Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

Speaker Diarization based on Bayesian HMM with Eigenvoice Priors

Mireia Diez, Lukáš Burget, Pavel Matějka

Brno University of Technology, Speech@FIT, Czechia

{mireia,burget,matejka}@fit.vutbr.cz

Abstract

Nowadays, most speaker diarization methods address the task in two steps: segmentation of the input conversation into (preferably) speaker homogeneous segments, and clustering. Generally, different models and techniques are used for the two steps. In this paper we present a very elegant approach where a straightforward and efficient Variational Bayes (VB) inference in a single probabilistic model addresses the complete SD problem. Our model is a Bayesian Hidden Markov Model, in which states represent speaker specific distributions and transitions between states represent speaker turns. As in the ivector or JFA models, speaker distributions are modeled by GMMs with parameters constrained by eigenvoice priors. This allows to robustly estimate the speaker models from very short speech segments. The model, which was released as open source code and has already been used by several labs, is fully described for the first time in this paper. We present results and the system is compared and combined with other state-of-the-art approaches. The model provides the best results reported so far on the CALLHOME dataset.

1. Introduction

Speaker Diarization (SD) is the task of determining speaker turns in an audio recording of a conversation or, as is it also commonly stated, finding "Who spoken when?". Speaker Diarization has been of interest for the research community since the late nineties, when the first works on speaker segmentation and clustering emerged [1, 2]. Nevertheless, SD has proven to be a complex task and, still nowadays, research seeks for systems that can be applied to real world scenarios.

In this paper, we present a Bayesian approach to SD, where the sequence of speech features representing a conversation is assumed to be generated from a Bayesian Hidden Markov Model (HMM). HMM states represent speakers and the transitions between the states correspond to speaker turns. The speaker (or HMM state) specific distributions are modeled by Gaussian Mixture Models (GMMs). In order to robustly learn the speaker specific distributions, a strong informative prior is imposed on the GMM parameters, which makes use of eigenvoices just like i-vectors [3] or Joint Factor Analysis (JFA) [4] – the standard techniques for speaker recognition. Such prior facilitates discrimination between speaker voices in an input recording. The proposed Bayesian model offers a very elegant approach to SD as a straightforward and efficient Variational Bayes (VB) inference in a single probabilistic model addresses the complete SD problem. The system contrasts with most of the conventional approaches, where different models, techniques and heuristics are used to address the individual subproblems of SD such as speaker turn detection, speaker clustering or determining the number of speakers in the conversation.

Early systems for Speaker Diarization [1, 2, 5, 6] proposed a step-by-step schema to address the SD task: First, the parts of the input conversation that are not of interest are removed (i.e. silence, music, overlapped speech). Speech regions are then divided into smaller segments with the aim of splitting the speech into (preferably) speaker homogeneous regions. These segments are then clustered together according to the speaker identity, while inferring the number of speakers in the conversation. Agglomerative Hierarchical Clustering (AHC) is by far the most common method for the clustering, while there is a wide range of modelings and stopping criteria used in the literature [7, 8]. Later works [9, 10], started performing a resegmentation step. This consists of first training speaker specific models (typically GMMs) from the obtained clusters, which are then used to refine the assignment of speech frames to speakers. Generally the reassignment is done by means of an ergodic HMM where the speaker models are used as the HMM state distributions, and Viterbi alignment is used to align frames to states. The resegmentation stages can be repeated iteratively to achieve best performance. Although, the models and techniques for the segmentation and clustering have evolved over time [7, 8, 11], the main approaches to SD still follow this general schema.

After the introduction of the resegmentation step, it was shown that the first segmentation can be addressed in a simpler way (e.g. uniformly splitting the input conversation into about 2 second fixed length segments) as the resegmentation would retrieve the missed speaker change points. Also, inspired by the success of i-vectors in speaker recognition [3], SD systems started using these low-dimensional representations of speech segments to facilitate the clustering step [12], which has become the standard practice. For example, the recent work [13] follows this schema: For each 2 second segment, an i-vector is extracted and the i-vectors are clustered using AHC. For the clustering, Probabilistic Linear Discriminant Analysis (PLDA) [14] is used to measure similarity between i-vectors, which is another standard technique borrowed from speaker recognition field [15]. This i-vector/PLDA-AHC based system will also serve as the baseline for our experiments.

The first VB approach to SD was proposed in [16, 17] and further extended in [18]. Our work, which is mainly inspired by [18], applies the same eigenvoice priors and similar VB inference, but incorporates HMMs to model speaker transitions. The resulting Bayesian HMM is similar to the *sticky HDP-HMM* presented in [19], except that we use a more practical setting with fixed number of HMM states and a more efficient VB inference. Moreover, [19] does not make use of the eigenvoice priors, which makes our model more robust.

The model presented in this work can be initialized by

The work was supported by the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 748097, and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

10.21437/Odyssey.2018-21

choosing the (maximum) number of speakers for the input conversation and using a random assignment of frames to speakers. Also, it can be initialized using labels obtained from an external diarization algorithm as in [20], where the previously mentioned i-vector/PLDA-AHC system output was used for the initialization. As it will be shown in the paper, the model is *good initialization hungry*, as better initializations will drive the algorithm into better solutions (i.e. avoiding local optima).

An open source Python implementation of our SD approach is available at [21]. It has been already used by different research labs in several published works [22, 20, 23]. However, the model has not been yet sufficiently described in any publication. In this paper, we give the full description of the model and provide insights on the inference in the model. Experiments are then carried out showing the effectiveness of the technique.

2. The model

Our model assumes that the sequence of observed speech features (e.g. Mel Frequency Cepstral Coefficients (MFCCs)) corresponding to an input conversation is generated from an HMM with speaker specific state distributions. The distribution of each speaker is modeled using a GMM with parameters constrained to live in the eigenvoice subspace (see section 2.2 for more details). This allows us to robustly model the distribution of speaker s using only a low dimensional vector $\mathbf{y}_s$. Our model does not consider any overlapped speech as each speech frame is assumed to be generated from one of the M HMM states corresponding to only one of the S speakers. In the simple case, we use an ergodic HMM with one-to-one correspondence between the HMM states and the speakers (i.e. $M = S$), where transitions from any state to any state are possible. However, the transition probabilities are set in a way that discourages too frequent transitions between speakers in order to reflect speaker turns duration of a natural conversation. More details on setting and learning the transition probabilities can be found in section 2.1, which also introduces slightly more complex HMM topology, where linear chains of HMM states are used for each speaker to impose minimum duration constraints on the speaker turns. In such case, multiple states correspond to one speaker, all of which, however, share the same speaker specific GMM.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ be the sequence of the observed feature vectors and $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T\}$ the corresponding sequence of latent variables defining the hard alignment of speech frames to HMM states. In our notation, $\mathbf{z}_t$ is the one-hot encoding vector (i.e. $z_{tm} = 1$ indicates that the mth HMM state is responsible for generating observation $\mathbf{x}_t$; otherwise $z_{tm} = 0$). For the notational convenience, we also define one-hot vectors $\bar{\mathbf{z}}_t$ indicating the speakers responsible for generating the observations. Note that vectors $\bar{\mathbf{z}}_t$ are fully defined in terms of the latent variables $z_{tm} = 1$ (i.e. if $z_{tm} = 1$ for any state m corresponding to speaker s then $\bar{z}_{ts} = 1$; otherwise $\bar{z}_{ts} = 0$). In the simple case of the one-to-one correspondence between the HMM states and the speakers $\bar{\mathbf{z}}_t = \mathbf{z}_t$. Note that, in our experiments, such simpler setting was found sufficient to obtain the best results. Therefore, the reader might consider such case during the first reading of the paper and disregard the difference between $\bar{\mathbf{z}}_t$ and $\mathbf{z}_t$.

To address the SD task using our model, the speaker distributions (i.e. the vectors $\mathbf{y}_s$) and the latent variables $\mathbf{z}_t$ are jointly estimated given an input sequence $\mathbf{X}$. The solution to the SD task is then given by the most likely sequence $\mathbf{Z}$, which encodes the alignment of speech frames to speakers.

2.1. HMM topology

Now, we describe our HMM topology and the setting of the transition probabilities, which model the speaker turn durations. In our model, there can be multiple HMM states per speaker, all of which share the same speaker specific distribution. HMM states corresponding to one speaker form a forward linear chain to impose a constraint on minimum speaker turn duration. Figure 1 shows an example of such topology for only $S = 2$ speakers and $D = 3$ states per speaker (i.e. the overall number of HMM states in this case is $M = SD = 6$ and the minimum speaker turn duration is $D = 3$). Each row of states in the figure is the linear chain corresponding to one speaker. We have chosen to set the transition probabilities as follows: For any but the last speaker's state, we transition forward to the next state in the chain with probability one. In the last state, we stay again in the same state with probability P_{loop}. This probability is one of the tunable parameters in our models and will be typically set to high value to discourage frequent speaker turns. The remaining probability $1 - P_{\text{loop}}$ is the probability of leaving the last state and entering any of the first states of any speaker (i.e. probability of changing speaker)[1]. When changing the speaker, the probability of choosing the speaker s as the new speaker is π_s. Therefore, the joint probability of leaving a last state and entering the first state of speaker s is $(1 - P_{\text{loop}})\pi_s$ as depicted on the corresponding transitions in Figure 1. The probabilities π_s also control the selection of the initial HMM state (i.e. the state generating the first observation) as depicted in the figure by arrows entering from the left to the first states of speaker chains.

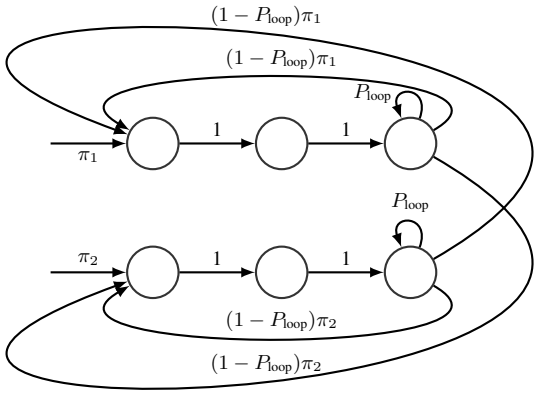

Figure 1: HMM model for 2 speakers with 3 states per speaker.

The probabilities π_s are inferred (jointly with the variables $\mathbf{y}_s$ and $\mathbf{z}_t$) from the input conversation. Thanks to the Automatic Relevance Determination (ARD) principle [24] stemming from our Bayesian model, zero probabilities will be learned for the π_s corresponding to redundant speakers, which effectively drops such speakers from the HMM model. Typically, we initialize the HMM with larger number of speakers and we make use of this behavior to learn the right[2] number of speakers in the conversation.

[1] For convenience, we allow to re-enter the same speaker as it leads to simpler update formulas.

[2] When saying "right number of speakers", we mean the speakers that are not considered redundant by our approximate Bayesian model, which does not necessarily have to be the correct number of speakers in the conversation.

In our experiments, we have found sufficient to use only a single HMM state per speaker provided that the speaker turn duration is properly modeled by other means (sufficiently large P_{loop}, *downsamplingFactor* described in section 3.3). This can be seen just as a special case of the HMM topology described above, where the only state of each speaker is at the same time the first and last state in the "speaker's chain". Note that, in this case, the probability of staying in the same state of speakers s is $P_{\text{loop}} + (1 - P_{\text{loop}})\pi_s$, which, in Figure 1, corresponds to probability of looping in the last state plus probability of leaving the last state and re-entering again the same speaker.

2.2. Speaker specific distributions

For each speaker, the distribution of speech features is modelled using a GMM. Like similar models for speaker recognition (e.g. i-vectors extraction [3] or JFA [4]), our model assumes that the speaker specific GMMs are all related to a single Universal Background Model (UBM-GMM). The UBM-GMM is an ordinary GMM typically trained on large amount of speech data from many speakers. All speaker specific GMMs have the same number of Gaussian components C as the UBM-GMM. Furthermore, there is a one-to-one correspondence between the components of the UBM-GMM and the components of each speaker model. All speaker specific GMMs share the same component weights w_c^{ubm} and covariance matrices $\boldsymbol{\Sigma}_c^{ubm}$, which are copied from the corresponding UBM-GMM components $c = 1..C$. Only the component mean vectors $\boldsymbol{\mu}_{sc}$ take speaker specific values, which are however still constrained as follows: Let $\boldsymbol{\mu}_s = [\boldsymbol{\mu}_{s1}^T \boldsymbol{\mu}_{s2}^T \dots \boldsymbol{\mu}_{sC}^T]^T$ be the super-vector of concatenated Gaussian component means for speaker s and let $\boldsymbol{\mu}^{ubm}$ be the similarly defined super-vector of concatenated UBM-GMM means. The high-dimensional super-vectors

$$\boldsymbol{\mu}_s = \boldsymbol{\mu}^{ubm} + \mathbf{V}\mathbf{y}_s \tag{1}$$

are constrained to live in a low-dimensional subspace around the origin given by $\boldsymbol{\mu}^{ubm}$. The subspace is spanned by the so-called eigenvoice basis, which are the columns of the low-rank matrix $\mathbf{V}$. This matrix is also shared by all speaker models. The only speaker specific parameters are then the low-dimensional vectors $\mathbf{y}_s$, which can be seen as coordinates of $\boldsymbol{\mu}_s$ in the low-dimensional subspace. All the speaker independent parameters $\boldsymbol{\mu}^{ubm}$, $\boldsymbol{\Sigma}_c^{ubm}$, w_c^{ubm} and $\mathbf{V}$ are pre-trained and fixed during the inference in our model when addressing the SD task. Therefore, the speaker specific distributions

$$p(\mathbf{x}_t|\mathbf{y}_s) = \text{GMM}(\mathbf{x}_t; \{\boldsymbol{\mu}_{sc}\}, \{\boldsymbol{\Sigma}_c^{ubm}\}, \{w_c^{ubm}\}) \tag{2}$$

can be expressed only in terms of the low-dimensional vectors $\mathbf{y}_s$, which can be robustly estimated from the limited amount of speech available in the input conversation.

To further improve robustness of the speaker model estimates, we treat $\mathbf{y}_s$ as a latent variable with standard normal prior

$$p(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s; \mathbf{0}, \mathbf{I}). \tag{3}$$

Inserting such prior into (1) translates to Gaussian prior imposed on speaker mean super-vectors

$$p(\boldsymbol{\mu}_s) = \mathcal{N}(\boldsymbol{\mu}_s; \boldsymbol{\mu}^{ubm}, \mathbf{V}\mathbf{V}^T), \tag{4}$$

which can be also seen as an informative prior on the possible speaker GMMs. To obtain such prior that correctly models the variability of the speaker mean super-vectors, the matrix $\mathbf{V}$ also needs to be pre-trained on speech data from a large number of speakers. Note, that the model for representing speaker specific distributions described above is essentially the same as the model for i-vector extraction [3] or JFA model [4]. Therefore, we do not provide a detailed description of the procedure for training $\mathbf{V}$ in this paper and we kindly refer the reader to the original sources. In our experiments, we train $\mathbf{V}$ using exactly the same procedure (EM algorithm) and the same code that we normally use to train the total variability matrix for i-vector extraction in the speaker recognition task. Note that, in this case, the resulting matrix $\mathbf{V}$ does not only model the between-speaker variability but also other inter-session (e.g. channel) variability, which, however, turned out to be helpful in our experiments as the channel variability can facilitate discrimination between speakers in a conversation.

2.3. Bayesian HMM

To summarize, our complete model for SD is a Bayesian HMM, which is defined in terms of transition probabilities $P(\mathbf{z}_t|\mathbf{z}_{t-1})$ and the state specific distributions (or so-called emission probabilities) $p(\mathbf{x}_t|\mathbf{z}_t)$. The transition probabilities are set as described in section 2.1). By abuse of notation, $P(\mathbf{z}_1|\mathbf{z}_0)$ will correspond to the initial state probability $P(\mathbf{z}_1)$ in the following formulas. The state distributions are

$$p(\mathbf{x}_t|\mathbf{z}_t) = \prod_s p(\mathbf{x}_t|\mathbf{y}_s)^{\bar{z}_{ts}}, \tag{5}$$

where the speaker indicator variable $\bar{z}_{ts}$ selects the correct speaker specific distribution (2) corresponding to the HMM state $\mathbf{z}_t$.

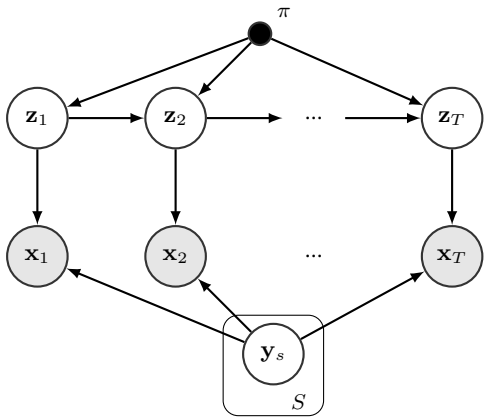

Figure 2: Directed graph of the Bayesian model used in the approach.

The Bayesian Network corresponding to our Bayesian HMM is depicted in Figure 2. The model assumes that the feature sequence corresponding to an input conversation is generated as follows:

for $s = 1..S$ **do**
  $\mathbf{y}_s \sim \mathcal{N}(0, \mathbf{I})$
  $\boldsymbol{\mu}_s = \boldsymbol{\mu}^{ubm} + \mathbf{V}\mathbf{y}_s$
for $t = 1..T$ **do**
  $\mathbf{z}_t \sim p(\mathbf{z}_t|\mathbf{z}_{t-1})$
  $\mathbf{x}_t \sim P(\mathbf{x}_t|\mathbf{z}_t)$

Here, a speaker specific GMM distribution is first sampled for each speaker s. This is achieved by sampling the low dimensional speaker vectors $\mathbf{y}_s$ from the standard normal prior and

then applying (1) to obtain the corresponding GMM means. Recall that the other GMM parameters are pre-trained and shared by all speaker models. Once the speaker models are generated for the conversation, the initial HMM state is selected according to the distribution $P(\mathbf{z}_1) = P(\mathbf{z}_1|\mathbf{z}_0)$ (i.e. as described in Section 2.1, one of the first states in the speaker chains is selected according to probabilities π_s). Given the selected state $\mathbf{z}_1$, the first observation $\mathbf{x}_1$ is samples from distribution $p(\mathbf{x}_1|\mathbf{z}_1)$ (i.e. from the speaker specific GMM corresponding to the state $\mathbf{z}_1$). Then, for each frame t, new HMM state is select according to $P(\mathbf{z}_t|\mathbf{z}_{t-1})$ and new observation $\mathbf{x}_t$ is sampled from $P(\mathbf{x}_t|\mathbf{z}_t)$.

We call our model "Bayesian" HMM as we impose a prior on the parameters of the state distributions (i.e. $\mathbf{y}_s$ is a latent variable with standard normal prior). However, unlike other "Fully Bayesian" HMM implementations [19, 25] we do not impose any prior on the transition probabilities.

Further note that, although our state distributions are GMMs, the Bayesian network in Figure 2 does not introduce any latent variables defining the alignments of observations to the Gaussian components. We assume that this alignment is exactly the same for all the speaker specific GMMs and UBM-GMM. This is possible thanks to the correspondence between the Gaussian components in these models. Therefore, we precalculate the alignments using the UBM-GMM and consider them observed during the inference in our model. More precisely, we calculate soft alignments (or responsibilities) as the posterior probabilities of UBM-GMM components given an observation $p_{ubm}(c|\mathbf{x}_t)$. Note that such approximation, which considerably simplifies the inference in the model, is also used in the similar models for speaker recognition [3, 4].

3. Inference

In this section, we will describe the inference in our model addressing the speaker diarization task. We give all the formulas necessary for implementing the described SD method or for understanding the implementation available at [21]. We only sketch the derivations of the update formulas and other quantities. For their full derivation, we kindly refer the reader to the supplementary material [26].

3.1. Definition of useful quantities

Let us first define some useful quantities. Using the UBM-GMM we collect per-frame zero, first and second order statistics:

$$\begin{aligned}\zeta_{tc} =& p_{ubm}(c|\mathbf{x}_t)\\ \boldsymbol{\rho}_t =& \sum_c \zeta_{tc}\left(\mathbf{x}_t - \boldsymbol{\mu}_c^{ubm}\right)^T {\boldsymbol{\Sigma}_c^{ubm}}^{-1}\mathbf{V}_c\\ \boldsymbol{\phi}_t =& \sum_c \zeta_{tc}\mathbf{V}_c^T{\boldsymbol{\Sigma}_c^{ubm}}^{-1}\mathbf{V}_c\end{aligned} \tag{6}$$

Then, assuming the fixed alignment of frames to Gaussian components given by the UBM responsibilities ζ_{tc} the speaker specific distributions can be evaluated in terms of the sufficient statistics as:

$$\ln p(\mathbf{x}_t|\mathbf{y_s}) = G(\mathbf{x}_t) + \boldsymbol{\rho}_t\mathbf{y}_s - \frac{1}{2}tr\left(\boldsymbol{\phi}_t\mathbf{y}_s\mathbf{y}_s^T\right), \tag{7}$$

where the speaker independent constant

$$\begin{aligned}G(\mathbf{x}_t) =& -\frac{D}{2}\ln 2\pi - \sum_c \frac{\zeta_{tc}}{2}\ln|\boldsymbol{\Sigma}_c^{ubm}| + \sum_c \zeta_{tc}\ln\frac{w_c^{ubm}}{\zeta_{tc}}\\ &- \frac{1}{2}\sum_c \zeta_{tc}\left(\mathbf{x}_t - \boldsymbol{\mu}_c^{ubm}\right)^T {\boldsymbol{\Sigma}_c^{ubm}}^{-1}\left(\mathbf{x}_t - \boldsymbol{\mu}_c^{ubm}\right).\end{aligned} \tag{8}$$

Equation (7) is only an approximation (lower bound) to the true speaker PDF. The value would be exact only if the responsibilities estimated with the speaker models were the same as the responsibilities ζ_{tc} obtained with UBM-GMM (see [26] for full derivation of equation (7)).

For the following inference, we need to define the joint probability distribution of all the random variables

$$\begin{aligned}&\ln p(\mathbf{X},\mathbf{Z},\mathbf{Y}) = \ln p(\mathbf{X}|\mathbf{Z},\mathbf{Y}) + \ln p(\mathbf{Z}) + \ln p(\mathbf{Y}) \qquad (9)\\ &= \sum_t\sum_s \ln p(\mathbf{x}_t|\mathbf{y}_s)^{\bar{z}_{ts}} + \sum_t \ln p(\mathbf{z}_t|\mathbf{z}_{t-1}) + \sum_s \ln p(\mathbf{y}_s)\end{aligned}$$

where $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_S\}$ is the set of all the speaker specific latent variables.

3.2. Variational Bayes inference

The diarization problem consists in finding the assignment of frames to speakers, which is represented by the latent sequence $\mathbf{Z}$. In order to find the most likely sequence $\mathbf{Z}$, we need to infer the posterior distribution $p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z},\mathbf{Y}|\mathbf{X})d\mathbf{Y}$. Unfortunately, the evaluation of this integral is intractable, and therefore, we will approximate it using Variational Bayes inference [24] where the distribution $p(\mathbf{Z},\mathbf{Y}|\mathbf{X})$ is approximated by $q(\mathbf{Z},\mathbf{Y})$. We search for such $q(\mathbf{Z},\mathbf{Y})$ that minimizes the Kullback-Liebler divergence $D_{KL}(q(\mathbf{Z},\mathbf{Y})||p(\mathbf{Z},\mathbf{Y}|\mathbf{X}))$. We use the mean-field approximation [24, 18] assuming that the approximate posterior distribution factorizes as

$$q(\mathbf{Z},\mathbf{Y}) = q(\mathbf{Z})q(\mathbf{Y}) = \prod_t p(\mathbf{z}_t|\mathbf{z}_{t-1})\prod_s p(\mathbf{y}_s). \tag{10}$$

With the mean-field approximation, the VB inference dictates that the distributions of the latent variables $q(\mathbf{Y})$ and $q(\mathbf{Z})$ need to be iteratively updated according to the formulas [24, 18]

$$\ln q(\mathbf{Y}) = E_Z[\ln p(\mathbf{X},\mathbf{Z},\mathbf{Y})] + \text{const} \tag{11}$$

$$\ln q(\mathbf{Z}) = E_Y[\ln p(\mathbf{X},\mathbf{Z},\mathbf{Y})] + \text{const}, \tag{12}$$

which guarantee to improve the VB objective (18).

By substituting equation (9) into (11) and solving the expectation w.r.t. the current $q(\mathbf{Z})$, we obtain the approximate posterior distribution for speaker latent variables

$$q(\mathbf{y}_s) = \mathcal{N}\left(\mathbf{y}_s|\boldsymbol{\alpha_s}, \mathbf{L}_s^{-1}\right) \tag{13}$$

with mean and precision

$$\boldsymbol{\alpha}_s = \mathbf{L}_s^{-1}\sum_t \gamma_{ts}\boldsymbol{\rho}_t \qquad \mathbf{L}_s = \mathbf{I} + \sum_t \gamma_{ts}\boldsymbol{\phi}_t, \tag{14}$$

where γ_{ts} are the responsibilities defining soft alignments of speech frames to speakers. That is, γ_{ts} are the posterior probabilities that frame t was generated by the speaker s, which are given by the marginal approximate posterior $q(z_{ts})$ derived from the current distribution $q(\mathbf{Z})$. The details on estimating these values are given below.

Note that equation (14) corresponds to the standard formulas for i-vector extraction, except for the responsibility term γ_{ts}, which is not present for i-vectors[3]. Furthermore, i-vectors are

[3]The standard speaker verification task assumes that the whole recording comes from a single speaker.

only MAP point estimates of the latent variable (i.e. the means $\boldsymbol{\alpha}_s$), whereas our approach considers the whole posterior distributions (including the precisions $\mathbf{L}_s$) with the aim of accounting for the uncertainty in the speaker model estimates.

To update the latent approximate posterior $q(\mathbf{Z})$, we similarly substitute equation (9) into (12) and solve the expectation w.r.t. the current distribution of speaker variables $q(\mathbf{Y})$, which results in

$$q(\mathbf{Z}) \propto \prod_t \overline{p}(\mathbf{x}_t|\mathbf{z}_t) + \prod_t p(\mathbf{z}_t|\mathbf{z}_{t-1}) \tag{15}$$

where

$$\ln \overline{p}(\mathbf{x}_t|\mathbf{z}_t) = \tag{16}$$
$$\sum_s \bar{z}_{ts} \left[\boldsymbol{\rho}_t \boldsymbol{\alpha}_s - \frac{1}{2} tr\left(\boldsymbol{\phi}_t \left[\mathbf{L}_s^{-1} + \boldsymbol{\alpha}_s \boldsymbol{\alpha}_s^T \right] \right) + G(x_t) \right]$$

Notice that equation (15) has exactly the same form as the posterior probability of the latent sequence $p(\mathbf{Z}|\mathbf{X})$ for in the standard (non Bayesian) HMM except that the standard emission probability $p(\mathbf{x}_t|\mathbf{z}_t)$ is replaced by $\overline{p}(\mathbf{x}_t|\mathbf{z}_t)$ from equation (16). Further note that we do not need to infer the full posterior distribution over all the possible latent sequences $q(\mathbf{Z})$, as only the responsibilities γ_{ts} are necessary in the update equations (14). This responsibilities can be calculated using the standard forward-backward algorithm (equation (13.33) in [24]), with the only difference that in the recursions for calculating the forward and backward probabilities (equations (13.36) and (13.38) in [24]) we substitute the term $p(\mathbf{x}_t|\mathbf{z}_t)$ with $\overline{p}(\mathbf{x}_t|\mathbf{z}_t)$. Similarly we can calculate $\xi_{t,m,n}$, the approximate marginal probabilities of transitioning from state m to state n at time t (equation (13.43) in [24]), which is used in the updates of speaker priors π_s (see equation (19)).

To monitor the convergence, we use the usual VB objective, the Evidence Lower Bound (ELBO)

$$\mathcal{L} = E_{\mathbf{Y},\mathbf{Z}} \left\{ \ln \left(\frac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{q(\mathbf{Y},\mathbf{Z})} \right) \right\} \leq \ln p(\mathbf{X}), \tag{17}$$

which can be efficiently evaluated (see page 95 of [25]) as

$$\mathcal{L} = \ln \overline{p}(\mathbf{X}) + \sum_s \frac{1}{2} \left(R + \ln |\mathbf{L_s^{-1}}| - tr(\mathbf{L}_s^{-1}) - \boldsymbol{\alpha}_s^T \boldsymbol{\alpha}_s \right) \tag{18}$$

where $\overline{p}(\mathbf{X})$ is the total forward probability calculated during the forward-backward algorithm (equation (13.42) in [24]) and R is the rank of the subspace eigenvoice matrix $\mathbf{V}$ (see [26] for the full derivation of the ELBO objective). Note, however, that such ELBO evaluation is valid only right after the update of the $q(\mathbf{Z})$ distribution.

Finally, the speaker priors π_s are updated as ML II estimates: We search for the values of π_s that maximize the the lower bound (17), which gives the following update formula

$$\pi_s \propto \gamma_{1s} + \sum_m \sum_t \xi_{t,m,init(s)}. \tag{19}$$

where $init(s)$ is the initial state in the linear chain of speaker s. Note that updating the priors modifies the transition probabilities in our HMM model as described in section 2.1. Because of the ARD principle, these updates tend to drive the π_s values of any redundant speaker to zero values, which effectively drops the speaker from the model and selects the right number of speakers in the input conversation.

The complete iterative VB inference for a single input conversation consists of the following steps:

initialize γ_{ts} either randomly, or from the output of an external diarization system (see section 4.1).
repeat
- update the speaker models $\mathbf{y}_s$ using eq. (14)
- update the responsibilities γ_{ts} using the forward-backward algorithm with emission probabilities calculated as eq. (16)
- update speaker priors π_s using eq. (19)

until the convergence of $\mathcal{L}$, eq. (18)

3.3. Tunable model parameters

The VB inference can be controlled by the following tunable parameters: In section 2.1, we have already described the probability of staying in the last speaker state P_{loop} and the the minimum duration constraint D, which can be seen as parameters of the speaker turn duration model.

The wrong assumption of statistical independence between observations made by HMM results to overconfident posterior distributions of the latent variables. To counteract this problem, we scale all the sufficient statistics in (6) by the factor $statScale$. This factor is typically set to a value between 0 and 1 in order to effectively reduce the number of observations, which makes the model believe that there is less evidence in the data for estimating the posterior distributions.

Another parameter in our model, which controls the speaker turn duration, is the positive integer valued $downsampligFactor$. Formally, we assume a modified HMM generative process where $downsampligFactor$ observations are generated at once from the current HMM state in each step (i.e. after each transition). To reflect this model modification in the VB inference, we simply accumulate the per-frame statistics (6) for each $downsampligFactor$ consecutive frames, which effectively reduces the frame rate of the statistics by this factor. This modification can significantly speed up the VB inference for the price of the reduced frame resolution leading to a coarser granularity of the output labeling. However, the reduced frame rate does not necessarily have to be seen as a disadvantage. In fact, it can help to improve modeling of speaker turn duration. With a single HMM state per speaker, HMM assumes geometrically distributed speaker turn durations. In the case 10 ms frame rate, as used for our MFCC features, such duration model does not reflect the reality very well. As pointed out in [19], however, for reduced frame rates (e.g. 250 ms corresponding to $downsampligFactor = 25$ used in our experimets), the geometric distribution becomes quite good match for the speaker turn duration modeling. Indeed, we have found $downsampligFactor = 25$ to be a good setting when tuning all the parameters for the best SD performance.

3.4. Complexity

To comment on the computational complexity of the inference in our model, we compare it to the the popular SD approach serving as our baseline [13], where i-vectors are extracted for short overlapping segments of fixed length (e.g. 2 seconds) and clustered using AHC. Here, the number of i-vectors that

needs to be extracted grows linearly with the length of the input conversation, which is typically hundreds of i-vectors for few minutes of speech. As described in section 3.2, the inference in our model iterates between the updates of the speaker models and update of the assignment of frames to speakers. The speaker model updates have essentially the same computational complexity as the i-vector extraction and the models need to be updated once per VB iteration. On the other hand, we update only a fixed number of speaker models (i.e. the maximum assumed number of speakers, which is 10 in our experiments) and the VB inference typically converges in only few (less than 10) iterations. The update of responsibilities γ_{ts}, based on the forward-backward algorithm, has similar computational complexity as the standard Viterbi re-segmentation using an HMM.

4. Experimental setup

4.1. System description

The features used in our experiments are standard 19 MFCCs plus energy, with no deltas. No mean nor variance normalization are applied in the feature extraction as these methods were found to harm the diarization performance in our experiments where channel information can help to discriminate between speakers. Our system employs gender independent UBM-GMM, with 1024 diagonal-covariance Gaussian components. The dimensionality of the speaker latent variable $\mathbf{y}_s$ is 400.

For the i-vector/PLDA-AHC system, which serves as a baseline, and which we also use to initialize the VB inference, we followed the configuration employed in [13]: 64 dimensional i-vectors are projected by means of PCA to 3 dimensions [27], and clustered using calibrated PLDA similarity score [15, 28].

To start the VB inference, we need some initial setting of the responsibilities γ_{ts} defining the soft alignments of speech frames to speakers. To initialize the responsibilities, we start from hard speaker labels, which can be obtained from an external diarization system or randomly. In the latter case, we choose the (maximum) number of speakers in the input conversation (10 speakers in our experiments) and randomly assign frames to the speakers. For each frame, the hard speaker label is then converted into the soft responsibilities by giving the selected speaker only a slightly higher probability than to the rest of the speakers.

The parameter controlling the VB inference were tuned for the best SD performance and set to the following values: $downsampligFactor = 25$, $P_{\text{loop}} = 0.9$, $statScale = 0.2$ and $D = 1$.

4.2. Datasets

Our experiments are evaluated on the NIST SRE 2000 CALLHOME dataset [29], consisting of 500 recordings of conversational telephone speech. The number of speakers per recording ranges between 2 and 7, although 87% of the files contain only 2 or 3 speakers.

To train the UBM-GMMs and ivector extractors (and PLDA model used for the baseline system), we use NIST SRE 2004-2008 datasets as in [13].

4.3. Evaluation metric

Diarization Error Rate (DER) as defined by NIST [30] is used to evaluate the system. As in most diarization works, we apply the standard 250 ms forgiveness collar around speaker change points. Also, as is the common practice, we use the oracle speech activity labels so that only speaker errors are accounted for in the DER (missed speech and false alarm speech errors are not taken into account), and overlapped speech is not evaluated.

5. Results

Let us provide a short overview of results for CALLHOME attained by the best performing systems found in the literature.

System	DER
VB with eigenvoice priors[18]	17.0
Speaker Factors [12]	13.7
VB-GMM [27]	14.5
Mean shift [31]	12.4
i-vector/PLDA-AHC [20]	13.7
DNN embeddings [23]	9.9

Table 1: DER for different speaker diarization approaches on the CALLHOME dataset.

The first line in Table 1 shows results obtained with our re-implementation of [18], which was not evaluated on CALLHOME in the original work. This model can be seen as a special case of our Bayesian HMM that transitions between speaker in exactly two second intervals. Note, however, that [18] reported results also with a final resegmentation step, which is not included in our case.

	DER	
Initialization	Only init.	After VB
Random init.	-	12.0
Random init. x5	-	9.0
i-vector/PLDA-AHC	13.7	9.7
Oracle labels	0	4.0

Table 2: DER attained with different initializations, before and after using VB inference, for the CALLHOME dataset

Table 2 shows the performance of the proposed diarization system when using different initializations. In all the cases, the same parameters and settings were used. The first line shows the result for the random initialization of the speaker responsibilities γ_{ts} described in section 4.1.

Note that, although the VB inference is deterministic when starting from given initial responsibilities, different random initializations of the responsibilities can lead to different solutions as the VB inference is not guaranteed to find the global optimum. Therefore, for the second line (Random init. x5), we repeat the VB inference for each input conversation 5 times[4], each time with different random initialization. At the end, the solution with the highest ELBO objective is selected. As can be seen, this strategy results to a significant 3% DER improvement as compared to the single initialization from the previous line. With 9% DER, it also outperforms all the previously published results from Table 1.

[4]We have observed that more than 5 random initializations do not lead to further significant improvements.

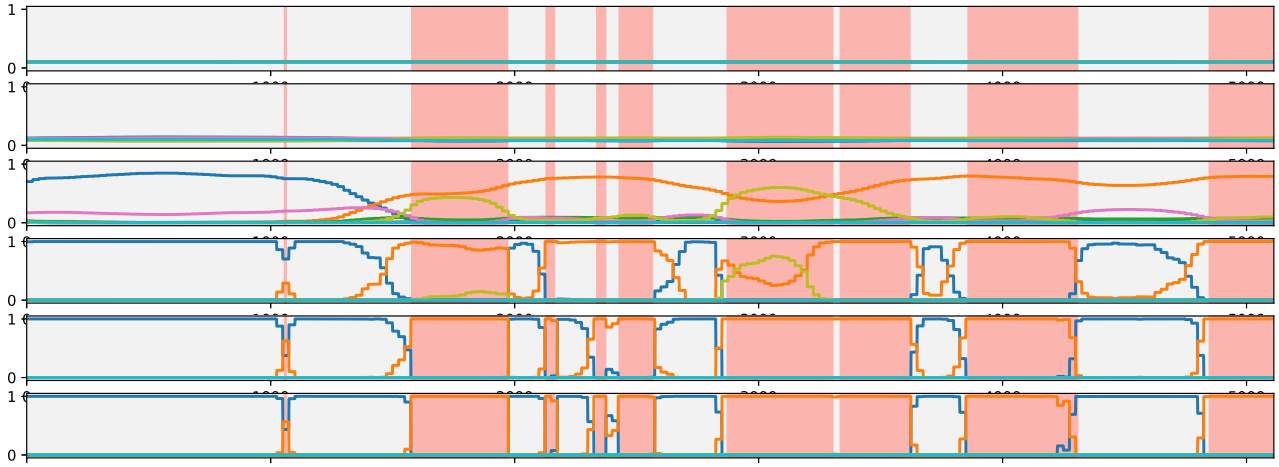

Figure 3: Convergence of the VB inference for a CALLHOME input conversation with 2 speakers.

The third line of Table 2 shows the results for the baseline i-vector/PLDA-AHC system (13.7% DER) and also for the proposed Bayesian HMM initialized from the i-vector/PLDA-AHC output labeling (9.7% DER). As can be seen, initializing the VB inference using the i-vector/PLDA-AHC output provides better performance than using the (single) random initialization, and it again outperforms all the previously published SD approaches. Still, the more computationally expensive 5 times repeated randomly initialized VB inference (Random init. x5) provides somewhat better performance.

Finally, the last line of Table 2 shows the result obtained when initializing the model with the oracle labels[5] to show how the inference diverges from the zero error solution. This value serves as a lower bound on the diarization performance that can be attained with the current version of the model.

5.1. Convergence of the algorithm

In Figure 3, we show an example of the algorithm convergence for a single conversation. The routine to make this plots is included in the implementation of the algorithm [21]. Each row corresponds to one iteration of the VB inference. Each row shows the whole conversation. The two different background colors represent the true labels for the two speaker present in the conversation. The colored lines show the frame-by-frame speaker responsibilities γ_{ts} as hypothesized for our speaker models in the individual VB iterations. As described in section 4.1, the initial responsibilities for all frames and all 10 initial speakers are set to almost the same value. In a few VB iterations, most of the speakers are dropped (i.e. their responsibilities for all frames have values close to 0) and the responsibilities for the remaining two speakers correctly separate the frames of the two speakers in the conversation. We can also see that very short speech segments are not always correctly attributed to the right speaker. Still, most of these segments lay in the forgiveness collar considered for measuring the DER, and therefore cause no performance penalization.

[5]Note that the algorithm could be tuned to make the inference rely more on the initialization and therefore the method could still achieve better results.

6. Conclusions

In this paper, we have presented the Bayesian Hidden Markov Model with Eigenvoice priors as a new probabilistic model for Speaker Diarization. Although the model has been already used in works by other authors [22, 20, 23], it has never been properly documented. The intention of the paper is to make this SD approach and its open source implementation [21] more accessible to the research community. The paper provides the necessary insights and give the full description of the model and of the VB inference in this model.

With a proper initialization, our model outperforms the previously published SD approaches on the CALLHOME dataset. In this work, we only show initial results with parameters tuned for the optimal performance. In our future work, we will provide more thorough analysis of different parameter settings, full derivation of the inference equations, description of the model variants and results on other datasets.

7. Acknowlegements

We would like to thank Daniel Garcia-Romero and Greg Sell for providing us the diarization outputs obtained with their systems and sharing details on their approaches.

8. References

[1] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[2] S. E. Johnson, "Who spoke when? - automatic segmentation and clustering for determining speaker turns," *PROC. EUROSPEECH*, vol. 5, pp. 2211–2214, 1999.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions Audio Speech Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[5] Hubert Jin, Francis Kubala, and Rich Schwartz, "Automatic speaker clustering," 1997.

[6] Scott Shaobing Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," 1998, pp. 127–132.

[7] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept 2006.

[8] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb 2012.

[9] Jean luc Gauvain, Lori Lamel, and Gilles Adda, "Partitioning and transcription of broadcast news data," in *ICSLP'98*, 1998, pp. 1335–1338.

[10] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Lluc Gauvain, "Improving speaker diarization," in *IN PROC. FALL 2004 RICH TRANSCRIPTION WORKSHOP (RT-04*, 2004.

[11] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-François Bonastre, and Laurent Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2, pp. 303 – 330, 2006, Odyssey 2004: The speaker and Language Recognition Workshop.

[12] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4133–4136.

[13] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 413–417.

[14] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[15] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," June 2010.

[16] F. Valente, P. Motlicek, and D. Vijayasenan, "Variational bayesian speaker diarization of meeting recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Con- ference on*, 2010, p. 4954 –4957.

[17] Fabio Valente, *Variational Bayesian methods for audio indexing*, Ph.D. thesis, Thesis, 09 2005.

[18] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," Tech. Rep., Montreal: CRIM, 2008.

[19] Emily B Fox, Erik B Sudderth, Michael Jordan, and Alan S Willsky, "The sticky hdp-hmm: Bayesian nonparametric hidden markov models with persistent states," 01 2007.

[20] G. Sell and D. Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4794–4798.

[21] Lukáš Burget, "VB Diarization with Eigenvoice and HMM Priors," http://speech.fit.vutbr.cz/software/vb-diarization-eigenvoice-and-hmm-priors, 2013, [Online; January-2017].

[22] Ondřej Novotný, Pavel Matějka, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, and Jan Černocký, "Analysis of speaker recognition systems in realistic scenarios of the sitw 2016 challenge," in *Proceedings of Interspeech 2016*. 2016, pp. 828–832, International Speech Communication Association.

[23] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4930–4934.

[24] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[25] M.J Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, 2003.

[26] Mireia Diez and Lukáš Burget, "Speaker diarization based on bayesian hmm with eigenvoice priors," http://www.fit.vutbr.cz/~mireia/technical_report_Odyssey2018.pdf, 2018, Technical report.

[27] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[28] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan Černocký, "Full-covariance UBM and heavy-tailed PLDA in I-vector speaker verification," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*. 2011, pp. 4828–4831, IEEE Signal Processing Society.

[29] "NIST SRE 2000 Evalplan," https://www.nist.gov/sites/default/files/documents/2017/09/26/spk-2000-plan-v1.0.htm_.pdf.

[30] "NIST Rich Transcription Evaluations," https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation.

[31] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 1, pp. 217–227, Jan. 2014.