, \mathbf{I})$ . We are interested in calculating the marginal likelihood  $P(\mathbf{X})$  (or ‘evidence’) defined by

$$P(\mathbf{X}) = \int P(\mathbf{X}|\theta)P(\theta)d\theta$$

and the posterior distribution  $P(\theta|\mathbf{X})$ .

That these two problems are intimately related can be seen from the following identity which holds for any distribution  $Q(\theta)$ :

$$\ln P(\mathbf{X}) = \mathcal{L}(Q) + D(Q(\theta)||P(\theta|\mathbf{X}))$$

where  $\mathcal{L}(Q)$  is an EM-type auxiliary function and  $D(\cdot||\cdot)$  denotes the Kullback-Leibler divergence:

$$\begin{aligned} \mathcal{L}(Q) &= \int Q(\theta) \ln \frac{P(\mathbf{X}, \theta)}{Q(\theta)} d\theta \\ D(Q(\theta)||P(\theta|\mathbf{X})) &= - \int Q(\theta) \ln \frac{P(\theta|\mathbf{X})}{Q(\theta)} d\theta. \end{aligned}$$

Recall that  $D(Q(\theta)||P(\theta|\mathbf{X})) \geq 0$  with equality holding iff  $P(\theta|\mathbf{X}) = Q(\theta)$ . Thus if the true posterior  $P(\theta|\mathbf{X})$  is tractable, setting  $Q(\theta) = P(\theta|\mathbf{X})$  and evaluating  $\mathcal{L}(Q)$  gives the evidence. Any other distribution  $Q(\theta)$  gives a lower bound on the evidence. The quantity  $-\mathcal{L}$  is known to physicists as the variational free energy [2].

We are concerned with the the situation where the true posterior is intractable and we want to find a tractable approximation and, simultaneously, a lower bound on the evidence. Variational Bayes is an solution to this type of problem which iteratively refines the approximate posterior in such a way as to increase the value of the lower bound on successive iterations. In practice, the mechanics are very similar to EM.

The basic assumption is that the approximate posterior factorizes as

$$Q(\mathbf{Y}, \mathbf{I}) = Q(\mathbf{Y})Q(\mathbf{I}).$$

As in the calculus of variations, no functional form for  $Q(\mathbf{Y})$  and  $Q(\mathbf{I})$  needs to be assumed. Note that it is typically the case that the hidden variables are statistically independent in the prior, so that

$$P(\mathbf{Y}, \mathbf{I}) = P(\mathbf{Y})P(\mathbf{I}),$$

but not in the posterior, so that

$$P(\mathbf{Y}, \mathbf{I}|\mathbf{X}) \neq P(\mathbf{Y}|\mathbf{X})P(\mathbf{I}|\mathbf{X}),$$

unless it happens that  $P(\mathbf{X}|\mathbf{Y}, \mathbf{I})$  also factorizes (which is almost never the case).

To work out the update formulas for  $Q(\mathbf{Y})$  and  $Q(\mathbf{I})$ , let us fix  $Q(\mathbf{Y})$  and regard  $\mathcal{L}$  as a functional of  $Q(\mathbf{I})$  alone. Thus

$$\begin{aligned} \mathcal{L} &= \int Q(\mathbf{I}) \left( \int Q(\mathbf{Y}) \ln \frac{P(\mathbf{X}, \mathbf{Y}, \mathbf{I})}{Q(\mathbf{Y})Q(\mathbf{I})} d\mathbf{Y} \right) d\mathbf{I} \\ &= \int Q(\mathbf{I}) \left( \int Q(\mathbf{Y}) \ln \frac{P(\mathbf{X}, \mathbf{Y}, \mathbf{I})}{Q(\mathbf{Y})} d\mathbf{Y} - \ln Q(\mathbf{I}) \right) d\mathbf{I} \\ &= \int Q(\mathbf{I}) \left( \ln \tilde{Q}(\mathbf{I}) - \ln Q(\mathbf{I}) \right) d\mathbf{I} \\ \text{where } \ln \tilde{Q}(\mathbf{I}) &= \int Q(\mathbf{Y}) \ln \frac{P(\mathbf{X}, \mathbf{Y}, \mathbf{I})}{Q(\mathbf{Y})} d\mathbf{Y} \\ &= E_{\mathbf{Y}} \left[ \ln \frac{P(\mathbf{X}, \mathbf{Y}, \mathbf{I})}{Q(\mathbf{Y})} \right] \text{ for short.} \end{aligned}$$

But for the fact that  $\tilde{Q}(\mathbf{I})$  doesn't integrate to one (so it doesn't qualify as a probability distribution), the formula that we have derived for  $\mathcal{L}$  is just the negative Kullback-Leibler divergence  $-D(Q(\mathbf{I})\|\tilde{Q}(\mathbf{I}))$ . Thus the maximum value of  $\mathcal{L}$  with respect to  $Q(\mathbf{I})$  is attained when  $Q(\mathbf{I})$  and  $\tilde{Q}(\mathbf{I})$  coincide. This gives the update formula for  $Q(\mathbf{I})$ , namely

$$\begin{aligned} \ln Q(\mathbf{I}) &= E_{\mathbf{Y}} \left[ \ln \frac{P(\mathbf{X}, \mathbf{Y}, \mathbf{I})}{Q(\mathbf{Y})} \right] + \text{const} \\ &= E_{\mathbf{Y}} [\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I})] + \text{const} \end{aligned}$$

where const is independent of  $\mathbf{I}$  (its value is determined by the condition that  $Q(\mathbf{I})$  integrates to 1). Similarly the update formula for  $Q(\mathbf{Y})$  is

$$\ln Q(\mathbf{Y}) = E_{\mathbf{I}} [\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I})] + \text{const}.$$

Because these formulas are coupled they have to be applied iteratively. Convergence is guaranteed [10]. Note that the update formulas are designed to ensure that the value of  $\mathcal{L}$  increases on each update.

### III. MODELING ASSUMPTIONS

We suppose that we are given a segmented speech file and that we wish to determine how many speakers there are in the file and which segments are associated with each of these speakers. We use the term segment to refer to a sequence of speech observation vectors which corresponds to a time interval in which a single speaker is talking. Although it could be incorporated into the type of framework that we will develop, we exclude the possibility that more than one speaker is talking at once. We denote acoustic feature vectors by  $X_1, X_2, \dots$  and we let  $F$  denote the dimension of these vectors.

**Assumption 1:** *Segment boundaries are given.*

A uniform segmentation (say into 1 second intervals) can be assumed to begin with; this can be refined iteratively once point estimates of speaker models have been calculated. Given a speech file divided into  $M$  segments  $\mathbf{x}_1, \dots, \mathbf{x}_M$ , set

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_M).$$

(The m in segments is intended to serve as a mnemonic: segments are indexed by  $m = 1, \dots, M$ .) We assume provisionally that the number of speakers  $S$  in the speech file is known.

For each  $m = 1, \dots, M$ , we associate with the segment  $\mathbf{x}_m$  an  $S \times 1$  indicator vector  $\mathbf{i}_m$  whose components are defined as follows: for  $s = 1, \dots, S$ ,  $i_{ms} = 1$  if speaker  $s$  is talking in the segment and  $i_{ms} = 0$  otherwise. For  $s = 1, \dots, S$ , set  $P(i_{ms} = 1) = \pi_s$  so that

$$P(\mathbf{i}_m) = \prod_{s=1}^S \pi_s^{i_{ms}}.$$

Again, we assume provisionally that the mixing coefficients  $\pi_1, \dots, \pi_S$  are known. Set

$$\mathbf{I} = (\mathbf{i}_1, \dots, \mathbf{i}_M)$$

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_S).$$

( $\mathbf{I}$  is not to be confused with the identity matrix.)

The likelihood function is given by

$$P(\mathbf{X}, \mathbf{Y}, \mathbf{I} | \boldsymbol{\pi}) = P(\mathbf{X} | \mathbf{Y}, \mathbf{I}) P(\mathbf{Y}) P(\mathbf{I} | \boldsymbol{\pi})$$

where

$$\begin{aligned}
 P(\mathbf{X}|\mathbf{Y}, \mathbf{I}) &= \prod_{m=1}^M \prod_{s=1}^S P(\mathbf{x}_m|\mathbf{y}_s)^{i_{ms}} \\
 P(\mathbf{Y}) &= \prod_{s=1}^S P(\mathbf{y}_s) \\
 P(\mathbf{I}|\boldsymbol{\pi}) &= \prod_{m=1}^M \prod_{s=1}^S \pi_s^{i_{ms}}.
 \end{aligned}$$

Here  $\mathbf{y}_s$  is vector of parameters which specify the model for speaker  $s$  and the distributions  $P(\mathbf{x}_m|\mathbf{y}_s)$  and  $P(\mathbf{y}_s)$  remain to be specified.

The assignment of segments to speakers is encoded in the posterior distribution  $P(\mathbf{I}|\mathbf{X}, \boldsymbol{\pi})$  which, up to a normalizing constant (namely  $1/P(\mathbf{X}|\boldsymbol{\pi})$ ), is given by marginalizing  $P(\mathbf{X}, \mathbf{Y}, \mathbf{I}|\boldsymbol{\pi})$ :

$$\begin{aligned}
 P(\mathbf{I}|\mathbf{X}, \boldsymbol{\pi}) &\propto P(\mathbf{X}, \mathbf{I}|\boldsymbol{\pi}) \\
 &= \int P(\mathbf{X}, \mathbf{Y}, \mathbf{I}|\boldsymbol{\pi}) d\mathbf{Y}.
 \end{aligned}$$

The calculation of the posterior  $P(\mathbf{I}|\mathbf{X}, \boldsymbol{\pi})$  is intractable in practice so we will approximate it by a variational posterior  $Q(\mathbf{I})$ . This will have the property that it factorizes in the same way as  $P(\mathbf{I})$ . Thus

$$Q(\mathbf{I}) = \prod_{m=1}^M \prod_{s=1}^S q_{ms}^{i_{ms}}$$

where  $q_{ms}$  is the posterior probability that speaker  $s$  is talking in segment  $m$ . Thus to make a hard decision concerning the identity of the speaker in segment  $m$  we find

$$\operatorname{argmax}_s q_{ms}.$$

**Assumption 2:** *An upper bound  $S$  on the number of speakers is given. The mixing coefficients  $(\pi_1, \dots, \pi_S)$  can be estimated by maximizing the marginal likelihood  $P(\mathbf{X}|\boldsymbol{\pi})$  with respect to  $\boldsymbol{\pi}$  (that is, by maximum likelihood II).*

The marginal likelihood of the data  $\mathbf{X}$  is given by

$$P(\mathbf{X}|\boldsymbol{\pi}) = \int P(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\pi}) d\boldsymbol{\theta} \text{ where } \boldsymbol{\theta} = (\mathbf{Y}, \mathbf{I}).$$

(The marginal likelihood of the data is sometimes referred to as the evidence.) Thus we propose to discover the actual number of speakers in the given speech file by counting the number of mixture coefficients assigned non-zero values by maximum likelihood II estimation. Since the integral is intractable we will use a variational approximation to evaluate it.

**Assumption 3:** *Speaker supervectors are distributed as follows: if  $\mathbf{s}$  is a randomly chosen supervector,*

$$\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} \quad (2)$$

where  $\mathbf{V}$  is of dimension  $CF \times R$  and the  $R \times 1$  random vector  $\mathbf{y}$  has a standard normal distribution.

Here  $C$  denotes the number of components in the UBM or a speaker GMM so that supervectors are of dimension  $CF \times 1$ ;  $R$  is the number of eigenvoices. For each mixture component  $c = 1, \dots, C$ , let  $m_c$  be the  $F \times 1$  subvector of  $\mathbf{m}$  which corresponds to the mixture component and let  $\Sigma_c$  be the corresponding covariance matrix. We assume that  $\Sigma_c$  is diagonal and we denote by  $\Sigma$  the supervector sized diagonal covariance matrix obtained by concatenating  $\Sigma_1, \Sigma_2, \dots, \Sigma_C$ .

An algorithm for estimating  $\mathbf{V}$  and  $\Sigma$  using the maximum likelihood II criterion is presented in [5]. In implementing this algorithm different recordings of a given speaker should be treated as different ‘speakers’. This is appropriate for the problem at hand since, in a telephone conversation involving different speakers, each speaker will be subject to different channel effects. (For some applications, additional eigenchannels will need to be appended to these eigenvoices.)

**Assumption 4:** *The alignment of frames with mixture components is given.*

This is perhaps the most questionable assumption that we make; it is motivated by the success we have had with it in text-independent speaker-recognition [4]. The idea is to use Viterbi or Baum-Welch alignments calculated with a UBM to simplify calculations with speaker dependent GMM’s. Given acoustic feature vectors  $X_1, X_2, \dots$  corresponding to a segment  $\mathbf{x}$ , for each mixture component  $c$  we define the first and second order Viterbi statistics in the usual way:

$$F_c = \sum_t X_t$$

$$S_c = \text{diag} \left( \sum_t X_t X_t^* \right)$$

where the sums extend over all frames  $t$  which are aligned with the mixture component  $c$ . Let  $N_c$  be the number of such frames. Viterbi alignments are carried out using the UBM.

Although we will refer to these statistics as Viterbi statistics in this paper, in our implementation we actually use Baum-Welch statistics instead. These are defined by

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

$$F_c = \sum_t \gamma_t(c) X_t$$

$$S_c = \text{diag} \left( \sum_t \gamma_t(c) X_t X_t^* \right)$$

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

We denote the centralized first- and second order Viterbi statistics by  $\tilde{F}_c$  and  $\tilde{S}_c$ :

$$\begin{aligned}\tilde{F}_c &= \sum_t (X_t - m_c) \\ \tilde{S}_c &= \text{diag} \left( \sum_t (X_t - m_c)(X_t - m_c)^* \right)\end{aligned}$$

where  $m_c$  is the subvector of  $\mathbf{m}$  in (1) which corresponds to the mixture component  $c$ . In other words,

$$\tilde{F}_c = F_c - N_c m_c \quad (3)$$

$$\tilde{S}_c = S_c - \text{diag} (F_c m_c^* + m_c F_c^* - N_c m_c m_c^*). \quad (4)$$

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

Given a segment  $\mathbf{x}$  and speaker factors  $\mathbf{y}$ , the conditional likelihood  $P(\mathbf{x}|\mathbf{y})$  is calculated as follows. Let  $\mathbf{N}$ ,  $\tilde{\mathbf{F}}$  and  $\tilde{\mathbf{S}}$  be the centralized Viterbi statistics extracted from the segment  $\mathbf{x}$  and define

$$G = \sum_{c=1}^C N_c \ln \frac{1}{(2\pi)^{F/2} |\Sigma_c|^{1/2}} - \frac{1}{2} \text{tr} (\Sigma^{-1} \tilde{\mathbf{S}}). \quad (5)$$

$$H(\mathbf{y}) = \mathbf{y}^* \mathbf{V}^* \Sigma^{-1} \tilde{\mathbf{F}} - \frac{1}{2} \mathbf{y}^* \mathbf{V}^* \mathbf{N} \Sigma^{-1} \mathbf{V} \mathbf{y}. \quad (6)$$

Then, by Lemma 1 in [5],

$$\ln P(\mathbf{x}|\mathbf{y}) = G + H(\mathbf{y}). \quad (7)$$

The posterior distribution  $P(\mathbf{y}|\mathbf{x})$  can be calculated by appealing to Proposition 1 in [5]. If  $\mathbf{N}$ ,  $\tilde{\mathbf{F}}$  and  $\tilde{\mathbf{S}}$  are the centralized Viterbi statistics extracted from  $\mathbf{x}$ , then the posterior distribution of  $P(\mathbf{y}|\mathbf{x})$  is Gaussian with mean  $\mathbf{a}$  and precision matrix  $\mathbf{l}$  given by

$$\mathbf{l} = \mathbf{I} + \mathbf{V}^* \Sigma^{-1} \mathbf{N} \mathbf{V} \quad (8)$$

$$\mathbf{a} = \mathbf{l}^{-1} \mathbf{V}^* \Sigma^{-1} \tilde{\mathbf{F}}. \quad (9)$$

Calculating the marginal likelihood of the segment,  $P(\mathbf{x})$  involves evaluating the Gaussian integral

$$\int P(\mathbf{x}|\mathbf{y}) P(\mathbf{y}) d\mathbf{y} \quad \text{where } P(\mathbf{y}) = N(\mathbf{y}|\mathbf{0}, \mathbf{I}).$$

(We use  $N(\cdot|\boldsymbol{\mu}, \mathbf{K})$  to indicate the Gaussian kernel with mean  $\boldsymbol{\mu}$  and covariance matrix  $\mathbf{K}$ .) A closed form expression in terms of the Viterbi statistics extracted from  $\boldsymbol{x}$  is derived in Proposition 2 in [5], namely

$$\ln P(\boldsymbol{x}) = G - \frac{1}{2}|\boldsymbol{l}| + \frac{1}{2}\tilde{\mathbf{F}}^* \boldsymbol{\Sigma}^{-1} \mathbf{V} \boldsymbol{l}^{-1} \mathbf{V}^* \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{F}}. \quad (10)$$

If the assignment of frames to mixture components was not given then a variational approximation would be needed to evaluate this marginal likelihood.

#### IV. AN EXAMPLE: SPEAKER CHANGE DETECTION

Since we are assuming that the speech file has been presegmented (Assumption 1) we will not propose a solution to the problem of detecting speaker change points. However we will show how a simple problem of this type can be solved in closed form thanks to Assumption 4. This is an illustration of the Bayesian Occam's razor principle.

Suppose we are given a sequence of frames  $X_1, \dots, X_T$  and a time  $T_1$  where  $1 < T_1 < T$  and it is required to test the hypothesis  $H_1$  that a single speaker is talking in the interval against the hypothesis  $H_2$  that one speaker is talking in the interval  $X_1, \dots, X_{T_1}$  and another speaker in the interval  $X_{T_1+1}, \dots, X_T$ . Denote these intervals by  $\boldsymbol{x}$ ,  $\boldsymbol{x}_1$  and  $\boldsymbol{x}_2$ .

We assume that the two hypotheses are *a priori* equally likely. Under the hypothesis  $H_2$ , the intervals  $\boldsymbol{x}_1$  and  $\boldsymbol{x}_2$  are statistically independent so  $P(\boldsymbol{x}_1, \boldsymbol{x}_2|H_2) = P(\boldsymbol{x}_1)P(\boldsymbol{x}_2)$ . Thus the the likelihood ratio (or Bayes factor) for the hypothesis test is

$$\frac{P(\boldsymbol{x})}{P(\boldsymbol{x}_1)P(\boldsymbol{x}_2)}.$$

Each of the terms in this expression can be evaluated in terms of the Viterbi statistics extracted from  $\boldsymbol{x}$ ,  $\boldsymbol{x}_1$  and  $\boldsymbol{x}_2$  using (10). If this ratio is greater than 1, we conclude that  $H_1$  is more likely than  $H_2$ . Note that this decision criterion is essentially the same as the batch likelihood ratio criterion for speaker recognition introduced in [11] and the 'identity variable' criterion for face recognition used in [12].

For another perspective, note that we can write

$$\frac{P(\boldsymbol{x})}{P(\boldsymbol{x}_1)P(\boldsymbol{x}_2)} = \frac{P(\boldsymbol{x}_2|\boldsymbol{x}_1)}{P(\boldsymbol{x}_2)} \quad (11)$$

where  $P(\boldsymbol{x}_2)$  is evaluated as before and

$$P(\boldsymbol{x}_2|\boldsymbol{x}_1) = \int P(\boldsymbol{x}_2|\boldsymbol{y})P(\boldsymbol{y}|\boldsymbol{x}_1) d\boldsymbol{y}. \quad (12)$$

Recall that  $P(\mathbf{y}|\mathbf{x}_1)$  is a non-standard Gaussian kernel whose mean  $\mathbf{a}$  and precision matrix  $\mathbf{l}$  are given in terms of the Viterbi statistics of  $\mathbf{x}_1$  by (8) and (9).

Setting  $\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y}$ , the posterior distribution of  $\mathbf{s}$  conditioned on  $\mathbf{x}_1$  can be put into canonical form by writing  $\mathbf{l}^{-1}$  in the form  $\mathbf{t}^*\mathbf{t}$  (Cholesky decomposition), and setting

$$\begin{aligned}\mathbf{m}_1 &= \mathbf{m} + \mathbf{V}\mathbf{a} \\ \mathbf{V}_1 &= \mathbf{V}\mathbf{t}^*\end{aligned}$$

so that

$$\mathbf{s} = \mathbf{m}_1 + \mathbf{V}_1\mathbf{y}_1 \quad (13)$$

where  $\mathbf{y}_1$  has a standard normal distribution.<sup>2</sup> This enables us to evaluate the integral in (12) using (10) by replacing  $\mathbf{m}$  and  $\mathbf{V}$  with  $\mathbf{m}_1$  and  $\mathbf{V}_1$ . (In particular the Viterbi statistics have to be centralized with respect to  $\mathbf{m}_1$  rather than  $\mathbf{m}$  in (3) and (4).)

Returning to the right hand side of (11), the question of whether the denominator or the numerator has the larger value depends on which of the distributions (2) and (13) better accounts for the segment  $\mathbf{x}_2$  and this in turn depends on how similar  $\mathbf{x}_2$  is to  $\mathbf{x}_1$ .

## V. VARIATIONAL POSTERIOR CALCULATIONS

### *Recapitulation*

We assumed that we are given a speech file divided into  $M$  segments  $\mathbf{x}_1, \dots, \mathbf{x}_M$ , and an upper bound  $S$  on the number of speakers in the file. We want to determine the actual number of speakers and the assignment of speakers to segments.

For each  $m = 1, \dots, M$ , we associated with the segment  $\mathbf{x}_m$  an  $S \times 1$  indicator vector  $\mathbf{i}_m$  whose components are defined as follows: for  $s = 1, \dots, S$ ,  $i_{ms} = 1$  if speaker  $s$  is talking in the segment and  $i_{ms} = 0$  otherwise. For  $s = 1, \dots, S$ , we set  $P(i_{ms} = 1) = \pi_s$  where the mixing coefficient  $\pi_s$  is the *a priori* probability that speaker  $s$  is speaking in a given segment. Thus

$$P(\mathbf{i}_m) = \prod_{s=1}^S \pi_s^{i_{ms}}. \quad (14)$$

<sup>2</sup>This is the minimum divergence estimation procedure of [4] (Section III-B3) applied to a single speaker.

Let  $N_m$ ,  $\tilde{\mathbf{F}}_m$  and  $\tilde{\mathbf{S}}_m$  be the centralized Viterbi statistics extracted from the segment  $\mathbf{x}_m$ . and define

$$G_m = \sum_{c=1}^C N_{mc} \ln \frac{1}{(2\pi)^{F/2} |\Sigma_c|^{1/2}} - \frac{1}{2} \text{tr} \left( \Sigma^{-1} \tilde{\mathbf{S}}_m \right). \quad (15)$$

$$H_m(\mathbf{y}) = \mathbf{y}^* \mathbf{V}^* \Sigma^{-1} \tilde{\mathbf{F}}_m - \frac{1}{2} \mathbf{y}^* \mathbf{V}^* \mathbf{N}_m \Sigma^{-1} \mathbf{V} \mathbf{y} \quad (16)$$

so that

$$\ln P(\mathbf{x}_m | \mathbf{y}_s) = G_m + H_m(\mathbf{y}_s). \quad (17)$$

We set

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_M)$$

$$\mathbf{I} = (\mathbf{i}_1, \dots, \mathbf{i}_M)$$

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_S)$$

$$\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_S).$$

so that the likelihood function is given by

$$P(\mathbf{X}, \mathbf{Y}, \mathbf{I} | \boldsymbol{\pi}) = P(\mathbf{X} | \mathbf{Y}, \mathbf{I}) P(\mathbf{Y}) P(\mathbf{I} | \boldsymbol{\pi}) \quad (18)$$

where

$$\begin{aligned} P(\mathbf{X} | \mathbf{Y}, \mathbf{I}) &= \prod_{m=1}^M \prod_{s=1}^S P(\mathbf{x}_m | \mathbf{y}_s)^{i_{ms}} \\ P(\mathbf{I} | \boldsymbol{\pi}) &= \prod_{m=1}^M \prod_{s=1}^S \pi_s^{i_{ms}} \end{aligned} \quad (19)$$

and  $P(\mathbf{Y})$  is the standard normal distribution.

We aim to estimate  $\boldsymbol{\pi}$ , and hence the actual number of speakers in the file, by maximizing the marginal likelihood  $P(\mathbf{X} | \boldsymbol{\pi})$ . We aim to assign speakers to segments by calculating the posterior distribution  $P(\mathbf{I} | \mathbf{X}, \boldsymbol{\pi})$ .

### *Variational approximation*

Neither the marginal likelihood  $P(\mathbf{X} | \boldsymbol{\pi})$  nor the posterior  $P(\mathbf{I} | \mathbf{X}, \boldsymbol{\pi})$  can be calculated exactly so we adopt a variational approach as outlined in Section II. Assuming provisionally that  $\boldsymbol{\pi}$  is known, we will approximate the true posterior  $P(\mathbf{Y}, \mathbf{I} | \mathbf{X}, \boldsymbol{\pi})$  by a factorized distribution which we will denote by  $Q(\mathbf{Y}, \mathbf{I})$ . Not only will this distribution enable us to make a probabilistic assignment of speakers to segments if  $\boldsymbol{\pi}$  is known, but it will enable us to calculate a lower bound on the marginal likelihood

$P(\mathbf{X}|\boldsymbol{\pi})$  which can serve as a criterion for estimating  $\boldsymbol{\pi}$ . The only assumption that we make concerning  $Q(\mathbf{Y}, \mathbf{I})$  is that

$$Q(\mathbf{Y}, \mathbf{I}) = Q(\mathbf{Y})Q(\mathbf{I}). \quad (20)$$

Writing the marginal likelihood in the form

$$P(\mathbf{X}|\boldsymbol{\pi}) = \int P(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\pi}) d\boldsymbol{\theta} \quad \text{where } \boldsymbol{\theta} = (\mathbf{Y}, \mathbf{I}),$$

we define the functional  $\mathcal{L}(Q|\boldsymbol{\pi})$  by

$$\mathcal{L}(Q|\boldsymbol{\pi}) = \int Q(\boldsymbol{\theta}) \ln \frac{P(\mathbf{X}, \boldsymbol{\theta}|\boldsymbol{\pi})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta}. \quad (21)$$

As explained in Section II,

$$\mathcal{L}(Q|\boldsymbol{\pi}) \leq \ln P(\mathbf{X}|\boldsymbol{\pi})$$

for all  $Q$ .

The update formulas for  $Q(\mathbf{Y})$  and  $Q(\mathbf{I})$  are

$$\ln Q(\mathbf{Y}) = E_{\mathbf{I}} [\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I}|\boldsymbol{\pi})] + \text{const} \quad (22)$$

$$\ln Q(\mathbf{I}) = E_{\mathbf{Y}} [\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I}|\boldsymbol{\pi})] + \text{const} \quad (23)$$

where,  $E_{\mathbf{I}}[\cdot]$  indicates the expectation with respect to  $\mathbf{I}$  calculated with the distribution  $Q(\mathbf{I})$  and similarly for  $E_{\mathbf{Y}}[\cdot]$  and the constants are chosen so as to ensure that the total probability mass of each distribution is 1. Since the update formulas are coupled they have to be applied alternately. Each update is guaranteed to increase the value of  $\mathcal{L}(Q|\boldsymbol{\pi})$  [10]. Thus evaluating  $\mathcal{L}(Q|\boldsymbol{\pi})$  on each update is a useful check on whether the update formulas have been correctly implemented.

We will see that  $Q(\mathbf{Y})$  and  $Q(\mathbf{I})$  both factorize in the same way as  $P(\mathbf{Y})$  and  $P(\mathbf{I})$ :

$$Q(\mathbf{I}) = \prod_{m=1}^M Q(\mathbf{i}_m) \quad (24)$$

$$\text{where } Q(\mathbf{i}_m) = \prod_{s=1}^S q_{ms}^{i_{ms}} \quad (m = 1, \dots, M) \quad (25)$$

and

$$Q(\mathbf{Y}) = \prod_{s=1}^S N(\mathbf{y}_s | \mathbf{a}_s, \boldsymbol{\Lambda}_s^{-1}). \quad (26)$$

where the means and precisions  $\mathbf{a}_s$  and  $\boldsymbol{\Lambda}_s$  and the probabilities  $q_{ms}$  remain to be specified.

### Updating $Q(\mathbf{I})$

To update  $Q(\mathbf{I})$ , we calculate  $E_{\mathbf{Y}} [\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I}|\boldsymbol{\pi})]$  as prescribed by (23). Ignoring terms independent of  $\mathbf{I}$ ,

$$\begin{aligned}
E_{\mathbf{Y}} [\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I}|\boldsymbol{\pi})] &= E_{\mathbf{Y}} [\ln P(\mathbf{X}|\mathbf{Y}, \mathbf{I}) + \ln P(\mathbf{I}|\boldsymbol{\pi})] \quad \text{by (18)} \\
&= \sum_{m=1}^M \sum_{s=1}^S i_{ms} E_{\mathbf{y}_s} [\ln P(\mathbf{x}_m|\mathbf{y}_s)] + \sum_{m=1}^M \sum_{s=1}^S i_{ms} \ln \pi_s \quad \text{by (19)} \\
&= \sum_{m=1}^M \sum_{s=1}^S i_{ms} \ln \tilde{q}_{ms} \tag{27}
\end{aligned}$$

where

$$\ln \tilde{q}_{ms} = E_{\mathbf{y}_s} [\ln P(\mathbf{x}_m|\mathbf{y}_s)] + \ln \pi_s. \tag{28}$$

To evaluate  $\tilde{q}_{ms}$ , we write

$$\begin{aligned}
E_{\mathbf{y}_s} [\ln P(\mathbf{x}_m|\mathbf{y}_s)] &= E_{\mathbf{y}_s} [G_m + H_m(\mathbf{y}_s)] \quad \text{by (17)} \\
&= G_m + E_{\mathbf{y}_s} [H_m(\mathbf{y}_s)] \\
&= G_m + E_{\mathbf{y}_s} \left[ \mathbf{y}_s^* \mathbf{V}^* \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{F}}_m - \frac{1}{2} \mathbf{y}_s^* \mathbf{V}^* \mathbf{N}_m \boldsymbol{\Sigma}^{-1} \mathbf{V} \mathbf{y}_s \right] \quad \text{by (16)} \\
&= G_m + E_{\mathbf{y}_s} \left[ \mathbf{y}_s^* \mathbf{V}^* \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{F}}_m \right] - \frac{1}{2} E_{\mathbf{y}_s} [\text{tr} (\mathbf{V}^* \mathbf{N}_m \boldsymbol{\Sigma}^{-1} \mathbf{V} \mathbf{y}_s \mathbf{y}_s^*)] \\
&= G_m + \mathbf{a}_s^* \mathbf{V}^* \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{F}}_m - \frac{1}{2} \text{tr} (\mathbf{V}^* \mathbf{N}_m \boldsymbol{\Sigma}^{-1} \mathbf{V} (\boldsymbol{\Lambda}_s^{-1} + \mathbf{a}_s \mathbf{a}_s^*))
\end{aligned}$$

where  $\mathbf{a}_s$  and  $\boldsymbol{\Lambda}_s$  are the variational posterior mean and precision of  $\mathbf{y}_s$  and we have used the fact that

$$E_{\mathbf{y}_s} [\mathbf{y}_s \mathbf{y}_s^*] = \boldsymbol{\Lambda}_s^{-1} + \mathbf{a}_s \mathbf{a}_s^*$$

(Note that  $\mathbf{a}_s$  and  $\boldsymbol{\Lambda}_s$  have yet to be calculated.) Thus

$$\ln \tilde{q}_{ms} = G_m + \mathbf{a}_s^* \mathbf{V}^* \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{F}}_m - \frac{1}{2} \text{tr} (\mathbf{V}^* \mathbf{N}_m \boldsymbol{\Sigma}^{-1} \mathbf{V} (\boldsymbol{\Lambda}_s^{-1} + \mathbf{a}_s \mathbf{a}_s^*)) + \ln \pi_s. \tag{29}$$

Normalizing so that probabilities sum to 1 gives  $Q(\mathbf{I})$ :

$$\begin{aligned}
Q(\mathbf{I}) &= \prod_{m=1}^M \prod_{s=1}^S q_{ms}^{i_{ms}} \\
\text{where } q_{ms} &= \frac{\tilde{q}_{ms}}{\sum_{s'=1}^S \tilde{q}_{ms'}}. \tag{30}
\end{aligned}$$

*Remark*

By (7), equation (29) can be rewritten in the form

$$\ln \tilde{q}_{ms} = \ln \pi_s P(\mathbf{x}_m | \mathbf{y}_s = \mathbf{a}_s) - \frac{1}{2} \text{tr}(\mathbf{V}^* \mathbf{N}_m \boldsymbol{\Sigma}^{-1} \mathbf{V} \boldsymbol{\Lambda}_s^{-1}). \quad (31)$$

The second term in this expression vanishes if there is no uncertainty in the point estimate  $\mathbf{a}_s$  of the speaker factors (that is, if the posterior covariance matrix  $\boldsymbol{\Lambda}_s^{-1}$  is zero). In this case the Variational Bayes formalism drops out and the calculation of the posterior probabilities  $q_{ms}$  reduces to a straightforward application of Bayes rule.

*Updating  $Q(\mathbf{Y})$* 

To update  $Q(\mathbf{Y})$ , we calculate  $E_{\mathbf{I}}[\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I} | \boldsymbol{\pi})]$  as prescribed by (22):

$$\begin{aligned} E_{\mathbf{I}}[\ln P(\mathbf{X}, \mathbf{Y}, \mathbf{I} | \boldsymbol{\pi})] &= E_{\mathbf{I}}[\ln P(\mathbf{X} | \mathbf{Y}, \mathbf{I})] + \ln P(\mathbf{Y}) + E_{\mathbf{I}}[\ln P(\mathbf{I})] \quad \text{by (18)} \\ &= E_{\mathbf{I}} \left[ \sum_{m=1}^M \sum_{s=1}^S i_{ms} \ln P(\mathbf{x}_m | \mathbf{y}_s) \right] + \sum_{s=1}^S \ln P(\mathbf{y}_s) \\ &\quad + E_{\mathbf{I}} \left[ \sum_{m=1}^M \sum_{s=1}^S i_{ms} \ln \pi_s \right] \quad \text{by (19)} \\ &= \sum_{m=1}^M \sum_{s=1}^S q_{ms} \ln P(\mathbf{x}_m | \mathbf{y}_s) + \sum_{s=1}^S \ln P(\mathbf{y}_s) + \sum_{m=1}^M \sum_{s=1}^S q_{ms} \ln \pi_s \end{aligned}$$

where we have used the fact that

$$E_{\mathbf{I}}[i_{ms}] = q_{ms}. \quad (32)$$

Thus

$$\begin{aligned} Q(\mathbf{Y}) &= \prod_{s=1}^S Q(\mathbf{y}_s) \\ \text{where } \ln Q(\mathbf{y}_s) &= \sum_{m=1}^M q_{ms} \ln P(\mathbf{x}_m | \mathbf{y}_s) + \ln P(\mathbf{y}_s) + \text{const} \quad (s = 1, \dots, S) \end{aligned}$$

Ignoring additive terms independent of  $\mathbf{y}_s$ ,

$$\begin{aligned}
\ln Q(\mathbf{y}_s) &= \sum_{m=1}^M q_{ms} \ln P(\mathbf{x}_m | \mathbf{y}_s) + \ln P(\mathbf{y}_s) \\
&= \sum_{m=1}^M q_{ms} H_m(\mathbf{y}_s) + \ln P(\mathbf{y}_s) \quad \text{by (17)} \\
&= \sum_{m=1}^M q_{ms} \left( \mathbf{y}_s^* \mathbf{V}^* \Sigma^{-1} \tilde{\mathbf{F}}_m - \frac{1}{2} \mathbf{y}_s^* \mathbf{V}^* \mathbf{N}_m \Sigma^{-1} \mathbf{V} \mathbf{y}_s \right) - \frac{1}{2} \mathbf{y}_s^* \mathbf{y}_s \quad \text{by (16)} \\
&= -\frac{1}{2} (\mathbf{y}_s - \mathbf{a}_s)^* \Lambda_s (\mathbf{y}_s - \mathbf{a}_s) \tag{33}
\end{aligned}$$

where

$$\Lambda_s = \mathbf{I} + \mathbf{V}^* \left( \sum_{m=1}^M q_{ms} \mathbf{N}_m \right) \Sigma^{-1} \mathbf{V} \tag{34}$$

$$\mathbf{a}_s = \Lambda_s^{-1} \mathbf{V}^* \Sigma^{-1} \left( \sum_{m=1}^M q_{ms} \tilde{\mathbf{F}}_m \right) \tag{35}$$

In other words, the variational posterior distribution of  $\mathbf{y}_s$  is Gaussian with mean  $\mathbf{a}_s$  and covariance matrix  $\Lambda_s^{-1}$ . Note the similarity with (8) and (9). Normalizing,

$$\ln Q(\mathbf{y}_s) = \ln(2\pi)^{-R/2} |\Lambda_s|^{1/2} - \frac{1}{2} (\mathbf{y}_s - \mathbf{a}_s)^* \Lambda_s (\mathbf{y}_s - \mathbf{a}_s) \tag{36}$$

*Evaluating  $\mathcal{L}(Q|\pi)$*

$$\begin{aligned}
\mathcal{L}(Q|\pi) &= \int Q(\boldsymbol{\theta}) \ln \frac{P(\mathbf{X}, \boldsymbol{\theta} | \pi)}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \quad \text{by (21)} \\
&= E_{\mathbf{I}} \left[ E_{\mathbf{Y}} \left[ \ln \frac{P(\mathbf{X}, \mathbf{Y}, \mathbf{I} | \pi)}{Q(\mathbf{Y}) Q(\mathbf{I})} \right] \right] \\
&= E_{\mathbf{I}} [E_{\mathbf{Y}} [\ln P(\mathbf{X} | \mathbf{Y}, \mathbf{I})]] - E_{\mathbf{Y}} \left[ \ln \frac{Q(\mathbf{Y})}{P(\mathbf{Y})} \right] - E_{\mathbf{I}} \left[ \ln \frac{Q(\mathbf{I})}{P(\mathbf{I} | \pi)} \right]
\end{aligned}$$

Using (32) and (27) to simplify the first and third terms gives

$$\begin{aligned}
E_{\mathbf{I}} [E_{\mathbf{Y}} [\ln P(\mathbf{X} | \mathbf{Y}, \mathbf{I})]] &= E_{\mathbf{I}} \left[ \sum_{m=1}^M \sum_{s=1}^S i_{ms} E_{\mathbf{y}_s} [\ln P(\mathbf{x}_m | \mathbf{y}_s)] \right] \\
&= \sum_{m=1}^M \sum_{s=1}^S q_{ms} E_{\mathbf{y}_s} [\ln P(\mathbf{x}_m | \mathbf{y}_s)] \\
\text{and } E_{\mathbf{I}} \left[ \ln \frac{Q(\mathbf{I})}{P(\mathbf{I})} \right] &= E_{\mathbf{I}} \left[ \sum_{m=1}^M \sum_{s=1}^S i_{ms} \ln \frac{q_{ms}}{\pi_s} \right] \\
&= \sum_{m=1}^M \sum_{s=1}^S q_{ms} \ln \frac{q_{ms}}{\pi_s}
\end{aligned}$$

For the second term:

$$\begin{aligned}
E_{\mathbf{Y}} \left[ \ln \frac{Q(\mathbf{Y})}{P(\mathbf{Y})} \right] &= \sum_{s=1}^S E_{\mathbf{y}_s} \left[ \ln \frac{Q(\mathbf{y}_s)}{P(\mathbf{y}_s)} \right] \\
&= \sum_{s=1}^S D(Q(\mathbf{y}_s) \| P(\mathbf{y}_s)) \\
&= -\frac{RS}{2} + \frac{1}{2} \sum_{s=1}^S \left( -\ln |\text{Cov}(\mathbf{y}_s, \mathbf{y}_s)| + \text{tr}(E_{\mathbf{y}_s}[\mathbf{y}_s \mathbf{y}_s^*]) \right) \\
&= -\frac{RS}{2} + \frac{1}{2} \sum_{s=1}^S \left( \ln |\mathbf{\Lambda}_s| + \text{tr}(\mathbf{\Lambda}_s^{-1} + E_{\mathbf{y}_s}[\mathbf{y}_s] E_{\mathbf{y}_s}[\mathbf{y}_s^*]) \right) \\
&= -\frac{RS}{2} + \frac{1}{2} \sum_{s=1}^S \left( \ln |\mathbf{\Lambda}_s| + \text{tr}(\mathbf{\Lambda}_s^{-1} + \mathbf{a}_s \mathbf{a}_s^*) \right)
\end{aligned}$$

by the formula for the Kullback-Leibler divergence of two multivariate Gaussians. Note that this is always positive. Collecting terms,

$$\begin{aligned}
\mathcal{L}(Q|\boldsymbol{\pi}) &= \sum_{m=1}^M \sum_{s=1}^S q_{ms} \ln \tilde{q}_{ms} \\
&\quad + \frac{1}{2} \left\{ RS - \sum_{s=1}^S \left( \ln |\mathbf{\Lambda}_s| + \text{tr}(\mathbf{\Lambda}_s^{-1} + \mathbf{a}_s \mathbf{a}_s^*) \right) \right\} - \sum_{m=1}^M \sum_{s=1}^S q_{ms} \ln q_{ms} \quad (37)
\end{aligned}$$

where  $q_{ms}$  and  $\tilde{q}_{ms}$  are given by (30) and (29) and  $\mathbf{\Lambda}_s$  and  $\mathbf{a}_s$  are given by (34) and (35). The second term in (37) is always negative; it can be viewed as a penalty term which ensures that increasing the number of speakers will not necessarily result in an increased value for  $\mathcal{L}(Q|\boldsymbol{\pi})$ .

### Estimating $\boldsymbol{\pi}$

In order to estimate  $\boldsymbol{\pi}$ , the means and precisions  $\mathbf{a}_s$  and  $\mathbf{\Lambda}_s$  as well as the probabilities  $q_{ms}$  are supposed to be known and the expression on the right hand side of (37) is regarded as a function of  $\boldsymbol{\pi}$ . The only dependency on  $\boldsymbol{\pi}$  is through the terms  $\tilde{q}_{ms}$  as in (29). Thus to estimate  $\boldsymbol{\pi}$  we have to choose  $\pi_1, \dots, \pi_S$  so as to maximize

$$\sum_{m=1}^M \sum_{s=1}^S q_{ms} \ln \pi_s$$

subject to the constraint

$$\sum_{s=1}^S \ln \pi_s = 1.$$

This gives

$$\pi_s = \frac{1}{M} \sum_{m=1}^M q_{ms} \quad (s = 1, \dots, S). \quad (38)$$

## VI. SUMMARY

The training algorithm that we have derived consists in cycling through the following steps:

- 1) Updating  $Q(\mathbf{Y})$  according to (36).
- 2) Updating  $Q(\mathbf{I})$  according to (30).
- 3) Updating  $\pi$  according to (38).

Each step is guaranteed to increase the value of  $\mathcal{L}(Q|\pi)$ .

## REFERENCES

- [1] F. Valente, “Variational Bayesian methods for audio indexing,” Ph.D. dissertation, Eurecom, Sep 2005.
- [2] D. MacKay, *Information theory, inference and learning algorithms*. New York, NY: Cambridge University Press, 2003.
- [3] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008. [Online]. Available: <http://www.crim.ca/perso/patrick.kenny>
- [5] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 345–359, May 2005.
- [6] F. Castaldo, D. Colibro, E. Dalmaso, P. Laface, and C. Vair, “Stream-based speaker segmentation using speaker factors and eigenvoices,” in *Proc. ICASSP*, Las Vegas, Nevada, Mar. 2008.
- [7] A. Corduneanu and C. Bishop, “Variational bayesian model selection for mixture distributions,” in *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, T. Richardson and T. Jaakkola, Eds. Morgan Kaufmann, 2001, pp. 27–34.
- [8] D. J. C. MacKay, “Comparison of approximate methods for handling hyperparameters,” *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.
- [9] P. Somervuo, “Speech modeling using variational Bayesian mixture of Gaussians,” in *Proc. ICSLP*, Denver, Colorado, Sept. 2002, pp. 1245–1248.
- [10] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, LLC, 2006.
- [11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and session variability in GMM-based speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.
- [12] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. ICCV 2007*, Rio de Janeiro, Brazil, Oct. 2007.