Diarization of Telephone Conversations using Factor Analysis

Patrick Kenny, Douglas Reynolds and Fabio Castaldo

EDICS Category: SPE-SPKR

***Abstract*— We report on work on speaker diarization of telephone conversations which was begun at the Robust Speaker Recognition Workshop held at Johns Hopkins University in 2008. Three diarization systems were developed and experiments were conducted using the summed-channel telephone data from the 2008 NIST speaker recognition evaluation. The systems are a Baseline agglomerative clustering system, a Streaming system which uses speaker factors for speaker change point detection and traditional methods for speaker clustering, and a Variational Bayes system designed to exploit a large number of speaker factors as in state of the art speaker recognition systems. The Variational Bayes system proved to be the most effective, achieving a diarization error rate of 1.0% on the summed-channel data. This represents an 85% reduction in errors compared with the Baseline agglomerative clustering system. An interesting aspect of the Variational Bayes approach is that it implicitly performs speaker clustering in a way which avoids making premature hard decisions. This type of soft speaker clustering can be incorporated into other diarization systems (although causality has to be sacrificed in the case of the Streaming system). With this modification, the Baseline system achieved a diarization error rate of 3.5% (a 50% reduction in errors).**

***Index Terms*— Diarization, speaker recognition, speaker segmentation, clustering, speaker factors, channel factors, variational Bayes**

I. Introduction

In recent years, factor analysis methods have proved to be very effective in speaker recognition. This is particularly true of telephone speech, thanks to the availability of large telephone speech corpora for training factor analysis models [1], [2], [3], [4]. It is therefore natural to try to bring factor analysis methods to bear on the problem of diarization of telephone conversations. This problem was chosen as one of the themes of the Robust Speaker Recognition Workshop held at Johns Hopkins University in the summer of 2008.

Three diarization systems were developed in the workshop: a Baseline system, and two factor analysis based systems which we refer to as the Streaming system and the Variational Bayes system. We will describe these systems in detail in Sections II, III and IV. The Baseline system uses the traditional approach of speaker segmentation with the Bayesian information criterion (BIC) followed by agglomerative speaker clustering [5]. It is non-causal in the sense that an entire speech file has to be processed before any diarization decisions can be made. The Streaming system performs speaker change point detection using a sliding window and a small number of speaker factors (e.g. 20) [6]. It has the advantage that it can be configured to run in real time (with a latency) and it determines the number of speakers dynamically. Like the Streaming system, the Variational Bayes system was inspired by the success of factor analysis in speaker recognition, but it is designed to exploit much larger numbers of speaker factors and it is non-causal. It incorporates some aspects of the Baseline system and it builds on Valente's pioneering work on Variational Bayes speaker diarization [7]. The Variational Bayes system proved to be (by far) the most effective of the three and it is the principal focus of this article.

Bayesian speaker diarization posits a hierarchical generative model for turn taking in a given conversation and aims to infer the number of participants and who is speaking at a given time using only probabilistic methods, all of which ultimately reduce to the sum and product rules for combining probabilities. In practice such an approach quickly runs into intractable integrals and posteriors so fast approximate inference methods such as Variational Bayes or Expectation Propagation have to be invoked [7], [8], [9]. Our principal contribution in this article is to show how the prior distribution on Gaussian Mixture Models (GMMs) used by Valente in constructing the hierarchical generative model for a conversation in his Variational Bayes approach to speaker diarization can be replaced by the eigenvoice and eigenchannel priors used in factor analysis based speaker recognition, and that this leads to excellent results in diarizing two-party telephone conversations. Although eigenchannels are widely used in speaker recognition, eigenvoices have a longer pedigree in speech recognition [10], and the results we will present indicate that eigenvoices are the key ingredient in our approach to diarization.

In the first four subsections of Section IV, we provide sufficient background material and mathematical detail to explain how our approach can be implemented. Readers unfamiliar with the Variational Bayes method will find a more leisurely account, including complete mathematical derivations, in the report [11] which was made available on-line prior to the workshop. Readers familiar with Variational Bayes will recognize that these derivations are essentially mechanical, as is the case in most applications of the Variational Bayes method [12].

Very similar derivations can be found in [13] where the

Patrick Kenny is with the Centre de recherche informatique de Montréal, Douglas Reynolds is with MIT Lincoln Labs and Fabio Castaldo is with the Politecnico di Torino.

Variational Bayes method is brought to bear on the problem of calculating the posterior distribution of the hidden variables in joint factor analysis. Joint factor analysis involves three types of hidden variable — speaker factors, channel factors and indicator variables which account for the alignment of frames with GMM mixture components. The alignment of frames with mixture components is assumed to be inherited from a Universal Background Model (UBM) in [1]; this has the advantage the joint posterior of the other hidden variables can be calculated in closed form (it is actually Gaussian) [14]. However the calculation is both complicated and computationally burdensome and the assumption that frames can be aligned using a UBM seems unsatisfactory — a speaker and channel dependent GMM ought to be used instead. The authors in [13] show how a Variational Bayes approach deals handily with both of these problems. Even in the case where the alignment is carried out with a UBM (for computational reasons, for example) the Variational Bayes approach turns out to be much less computationally demanding than the exact (non-iterative) approach and the Variational approximation is of very high quality in practice. (It can be shown that, at convergence, the Variational posterior has the same mean as the exact posterior. It is only the posterior covariance that is not calculated exactly.) In fact, the Variational posterior calculations in this situation are the same as the Gauss-Seidel calculations presented by Vogt in [4] (but casting these calculations in the Variational Bayes framework has the advantage of guaranteeing convergence, a question which was left open in [4]). Thus Variational Bayes seems to provide the best way of formulating joint factor analysis for speaker recognition as well as for speaker diarization.

As for our experiments, we used summed channel telephone data from the NIST 2008 speaker recognition evaluation (SRE) as a test set. This consists of 2215 telephone conversations, each involving just two speakers, of approximately five minutes duration (≈ 200 hours in total). This data was chosen since it enabled us to derive reference diarizations, needed for measuring diarization error rates, by using time marks from the speech recognition transcripts produced on each channel separately. Since this data served as the test set for one of the speaker detection tasks in the 2008 SRE we were also able to measure the effect of diarization errors on speaker recognition performance (Section V).

II. Baseline System

The Baseline system consists of three stages: speaker change point detection, speaker clustering and Viterbi re-segmentation. This is the most widely used approach to speaker diarization [5]. The acoustic features for the Baseline system consisted of 13 raw cepstral coefficients $c_0, \ldots, c_{12}$ (without any type of normalization).

A. Speaker change point detection

In the first stage, speaker change points are detected using a Bayesian Information Criterion (BIC) based distance between abutting windows of feature vectors. This technique searches for change points within a window using a penalized likelihood ratio test to determine whether the data in the window is better modeled by a single distribution (no change point) or by two different distributions (change point). If a speaker change point is found, the window is reset and the search restarted. If no change point is found, the window is increased and the search is redone. Full covariance Gaussians are used as distribution models.

B. Agglomerative speaker clustering

The purpose of the speaker clustering stage is to associate or cluster segments from the same speaker together. The clustering ideally produces one cluster for each speaker in the audio with all segments from a given speaker in a single cluster. Agglomerative speaker clustering is a hierarchical procedure consisting of the following steps:

0) Initialize leaf clusters of tree with speech segments.
1) Compute pair-wise distances between each cluster.
2) Merge closest clusters.
3) Update distances of remaining clusters to new cluster.
4) Iterate steps 1-3 until a stopping criterion is met.

The clusters are represented by a single full covariance Gaussian. Since we have prior knowledge that there are just two speakers present in the audio, we stop when we reach two clusters. (Otherwise a BIC based stopping criterion is used.)

C. Viterbi re-segmentation

The Viterbi re-segmentation stage uses the Baum-Welch algorithm and the speaker clusters produced by agglomerative clustering to train Gaussian Mixture Models for each speaker, re-segments the data using the Viterbi algorithm and iterates this procedure until convergence. We used 32 Gaussians for these GMMs. (It is well known in speaker diarization that a small number of Gaussians is sufficient to characterize a speaker in a given acoustic environment. This is different from the situation in speaker recognition where very large numbers of Gaussians have proved to be necessary to characterize speakers in arbitrary acoustic environments.) The total number of re-segmentation iterations performed was 20 and, on each Viterbi iteration, 5 iterations of Baum-Welch re-estimation were performed.

Viterbi re-segmentation (using unnormalized cepstral coefficients without derivatives) was found to significantly help all three diarization systems, although the improvement was less dramatic in the case of the Streaming system.

D. Soft speaker clustering

Since it is a greedy method, agglomerative speaker clustering is prone to making premature hard decisions and so can lead to sub-optimal results. This tendency can be mitigated by a soft speaker clustering method inspired by the Variational Bayes framework (equations (16) and (17) below).

After an initial diarization has been carried out we have two GMMs (each having 32 components), one for each speaker. Using these GMMs, for each speaker segment we can calculate a likelihood for each of the two speakers participating in the

conversation. By applying Bayes rule and assuming equal prior probabilities, we can convert these likelihoods into posterior probabilities and use these posteriors to weight the Baum-Welch statistics extracted from the segment.

In this way we obtain, for each speaker, a set of weighted Baum-Welch statistics, one for each segment. Pooling across segments, we synthesize a set of Baum-Welch statistics for the speaker and use these synthetic Baum-Welch statistics to re-estimate the speaker's GMM.

Whereas Viterbi re-segmentation makes hard decisions as to which segments are assigned to each speaker, this type of GMM training avoids such hard decisions. In practice, its effectiveness depends critically on appropriately normalizing the segment likelihoods referred to above before converting them to posterior probabilities using Bayes rule. The normalization procedure that we actually used was to divide the log likelihoods by the number of frames in the segment and multiply the result by 3.

E. Protocol

The primary performance measure that we used in evaluating the diarization systems is the NIST diarization error rate (DER). This is calculated by aligning a reference diarization output with a system diarization output and computing a time weighted combination of miss, false alarm and speaker error. [1] In evaluating DERs we took the reference speech activity marks as given, we ignored intervals containing overlapped speech and we ignored errors of less than 250 ms in the locations of segment boundaries. These are the traditional conventions used in evaluating diarization performance on two-way on telephone conversations [15]. Note that, although overlapped speech intervals do not count in evaluating DERs, the diarization systems do have to contend with overlapped speech in performing speaker segmentation and clustering.

It is well known that diarization systems typically exhibit wide variations in performance across test files. Thus In reporting results for a given system on the NIST 2008 summed channel data, we will show the standard deviation of the DER (calculated over all test files) in addition to the mean DER.

F. Baseline Results

DER results for the Baseline system are reported in Table I. Incorporating the soft clustering procedure in the Baseline system is seen to reduce the diarization error rate by almost 50%.

TABLE I

MEAN AND STANDARD DEVIATION OF DIARIZATION ERROR RATES (DER) ON THE NIST 2008 SUMMED CHANNEL TELEPHONE DATA FOR THE BASELINE SYSTEM; σ REFERS TO THE STANDARD DEVIATION OF THE DIARIZATION ERRORS.

	mean DER (%)	σ (%)
Baseline without soft-clustering	6.8	12.3
Baseline with soft-clustering	**3.5**	**8.0**

[1]DER scoring code available at www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl.

III. STREAMING SYSTEM

Speaker diarization using factor analysis was first introduced in [6] using a stream-based approach. This technique performs an on-line diarization where a conversation is seen as a stream of fixed duration time slices. The system operates in a causal fashion by producing segmentation and clustering for a given slice without requiring the following slices. Speakers detected in the current slice are compared with previously detected speakers to determine if a new speaker has been detected or previous models should be updated.

Given an audio slice, a stream of cepstral coefficients and their first derivatives are extracted. With a small sliding window (about one second), a new stream of speaker factors is computed and used to perform the slice segmentation. The dimension of the speaker factor space is quite small (e.g. 20) compared with the number used in speaker recognition (e.g. 300) due to the short estimation window. Figure 1 shows the stream of speaker factors in a slice where two different speakers are present.

In this new space, a clustering of the stream of speaker factors is done, producing a single multivariate Gaussian for each speaker. A BIC criterion is used to determine how many speakers there are in the slice. A Hidden Markov Model (HMM) using the Gaussian for each state associated to a speaker is built and through the Viterbi algorithm a slice segmentation is obtained.

In addition to the segmentation, a 256-component Gaussian Mixture Model (GMM) in the acoustic space is created for each speaker found in the audio slice. These models are used in the last step, slice clustering, where we determine if a speaker in the current audio slice was present in previous slices, or is a new one.

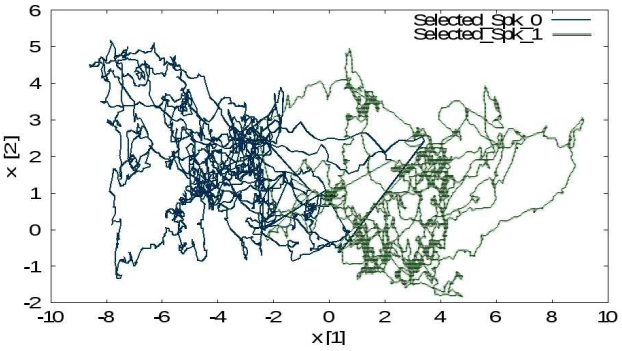

Fig. 1. The Streaming system computes the trajectory of a vector of speaker factors in a slice using a sliding window (the first two parameters are shown here). The distribution of points shows a bimodal behavior when two speakers are present so each speaker can be effectively modeled by means of a single Gaussian. (In this instance, the trajectory for one of the speakers is concentrated on the left and the trajectory for the other on the right.) The final slice labeling is obtained using the Viterbi algorithm.

Using an approximation to the Kullback-Leibler divergence, we find the closest speaker model built in previous slices to each speaker model in the current slice. If the divergence is below a threshold the previous model is adapted using the model created in the current slice, otherwise the current model is added to the set of speaker models found in the audio.

The final segmentation and speakers found from the on-line processing can further be refined using Viterbi re-segmentation over the entire file, as explained in Section II-C.

A. Streaming Results

The Streaming system was implemented using unnormalized mel cepstral coefficients and their derivatives $c_1, \ldots, c_{12}, \Delta c_0, \ldots, \Delta c_{12}$ as features.

DER results obtained with the Streaming system are shown in Table II. Of the three systems tested, the Streaming system had the best performance out of the box (first line), with some further gains with the non-causal Viterbi re-segmentation (second line). The Viterbi re-segmentation step was added to ensure a fair comparison with the other two systems although it is inconsistent with the primary purpose of the Streaming system, namely incremental on-line decision making. We did not test the (equally non-causal) soft-clustering technique with the Streaming system.

TABLE II
MEAN AND STANDARD DEVIATION OF DIARIZATION ERROR RATES (DER) ON THE NIST 2008 SUMMED CHANNEL TELEPHONE DATA FOR THE STREAMING SYSTEM; σ REFERS TO THE STANDARD DEVIATION OF THE DIARIZATION ERRORS.

	mean DER (%)	σ (%)
Streaming without Viterbi	5.8	11.1
Streaming + Viterbi	**4.6**	**8.8**

IV. VARIATIONAL BAYES SYSTEM

A. Motivation

As we have just seen, effective speaker change point detection can be achieved using as little as 20 speaker factors. However the number of speaker factors used in state of the art speaker recognition systems is typically much larger, e.g. 300 [1]. With factor analysis models of this size, it is difficult to process speech data incrementally; in practice factor analysis methods have to be applied in batch mode and extracting speaker and channel factors from a speech file is computationally burdensome. On the other hand, most speaker diarization algorithms work with very short intervals of speech, typically a few seconds long. For example it may be required to determine whether a speaker change point occurs at a given frame or whether the same speaker is talking in two short segments. Thus it is not obvious how the two technologies — speaker diarization and large scale factor analysis — can be integrated.

The Variational Bayes method of speaker diarization developed by Valente [7] is a natural way of solving this problem. In the technical report [11], we modified Valente's formulation in order to incorporate the factor analysis priors defined by eigenvoices and eigenchannels [1], and we simplified it to take account of the fact that we are dealing with a diarization problem in which the number of speakers is given.

The advantages of the Variational Bayes approach are that it is fully probabilistic, it comes with EM-like convergence guarantees and it avoids making premature hard decisions as in agglomerative speaker clustering. Furthermore Bayesian methods are automatically regularized in the sense that, in theory at least, they are not subject to the overfitting problems which maximum likelihood methods are prone to. Thus Bayesian model selection can be used to determine the number of speakers participating in a conversation without having to resort to BIC-like fudge factors [7]. Since our test bed consists of two-way telephone conversations we did not get to explore this possibility but we will return to this question briefly in Section IV-D.

B. Informal description

We assume at the outset that we are given a conversation involving just two speakers and that speaker change points are given. The diarization problem is formulated as one of calculating, for each speaker segment, the posterior probabilities of the events that one speaker or the other is talking in the segment, as illustrated in Figure 2. The initial speaker segmentation need not be very accurate. (To begin with, a uniform segmentation into 1 second intervals after removing silences can be used; this assumption can be relaxed in a second pass after Viterbi re-segmentation, as described in Section II-C.) We refer to these posterior probabilities as *segment posteriors*. Once these segment posteriors have been calculated, it is a straightforward matter to make a hard decision as to which of the two speakers is talking at any given time and this gives a solution to the diarization problem.

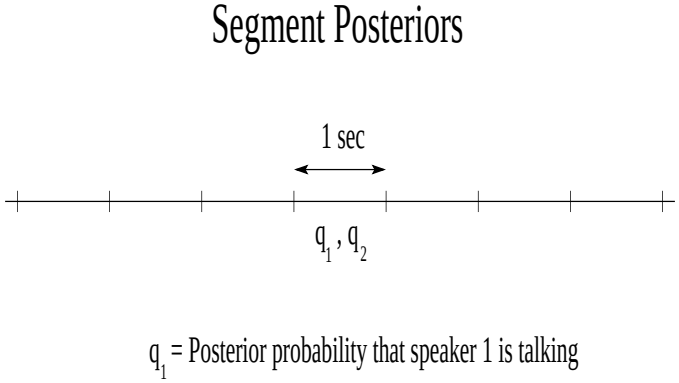

Fig. 2. Assume that the speech file has been partitioned into segments in which just one speaker is talking; these segments are shown here as 1 sec intervals. The object of the Variational Bayes algorithm is to estimate a pair of posterior probabilities for each speaker segment. In the text, the posterior probability that speaker s is talking in segment m is denoted by q_{ms}.

The Variational Bayes approach is fully probabilistic. It aims to solve the diarization problem by a consistent application of the rules of probability (marginalization and conditioning) using a hierarchical generative model of the speech data that contains three types of hidden random variable whose roles are to specify

1) The assignment of segments to speakers.
2) The parameters of speaker GMMs.
3) The assignment of frames to Gaussians in the speaker GMMs.

In his handling of point 2), Valente used a fully Bayesian treatment of the problem of GMM estimation in which all

of the GMM parameters — mixture weights, mean vectors and covariance matrices — are treated as random variables having appropriate prior distributions. (Fully Bayesian GMM estimation is the paradigmatic example of Variational Bayesian inference [9], [16].) Our reason for borrowing the Variational Bayesian framework is that it enables us to substitute eigenvoice and eigenchannel priors on GMMs in place of 2) and hence to bring large scale factor analysis methods to bear on the speaker diarization problem.

The collective experience of workers in speaker recognition using speaker GMMs with very large numbers of Gaussians has been that mixture weights and covariance matrices can be treated as speaker-independent but a large body of research has been devoted to finding powerful prior distributions for Gaussian mean vectors [1], [2], [3], [4]. We describe briefly how these priors are constructed and how they are used in speaker recognition. Recall that the term *supervector* is used to refer to the concatenation of the mean vectors in a Gaussian mixture model. The assumption in eigenvoice modeling is that speaker supervectors have a Gaussian distribution of the form

$$\boldsymbol{s} \quad = \quad \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}. \tag{1}$$

Here $\boldsymbol{s}$ is a randomly chosen speaker dependent supervector; $\boldsymbol{m}$ is a speaker-independent supervector; $\boldsymbol{V}$ is a rectangular matrix of low rank whose columns are referred to as eigenvoices; the vector $\boldsymbol{y}$ has a standard normal distribution; and the entries of $\boldsymbol{y}$ are the speaker factors. In Bayesian terms, (1) is a highly informative prior distribution: supervectors are of extremely high dimension in practice but (1) confines speaker supervectors to a low dimensional affine subspace of the supervector space. On the other hand, the factorial priors on GMM parameters used by Valente impose relatively weak constraints on speaker models.

To model channel effects in speaker recognition, (1) is modified as follows:

$$\boldsymbol{s} \quad = \quad \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{V}\boldsymbol{y}. \tag{2}$$

Here $\boldsymbol{s}$ is a randomly chosen speaker and *channel* dependent supervector; $\boldsymbol{U}$ is a rectangular matrix of low rank whose columns are referred to as eigenchannels; the vector $\boldsymbol{x}$ has a standard normal distribution; and the entries of $\boldsymbol{x}$ are the channel factors.

In speaker recognition (2) is used in different ways in enrollment and testing. In enrolling a given target speaker, both $\boldsymbol{x}$ and $\boldsymbol{y}$ are estimated from the enrollment data but $\boldsymbol{x}$ is discarded so as to obtain an estimate of the speaker's supervector which is independent of channel effects; in matching a target speaker against a given test utterance, $\boldsymbol{x}$ is treated as random. It turns out that incorporating $\boldsymbol{U}\boldsymbol{x}$ into the Variational Bayes diarization framework presents no extra difficulty since (2) is formally equivalent to (1) as can be seen by writing it as

$$\boldsymbol{s} \quad = \quad \boldsymbol{m} + \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}. \tag{3}$$

So although we will perform experiments with eigenchannels as well as eigenvectors, we will only refer to equation (1) in subsequent development.

In our version of point 2), each of the two speakers in the given conversation will be represented by a hidden vector of speaker factors. In the course of calculating the segment posteriors referred to in Figure 2, we will also calculate two *speaker posteriors* as illustrated in Figure 3. For each of the two speakers, the corresponding speaker posterior is a multivariate Gaussian distribution on speaker factors which models the location of the speaker in the speaker factor space. The mean of this distribution can be thought of as a point estimate of the speaker's location and the covariance matrix as a measure of the uncertainty in this point estimate. The multivariate Gaussian assumption here is the same as in the Streaming system (illustrated in Figure 1) but it is supposed to hold over the whole speech file rather than on a slice-by-slice basis. Other differences are that the Variational Bayes system uses Gaussians of dimension 300 rather than 20 and covariance matrices which are full rather than diagonal.

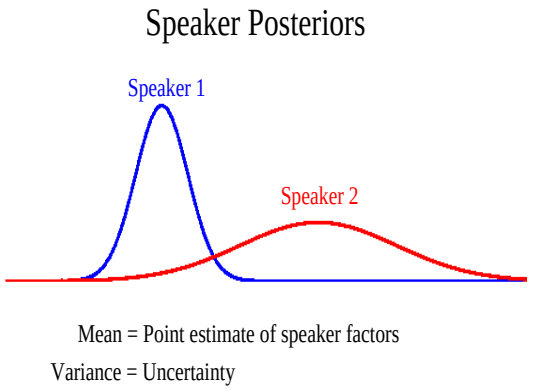

Fig. 3. Each speaker is represented by a posterior distribution on the space of speaker factors. The mean of this distribution provides a point estimate of the speaker's supervector, according to (1). The covariance matrix models the uncertainty of the speaker's location in the speaker factor space. The mean vector and precision matrix of the posterior for speaker s are denoted by $\boldsymbol{a}_s$ and $\boldsymbol{\Lambda}_s$ in the text.

We referred in point 3) above to a third type of hidden variable whose role is to specify the alignment of speech frames with Gaussians in speaker GMMs. Unlike Valente, we use a Universal Background Model to carry out the alignment (thus we treat the alignment of frames with Gaussians in speaker GMMs as deterministic). This strategy has proved to be very successful in speaker recognition [1]. It has the advantage that Baum-Welch statistics extracted from each of the speaker segments with the UBM are sufficient statistics for all of the probability calculations and greatly alleviates the computational burden of our version of Variational Bayes speaker diarization.

It is beyond the scope of this article to attempt a full exploration of the ways in which the Variational Bayes method can be applied to the diarization problem but it may be of interest to sketch some directions for future research. A glance at the paper [17] will be enough to convince the reader of the power of Variational Bayes for speech processing applications. It is shown there how a unified probabilistic framework can be developed for denoising and dereverberation of speech signals using an informative prior distribution on clean speech signals and Bayesian inference. A particularly striking aspect

of this work is the ease and naturalness with which multiple microphones can be handled. Our version of Variational Bayes speaker diarization can likewise handle multiple microphones without difficulty, at least in principle. All that is required is to associate with each speaker a vector of speaker factors ($\boldsymbol{x}$) and a collection of channel factor vectors ($\boldsymbol{y}$), one for each microphone (rather than a single channel factor vector as in equation (2)).

Detecting segments in which two or more speakers are talking is currently an active area of research in speaker diarization and one can speculate about how our approach might be extended to provide a principled solution to this problem. The question here is how to construct a probabilistic model for supervector $\boldsymbol{s}$ associated with such a segment. Assuming that there are just two speakers talking in the segment, that the corresponding speaker factor vectors are $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ and ignoring channel effects, it is reasonable to postulate a model of the form

$$\boldsymbol{s} = \boldsymbol{m}' + \boldsymbol{V}'(\boldsymbol{y}_1 + \boldsymbol{y}_2)$$

in place of equation (1). Estimating the matrix $\boldsymbol{V}'$ presents no difficulty in principle; all that would be required would be to synthesize a training set consisting of speech files obtained by summing pairs of recordings of different speakers.

In developing our Variational Bayes system, we followed Valente's lead in abstracting the problem of speaker change point detection rather than attempt to incorporate it into the variational posterior calculation. The practice of using a uniform segmentation to begin with and then refining it with Viterbi re-segmentation is well established but, in the context of a Variational Bayes system, it is a dereliction of the Bayesian philosophy (*lèse-Bayes* so to speak). In order to deal with the segmentation problem in a fully Bayesian manner, a natural solution would be to incorporate an extra level into the generative model, namely a hidden Markov model having one state per speaker, and enforce a minimum duration constraint on speaker segments. To be realistic, the minimum duration would have to be quite short, say 250 ms. This may lead to problems however. Referring to Fig. 2, we found in some preliminary experiments that uniformly partitioning a speech file into segments of duration 500 ms (rather than 1 second) led to unsatisfactory results. So it seems that research would be needed to get this type of strategy to work in practice.

Admittedly, it is not clear whether investigating this problem would lead to improved diarization error rates. In the early stages of development (as described in Section IV-F below) our best results were obtained by doing two passes on each speech file, each pass consisting of a run of Variational Bayes followed by Viterbi re-segmentation but, in the final version of the system (as described in Section IV-H), Viterbi re-segmentation in the second pass turned out not to be helpful and was suppressed. So, in the end, the only role played by Viterbi re-segmentation seems to be to correct the crude initial segmentation of the data. Where a fully Bayesian approach might be advantageous is that it would obviate the need for a second pass altogether.[2]

C. Modeling assumptions

As we have explained, our version of Variational Bayes contains just two types of hidden random variable, one which specifies the assignment of segments to speakers and the other which specifies the location of speakers in the space of speaker factors. To proceed we need to introduce some notation. In this and the following section we denote the number of speakers participating in the conversation by S and we drop the requirement that $S = 2$.

We are assuming that speaker segment boundaries are given. Denote the speaker segments by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ and set $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$. With each segment $m = 1, \ldots, M$, we associate an $S \times 1$ indicatorr vector $\boldsymbol{i}_m$ whose components are defined as follows: for $s = 1, \ldots, S$, $i_{ms} = 1$ if speaker s is talking in the segment and $i_{ms} = 0$ otherwise. We are assuming that there is just one speaker in each segment so that the vector $\boldsymbol{i}_m$ has just one non-zero component. Set $\boldsymbol{I} = (\boldsymbol{i}_1, \ldots, \boldsymbol{i}_M)$. (This is not to be confused with the identity matrix.)

We assign a prior probability to the event that a speaker s is talking in a given segment; we denote this by π_s and set

$$\pi_s = \frac{1}{S}. \tag{4}$$

Let $\boldsymbol{y}_s$ be the vector of speaker factors which defines the location of speaker s in the space of speaker factors. As in (1) we assume a standard normal prior $N(\cdot|\boldsymbol{0}, \boldsymbol{I})$ for $\boldsymbol{y}_s$. Set $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_S)$. Thus $\boldsymbol{Y}$ and $\boldsymbol{I}$ the hidden variables in our probability model and $\boldsymbol{X}$ is the observable data.

We explained in Lemma 1 in [19] how to calculate the probability $P(\boldsymbol{x}|\boldsymbol{y})$ of a segment $\boldsymbol{x}$ given a vector of speaker factors $\boldsymbol{y}$. In order to explain how to calculate the likelihood function for our probability model, it is necessary to recapitulate this result here. For each mixture component c, we denote the centered first and second order Baum-Welch statistics extracted from the segment with the UBM by $\tilde{F}_c$ and $\tilde{S}_c$:

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

where m_c is the subvector of $\boldsymbol{m}$ in (1) which corresponds to the mixture component c and $\gamma_t(c)$ is the posterior probability of the event that the observation at time t is generated by

[2]Since completing this work, we learned that a fully Bayesian solution to the speaker diarization problem has been proposed [18]. The authors achieve state of the art results on a NIST meeting task using a Bayesian method to detect speaker change points and a novel front end. To deal with the fact the number of speakers is unknown (in general), the authors posit a HMM with a countably infinite state space whose transition probabilities are drawn from a hierarchical Dirichlet process. Posterior calculations are performed using Gibbs sampling rather than Variational Bayes. Non-overlapping 250 ms frames are used in the front end and a minimum duration of 500 ms is imposed on speaker segments.

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

Define

$$\begin{aligned} G &= \sum_{c=1}^{C} N_c \ln \frac{1}{(2\pi)^{F/2}|\Sigma_c|^{1/2}} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}\right) \\ \text{and } H(\boldsymbol{y}) &= \boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}} - \frac{1}{2}\boldsymbol{y}^*\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{y} \end{aligned} \quad (5)$$

where Σ_c is the covariance matrix associated with mixture component c and $\boldsymbol{\Sigma}$ is the $CF \times CF$ dimensional covariance matrix whose diagonal blocks are $\Sigma_1, \ldots, \Sigma_C$. Then

$$\ln P(\boldsymbol{x}|\boldsymbol{y}) = G + H(\boldsymbol{y}). \quad (6)$$

We can now write down the complete likelihood function for our probability model:

$$P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}) = P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{I})P(\boldsymbol{Y})P(\boldsymbol{I}) \quad (7)$$

where

$$\begin{aligned} P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{I}) &= \prod_{m=1}^{M}\prod_{s=1}^{S} P(\boldsymbol{x}_m|\boldsymbol{y}_s)^{i_{ms}} \\ \text{and } P(\boldsymbol{I}) &= \prod_{m=1}^{M}\prod_{s=1}^{S} \pi_s^{i_{ms}} \end{aligned} \quad (8)$$

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

Note that if the posterior $P(\boldsymbol{I}|\boldsymbol{X})$ were tractable we would already be done since the *maximum a posteriori* assignment of segments to speakers — that is, the output of the diarization system — would be given by

$$\operatorname{argmax}_{\boldsymbol{I}} P(\boldsymbol{I}|\boldsymbol{X}). \quad (9)$$

D. Variational posterior calculations

Variational Bayes is a principled way of approximating intractable posteriors as we now briefly explain. Set $\theta = (\boldsymbol{Y}, \boldsymbol{I})$. For any probability distribution $Q(\theta)$, define

$$\mathcal{L}(Q) = \int Q(\theta) \ln \frac{P(\boldsymbol{X}, \theta)}{Q(\theta)} \mathrm{d}\,\theta.$$

It is a well known consequence of the non-negativity of Kullback-Leibler divergences [9] that $\mathcal{L}(Q)$ is a lower bound on the log evidence:

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

with equality holding iff

$$Q(\theta) = P(\theta|\boldsymbol{X}).$$

Thus the value of the lower bound is a measure of how well the distribution $Q(\boldsymbol{Y}, \boldsymbol{I})$ approximates the true posterior $P(\boldsymbol{Y}, \boldsymbol{I}|\boldsymbol{X})$. In Variational Bayes, a factorization of the form

$$Q(\boldsymbol{Y}, \boldsymbol{I}) = Q(\boldsymbol{Y})Q(\boldsymbol{I}) \quad (10)$$

is assumed and the marginals $Q(\boldsymbol{Y})$ and $Q(\boldsymbol{I})$ are iteratively refined in such a way as to increase the value of the lower bound on successive iterations. The update rules are

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

where $E_{\boldsymbol{Y}}[\cdot]$ indicates an expectation calculated with respect to $Q(\boldsymbol{Y})$ and similarly for $E_{\boldsymbol{I}}[\cdot]$. (The constants are chosen so as to ensure that the marginals integrate to 1.) Because these formulas are coupled they have to be applied iteratively. Convergence is guaranteed [9].

It turns out that these update formulas imply that $Q(\boldsymbol{Y})$ and $Q(\boldsymbol{I})$ both factorize in the same way as $P(\boldsymbol{Y})$ and $P(\boldsymbol{I})$:

$$Q(\boldsymbol{I}) = \prod_{m=1}^{M} Q(\boldsymbol{i}_m) \quad (11)$$

where, for each segment m,

$$Q(\boldsymbol{i}_m) = \prod_{s=1}^{S} q_{ms}^{i_{ms}} \quad (12)$$

and

$$Q(\boldsymbol{Y}) = \prod_{s=1}^{S} N(\boldsymbol{y}_s|\boldsymbol{a}_s, \boldsymbol{\Lambda}_s^{-1}). \quad (13)$$

where the means and precisions $\boldsymbol{a}_s$ and $\boldsymbol{\Lambda}_s$ and the probabilities q_{ms} remain to be specified. We state the update rules without proof, referring the reader to [11] for the derivations.

1) Updating the segment posteriors: For each segment m and speaker s, define $\tilde{q}_{ms}$ by setting $\ln \tilde{q}_{ms}$ equal to

$$\ln \pi_s P(\boldsymbol{x}_m|\boldsymbol{a}_s) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{V}^*\boldsymbol{N}_m\boldsymbol{\Sigma}^{-1}\boldsymbol{V}\boldsymbol{\Lambda}_s^{-1}\right). \quad (14)$$

where the first term in this expression is given by (6). Then the update formula for the segment posteriors is

$$q_{ms} = \frac{\tilde{q}_{ms}}{\sum_{s'=1}^{S}\tilde{q}_{ms'}}. \quad (15)$$

To interpret this, notice that the second term in (14) vanishes if there is no uncertainty in the point estimate $\boldsymbol{a}_s$ of the speaker factors (that is, if the posterior covariance matrix $\boldsymbol{\Lambda}_s^{-1}$ is zero). In this case the calculation of the segment posteriors reduces to a straightforward application of Bayes rule.

2) Updating the speaker posteriors: For each speaker s, synthesize speaker dependent Baum-Welch statistics $\boldsymbol{N}(s)$ and $\tilde{\boldsymbol{F}}(s)$ by setting

$$\boldsymbol{N}(s) = \sum_{m=1}^{M} q_{ms}\boldsymbol{N}_m \quad (16)$$

$$\tilde{\boldsymbol{F}}(s) = \sum_{m=1}^{M} q_{ms}\tilde{\boldsymbol{F}}_m. \quad (17)$$

The update formulas for $\boldsymbol{a}_s$ and $\boldsymbol{\Lambda}_s$ are

$$\boldsymbol{\Lambda}_s = \boldsymbol{I} + \boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(s)\boldsymbol{V} \quad (18)$$

$$\boldsymbol{a}_s = \boldsymbol{\Lambda}_s^{-1}\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}(s) \quad (19)$$

These formulas are formally identical with those in Proposition 1 of [19]. Equations (16) and (17) implement the soft speaker clustering referred to in Section II-D.

3) The role of $\mathcal{L}$: An expression for $\mathcal{L}(Q)$ in terms of $q_{ms}, \tilde{q}_{ms}, \mathbf{\Lambda}_s$ and $\boldsymbol{a}_s$ is given in equation (37) of [11]. The speaker and segment posteriors are updated alternately until the value of $\mathcal{L}$ converges. On convergence, diarization is performed by assigning each segment m to the speaker given by

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

As for initialization, it is necessary to initialize the speaker posteriors or the segment posteriors, but not both. We found that assigning random values to the segment posteriors is effective. This avoids the tricky problem of initializing the speaker posteriors and, except where otherwise indicated, this is the only type of initialization that we used.

In situations where the number of speakers is given (as in the case of two-way telephone conversations), the role of $\mathcal{L}$ is limited to monitoring convergence and checking that the posterior update formulas have been correctly implemented. In other situations it is possible, in theory at least, to use $\mathcal{L}$ to determine the number of speakers, without having recourse to tunable fudge factors to counter the overfitting tendency which maximum likelihood methods are prone to.

This is known as Bayesian model selection. Assuming that an upper bound on the number of speakers is given, the idea is to fit several models each containing different numbers of speakers to the speech data and use $\mathcal{L}$ as a criterion for determining which model gives the best fit [7], [9], [20].

E. Feature sets and factor analysis configurations

The principal question confronting us in building a Variational Bayes diarization system was which acoustic feature set to use. The question of whether or not to use short-term Gaussianization [21] was particularly unclear. Short-term Gaussianization is very widely used in speaker recognition where it is viewed as an effective antidote to channel effects (non-stationary as well as stationary). It has been our experience that there is a very strong synergy between factor analysis and Gaussianization in speaker recognition [22].

However, in speaker diarization of telephone speech, channel effects are if anything helpful rather than hurtful since the problem is not so much to distinguish between two speakers as to distinguish between two conversation sides, each of which may be subject to different channel effects. So mitigating channel effects may not be a good idea. Furthermore Gaussianization using a 3 second sliding window is apt to blur speaker change points. For these reasons, the question of whether short-term Gaussianization is helpful or harmful in speaker diarization is controversial [23], [24], [25].

Thus we ended up experimenting with several acoustic feature sets which we now describe:

1) BUT features: Like other researchers participating in the Robust Speaker Recognition workshop, we undertook to experiment in the first instance with a 39-dimensional acoustic feature set that had been optimized by Brno University of Technology for speaker recognition [3]. This feature set was derived from the cepstral coefficients $c_0, \ldots, c_{12}$ by calculating first, second and third derivatives; then short term Gaussianization was applied and a HLDA projection was used to reduce the dimensionality from 52 to 39.

As for the factor analysis configuration, we used a universal background model (UBM) with 512 Gaussians. We trained two gender-independent factor analysis models, one with 200 eigenvoices and no eigenchannels, the other with 200 eigenvoices and 50 eigenchannels. (This is not the configuration used by BUT for speaker recognition.)

We used about 1500 hours of speech data drawn from the Switchboard and Mixer corpora for training these factor analysis models. Essentially the same data was used for training the other factor analysis models described below.

2) CRIM features: The second feature set that we experimented with was the 40 dimensional feature set used in the "small" factor analysis systems described in [1]. This feature set consists of 20 Gaussianized cepstral coefficients $c_0, \ldots, c_{19}$ together with their first derivatives. Using this feature set, we trained a UBM with 1024 Gaussians and two gender-independent factor analysis models, one with 300 eigenvoices and no eigenchannels, the other with 300 eigenvoices and 100 eigenchannels. Because the factor analysis configurations are different, the comparison with the BUT features is not back-to-back but we did not explore this issue as neither feature set seems to be ideal for speaker diarization.

3) Raw cepstral features: The third feature set that we used consists of 20 *un*-Gaussianized cepstral coefficients $c_0, \ldots, c_{19}$ without first derivatives. This was inspired by the feature set used in the Baseline system, but we chose to use 20 coefficients rather than 13 because this has proved useful in speaker recognition [1]. (The streaming system uses first derivatives as well as cepstral coefficients but opinion is divided as to whether derivatives are helpful in speaker diarization.) Again we trained a UBM with 1024 Gaussians and two gender-independent factor analysis models, one with 300 eigenvoices and no eigenchannels, the other with 300 eigenvoices and 100 eigenchannels.

F. Variational Bayes Results

Results obtained with the Variational Bayes system and the various feature sets are reported in Table III; "EV without EC" refers to factor analysis models containing eigenvoices but no eigenchannels; eigenchannels were used to obtain the results in the "EV with EC" column.

The first thing to note (lines 1 and 2) is that Viterbi re-segmentation results in a 50% reduction in diarization errors. The results in line 1 were obtained by segmenting the speech files with GMMs adapted from the UBM by eigenvoice MAP adaptation [19] using the point estimates of the speaker factors provided by the Variational Bayes algorithm; on the other hand, the results in line 2 were obtained with the Viterbi re-segmentation procedure described in Section II-C. Another difference is that BUT features were used in obtaining the results in line 1 but raw cepstral features were used for the Viterbi re-segmentation results in line 2. This was the first indication that short-term Gaussianization might be harmful.

For the second pass referred to in line 3, the speaker change points found by Viterbi re-segmentation were used to initialize

a second run of Variational Bayes and this was followed by another Viterbi re-segmentation. The improvements obtained in this way show (unsurprisingly) that the initial segmentation into 1 second intervals is sub-optimal.

The results in line 4 were obtained in the same way as those in line 3 with CRIM features substituted for BUT features; an improvement in mean DER is noted although the standard deviation is larger in the "EV without EC" column.

The best results, namely a DER of 1.9% with standard deviation 5.6%, were obtained with the raw cepstral features (line 6) using using two passes (each consisting of Variational Bayes followed by Viterbi re-segmentation, as in line 3). For completeness, we ran an additional experiment without Viterbi re-segmentation in either the first or second pass (line 5) to see how much of the improvement in going from line 1 to line 6 could be attributed to the change in the acoustic features and how much to using Viterbi re-segmentation. Looking at the results in the "EV without EC" column it appears that most of the improvement is due to using raw cepstral features but the Viterbi re-segmentation is clearly helpful. The result in line 5 can be compared to the "Streaming without Viterbi" result presented in the first line of Table II. It is worth noting that the best feature set for the Variational Bayes system is very similar to the feature sets used by the Baseline and Streaming systems.

TABLE III
DIARIZATION ERROR RATES (DER) ON NIST 2008 SRE SUMMED CHANNEL TEST DATA OBTAINED USING VARIATIONAL BAYES (VB) AND EIGENVOICES; σ REFERS TO THE STANDARD DEVIATION OF THE DIARIZATION ERRORS; EV STANDS FOR EIGENVOICES, EC FOR EIGENCHANNELS.

		EV without EC		EV with EC	
		DER	σ	DER	σ
BUT features					
1	VB without Viterbi	9.1%	11.9%	9.2%	11.9%
2	VB + Viterbi	4.5%	8.5%	4.6%	8.6%
3	VB + Viterbi + 2^{nd} pass	3.8%	7.6%	3.9%	7.9%
CRIM features					
4	VB + Viterbi + 2^{nd} pass	3.3%	7.8%	3.5%	7.7%
Raw cepstral features					
5	VB + 2^{nd} pass, no Viterbi	2.2%	5.8%	2.9%	7.4%
6	VB + Viterbi + 2^{nd} pass	**1.9**%	**5.6**%	2.3%	6.4%

G. Why are eigenchannels ineffective?

Comparing the "EV with EC" and "EV without EC" columns in Table III shows that adding eigenchannels leads to slight but consistent degradations in performance. The results obtained with eigenvoices alone are so good that this may not seem to be worth worrying about. However the eigenvoices were estimated on a large quantity of telephone speech and it is doubtful that a Variational Bayes diarization system trained exclusively on telephone speech would perform well on other types of speech such as the interview data in the NIST 2008 SRE. In speaker recognition, it turns out to be quite easy to port a large factor analysis system trained exclusively on telephone speech to handle other types of microphones and channels by estimating supplementary eigenchannels on appropriate development data [1]. One would hope that the same strategy would prove to be successful in speaker diarization but given the results we have reported in Table III this is not a foregone conclusion. Considering that eigenchannels have proved to be very successful in speaker recognition with telephone speech, the question of why eigenchannels did not prove helpful in speaker diarization in our experiments seemed to be worth investigating.

On the face of it, adding eigenchannels ought to be helpful in Variational Bayes speaker diarization because they give a more realistic prior on conversation sides than eigenvoices alone. This observation suggested an experiment with another way of estimating a prior on conversation sides, namely training a factor analysis model in which each utterance is treated as a different "speaker". Since speaker and channel effects are not distinguished in this type of training, this is just a principal components model. As in the experiment reported in line 6 of Table III, we used raw cepstra as acoustic features and set the number of principal components to 300. We obtained a mean DER of 2.1% and a standard deviation of 5.6% with this type of factor analysis. Comparing with line 6 of Table III we see that this is an improvement on the results obtained using eigenchannels but they are still not as good as those obtained with eigenvoices alone.

There is a potential weakness in the way both the principal components and eigenchannels are estimated, namely that the factor analysis training data that we used consists of 4-wire rather than 2-wire or summed channel data where a new type of channel effect emerges, namely cross-talk. To see if this could explain why eigenchannels were not helping we artificially removed all of the overlapping speech in the NIST 2008 summed channel data and replicated the experiments in line 6 of Table III. Somewhat better results are obtained under these conditions: a mean DER of 1.7% with standard deviation of 5.4% in the case of eigenvoices alone and a mean DER of 2.0% with standard deviation of 6.2% when the eigenchannels are added. Of course these improvements were to be expected but they show that eigenchannels are unhelpful even under these idealized conditions.

We note in passing that since these improvements are rather small they suggest that, in diarizing telephone conversations as distinct from meetings, detecting speech overlaps is a less promising line of research than the oracle results presented in [26] would indicate. Note however that the telephone data provided by NIST involves conversations between strangers which tend to be less animated than conversations between people who know each other well. Also, care is needed in comparing the results in [26] with ours since diarization errors are not measured in exactly the same way as in Section II-E.

Another potential weakness in traditional eigenchannel modeling is that all channel effects (or, more generally, session effects) are assumed to be stationary rather than transient. Some evidence that this assumption may not be ideal for speaker recognition is presented in [27]. In [28], it was shown how explicit modeling of transient session effects could be used to good effect in speaker diarization. This seems to merit

more research.

H. Algorithmic refinements

Like any EM-type algorithm, Variational Bayes may fail to converge to a good local maximum of $\mathcal{L}$. We used the opportunity afforded by a second implementation to take some measures to protect against this type of behavior. This led to some surprisingly large performance gains.

Initializing the Variational Bayes algorithm by assigning random values to the segment posteriors q_{ms} (as described in Section IV-D.3) is clearly not ideal. This type of initialization seems to be particularly deleterious in the case of recordings where one speaker dominates the conversation: it frequently happens in such cases that the two speaker posteriors found by the Variational Bayes algorithm serve only to model the dominant speaker and the diarization error rate for the recording may be 40% or more.

In our second implementation, we ran the Variational Bayes algorithm on each speech file with four initializations in addition to the initialization described in Section IV-D.3, retaining only the results of the run which produced the highest value of $\mathcal{L}$. In each case, we initialized two speaker posteriors (rather than initializing segment posteriors as in Section IV-D.3). In the first instance, we produced three speaker posteriors by running the Variational Bayes algorithm (using the initialization described in Section IV-D.3) hypothesizing three speakers rather than two; taking these three speaker posteriors two at a time gave us three new initializations. We produced one further initialization by running the Variational Bayes algorithm hypothesizing a single speaker and combining the resulting speaker posterior with the speaker posterior defined by the standard normal distribution on the speaker factors. Thus, in our second implementation, we run the Variational Bayes algorithm five times in the first pass and five times in the second pass rather than once in each case as in our first implementation.

We also observed that the Variational Bayes algorithm tends to converge very quickly (note that there are only 2 terms in the variational factorization (10)); as a result, the local maximum of $\mathcal{L}$ which is found tends to depend on the initialization. This type of behavior can be inhibited by scaling the Baum-Welch statistics by a factor of α where $\alpha \ll 1$; for the raw cepstral features described in Section IV-E.3 setting $\alpha = 0.1$ seems to give the best results. This works because it effectively decreases the number of observation vectors in each speech file, thereby flattening the speaker and segment posterior distributions and slowing down the rate of convergence. This type of scaling can be viewed as compensating for the frame independence assumption underlying Gaussian mixture modeling. As a rationale for the surprisingly low value of α, we remark that such compensation really ought to be performed not only at diarization time but also when the matrix $\boldsymbol{V}$ is estimated so that α has to do double duty. (We note in passing that scaling the Baum-Welch statistics — or, equivalently, scaling the matrix $\boldsymbol{V}$ — also seems to be essential to get the Bayesian model selection procedure described in Section IV-D.3 working properly.)

In our second implementation, we also paid closer attention to the initial segmentation of each speech file in the first pass. We simply removed the silences and chopped the speech up into 1 second intervals in our first implementation but, for the second implementation, we used the information contained in the speech/silence boundary markers to segment the speech data, inserting additional potential speaker change points at speech/silence boundaries.

With these refinements we found that Viterbi re-segmentation was helpful only if it is applied in the first pass so we removed it from the second pass.

When we used the re-implemented Variational Bayes system to replicate the experiment that gave us the best results in Table III we obtained a substantial improvement: the DER was reduced from 1.9% to 1.0% and the standard deviation from 5.6% to 3.5%. We also got a very good result using channel factors as well as speaker factors, namely a DER of 1.1% with a standard deviation of 3.7%, but we observe again that there is no improvement over using speaker factors alone.

In speaker recognition, a diagonal term, $\boldsymbol{Dz}$, is also included in equation (1), [1]. We performed an experiment to see if this term would be helpful in speaker diarization. In order to include this term, it is necessary to associate with each speaker s a second type of posterior distribution $\boldsymbol{Q}(\boldsymbol{z}_s)$ which, unlike $\boldsymbol{Q}(\boldsymbol{y}_s)$, has a diagonal covariance matrix. This can be achieved by modifying [11] in a straightforward way. (This second type of posterior also arises in the Variational Bayes treatment of Joint Factor Analysis [13].) It turns out however that including the diagonal term $\boldsymbol{Dz}$ makes no difference whatever: we obtained a DER of 1.0% with a standard deviation of 3.5%, just as in the case where we used speaker factors alone to model inter-speaker variation.

V. Comparisons between systems

To compare the Baseline, Streaming and Variational Bayes systems, the best results from Tables I and II and Section IV-H are assembled in Table IV. As measured by DER, the Variational Bayes system is seen to have by far the best performance. Note that the result reported for the Streaming system was obtained with Viterbi re-segmentation; this ensures that the comparison with the other two systems is fair (although the Streaming system would probably have benefited from soft speaker clustering as well). To underscore the effectiveness of the Variational Bayes system we can add another result, namely DER = 2.3% and $\sigma = 3.9\%$, obtained in a single pass without Viterbi re-segmentation. This is seen to be better than both the Baseline and Streaming results.

TABLE IV
BEST RESULTS OBTAINED WITH ON THE NIST 2008 SUMMED CHANNEL TELEPHONE DATA WITH THE BASELINE, STREAMING AND VARIATIONAL BAYES (VB) SYSTEMS; σ REFERS TO THE STANDARD DEVIATION OF THE DIARIZATION ERRORS.

	mean DER (%)	σ (%)
Baseline with soft clustering	3.5	8.0
Streaming + Viterbi	4.6	8.8
VB with raw cepstra + Viterbi + 2^{nd} pass	**1.0**	**3.5**

In our second implementation, we also paid closer attention to the initial segmentation of each speech file in the first pass. We simply removed the silences and chopped the speech up into 1 second intervals in our first implementation but, for the second implementation, we used the information contained in the speech/silence boundary markers to segment the speech data, inserting additional potential speaker change points at speech/silence boundaries.

With these refinements we found that Viterbi re-segmentation was helpful only if it is applied in the first pass so we removed it from the second pass.

When we used the re-implemented Variational Bayes system to replicate the experiment that gave us the best results in Table III we obtained a substantial improvement: the DER was reduced from 1.9% to 1.0% and the standard deviation from 5.6% to 3.5%. We also got a very good result using channel factors as well as speaker factors, namely a DER of 1.1% with a standard deviation of 3.7%, but we observe again that there is no improvement over using speaker factors alone.

In speaker recognition, a diagonal term, $\boldsymbol{Dz}$, is also included in equation (1), [1]. We performed an experiment to see if this term would be helpful in speaker diarization. In order to include this term, it is necessary to associate with each speaker s a second type of posterior distribution $\boldsymbol{Q}(\boldsymbol{z}_s)$ which, unlike $\boldsymbol{Q}(\boldsymbol{y}_s)$, has a diagonal covariance matrix. This can be achieved by modifying [11] in a straightforward way. (This second type of posterior also arises in the Variational Bayes treatment of Joint Factor Analysis [13].) It turns out however that including the diagonal term $\boldsymbol{Dz}$ makes no difference whatever: we obtained a DER of 1.0% with a standard deviation of 3.5%, just as in the case where we used speaker factors alone to model inter-speaker variation.

V. Comparisons between systems

To compare the Baseline, Streaming and Variational Bayes systems, the best results from Tables I and II and Section IV-H are assembled in Table IV. As measured by DER, the Variational Bayes system is seen to have by far the best performance. Note that the result reported for the Streaming system was obtained with Viterbi re-segmentation; this ensures that the comparison with the other two systems is fair (although the Streaming system would probably have benefited from soft speaker clustering as well). To underscore the effectiveness of the Variational Bayes system we can add another result, namely DER = 2.3% and $\sigma = 3.9\%$, obtained in a single pass without Viterbi re-segmentation. This is seen to be better than both the Baseline and Streaming results.

TABLE IV
BEST RESULTS OBTAINED WITH ON THE NIST 2008 SUMMED CHANNEL TELEPHONE DATA WITH THE BASELINE, STREAMING AND VARIATIONAL BAYES (VB) SYSTEMS; σ REFERS TO THE STANDARD DEVIATION OF THE DIARIZATION ERRORS.

	mean DER (%)	σ (%)
Baseline with soft clustering	3.5	8.0
Streaming + Viterbi	4.6	8.8
VB with raw cepstra + Viterbi + 2^{nd} pass	**1.0**	**3.5**

The application of speaker diarization which is of greatest interest to the present authors is that it can serve as a pre-processing step for speaker detection in two-party telephone conversations. It turns out that, at least as far as the NIST 2008 summed channel data is concerned, speaker detection performance is relatively insensitive to diarization errors. Other authors have had similar experience [23].

To measure the effect of diarization errors on a speaker detection task, we used the diarization output in the recognition phase of one of the summed-channel telephone tasks from the 2008 SRE. In the *short2-summed* task, the speaker models are trained with a one-channel recording and tested with a summed channel conversation. (Thus diarization was performed at verification time but not at enrollment time.) The diarization output is used to split the test conversation into two speech files (presumably each from a single speaker). These files are scored separately and the maximum of the two scores is taken as the final detection score. A 512-Gaussian Factor Analysis speaker detection system developed by Loquendo and the Politecnico di Torino [29] was used to test all diarization systems. Results are reported in terms of the speaker verification equal error rate (EER).

In Figure 4 we show EER's for the 1conv-summed task for different configurations of the diarization systems that we developed. The end point DER values of 0% and 35% represent using reference diarization (with overlapped speech excluded) and no diarization, respectively. We see that there is some correlation between EER and DER, but this is relatively weak. It appears that systems with a DER of less than 10% produce EER's within about 1% of the "perfect" diarization. To sweep out one more point with a higher DER, we ran the Baseline system with no Viterbi re-segmentation (DER = 20%). While the EER did increase to 10.5%, it was still better than the EER obtained with no diarization, namely 14.1%.

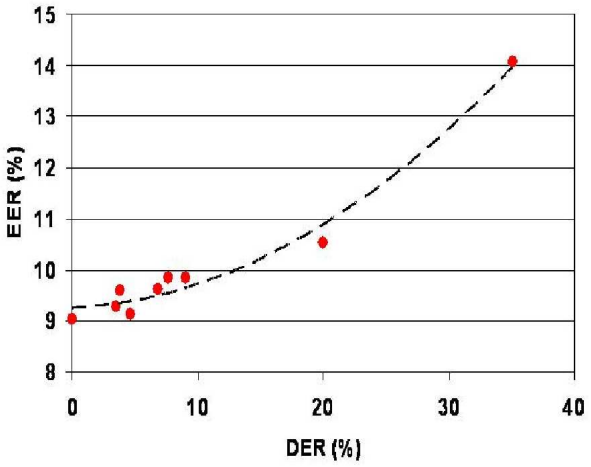

Fig. 4. EER vs DER for several diarization systems.

VI. Conclusions

In this article we have reported results obtained with three diarization systems developed during and after the 2008 Summer Workshop on Robust Speaker Recognition. While each of the systems had a different approach to speaker diarization, we found that ideas and techniques proved out in one system could be successfully applied in other systems. The Viterbi re-segmentation used in the Baseline system was found to be very useful in the other systems; the primary motivation for the Variational Bayes system was to exploit on a large scale the success demonstrated by the Streaming system in using speaker factors; and when the idea of soft speaker clustering from the Variational Bayes approach was incorporated into the Baseline system, an error rate reduction of almost 50% was achieved. The Variational Bayes system proved to be the most successful, achieving a DER of 1.0% on the NIST 2008 summed channel data. This represents an 85% error rate reduction compared with the agglomerative clustering Baseline system.

We examined the impact of using different diarization systems with varying DERs on a speaker recognition task using a standalone Factor Analysis speaker recognition system. We found that there was a weak correlation between speaker recognition errors and diarization errors, but the relationship between the two error rates was not as direct as we had expected. Our speaker recognition results on NIST 2008 summed channel data suggest that DER may not be a good criterion for optimizing a diarization system intended as a pre-processor for speaker detection.

Of course it may turn out in the long run that blind speaker diarization is a suboptimal strategy for speaker recognition applications and Variational Bayes allows other possibilities to be explored. In particular, a close integration of factor analysis based speaker recognition and diarization is feasible. For example, if enrollment data is available for a given speaker, that data can be used to initialize a speaker posterior before running Variational Bayes diarization on a 2-wire recording. This perspective would lead to the formulation of new decision criteria for speaker detection in multi-speaker environments.

A challenge of immediate interest would be to port the Variational Bayes system to the task of diarizing the interview recordings in the 2008 SRE. This domain presents several new challenges, including variable acoustics due to microphone type and placement as well as different speech styles and dynamics between interviewer and interviewee. As we discussed in Section IV-G, it is likely that further research will be needed to get the eigenchannel component of Variational Bayes speaker diarization working properly under these conditions.

VII. Acknowledgments

The authors would like to thank the Center for Spoken Language Processing at Johns Hopkins University for their hospitality and Pierre Ouellet, Ondrej Glembek, Lukas Burget, Pietro Laface and Ciprian Costin for their help with software development and data preparation.

Patrick Kenny was supported in part by the Natural Science and Engineering Research Council of Canada and by the Ministère du Développement Économique et Régional et de la Recherche du Gouvernement du Québec.

Douglas Reynolds was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions,

interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

REFERENCES

[1] 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 Lang. Process.*, vol. 16, no. 5, pp. 980–988, July 2008. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[2] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1969 – 1978, 2007.

[3] L. Burget, P. Matejka, O. Glembek, P. Schwarz, and J. Cernocky, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

[4] R. J. Vogt and S. Sridharan, "Explicit modeling of session variability for speaker verification," *Computer Speech and Language*, vol. 22, no. 1, pp. 17 – 38, 2008.

[5] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[6] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proc. ICASSP*, Las Vegas, Nevada, Mar. 2008, pp. 4133 – 4136.

[7] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, Eurecom, Sep 2005.

[8] Y. E. Kim, J. M. Walsh, and T. M. Doll, "Comparison of a joint iterative method for multiple speaker identification with sequential blind source separation and speaker identification," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008, pp. 283 – 286.

[9] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, LLC, 2006.

[10] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 695–707, Nov. 2000.

[11] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," 2008. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[12] V. Smidl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Berlin: Springer-Verlag, 2006.

[13] X. Zhao, Y. Dong, J. Zhao, L. Lu, J. Liu, and H. Wang, "Variational Bayesian joint factor analysis for speaker verification," in *Proc. ICASSP 2009*, Taipei, Taiwan, Apr. 2009.

[14] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13," 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[15] (2001) The NIST year 2001 speaker recognition evaluation plan. [Online]. Available: http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf

[16] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[17] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems 13*, Apr. 2001, pp. 758–764.

[18] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Advances in Neural Information Processing Systems 21*. Cambridge, MA: MIT Press, 2009.

[19] 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.

[20] D. MacKay, *Information theory, inference and learning algorithms*. New York, NY: Cambridge University Press, 2003.

[21] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, June 2001, pp. 213–218.

[22] 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.

[23] H. Aronowitz and Y. Solewicz, "Speaker recognition in two wire test sessions," in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 865 – 868.

[24] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1505–1512, Sept. 2006.

[25] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining Gaussianized/non-Gaussianized features to improve speaker diarization of telephone conversations," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1040 – 1043, Dec. 2007.

[26] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. Interspeech*, Brisbane, Australia, Sept. 2008, pp. 32–35.

[27] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 853 – 856.

[28] H. Aronowitz, "Trainable speaker diarization," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007, pp. 1861 – 1864. [Online]. Available: http://aronowitzh.googlepages.com/publicationlist

[29] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Loquendo – Politecnico di Torino's 2006 NIST speaker recognition evaluation system," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1238 – 1241.